Monte Carlo methods for short polypeptides

Jeremy Schofield and Mark A. Ratner
Department of Chemistry, Northwestern University, Evanston, Illinois 60208

(Received 30 June 1998; accepted 17 August 1998)

Nonphysical sampling Monte Carlo techniques that enable average structural properties of short in vacuo polypeptide chains to be calculated accurately are discussed. Updating algorithms developed for Monte Carlo studies of flexible polymer chains are modified and adapted for polypeptide chain systems to improve conformational sampling. Utilizing these methods, the effect of bond angle and bond length constraints in Monte Carlo simulations are examined and it is demonstrated that angle constraints bias structural averages without greatly reducing the computational work. © 1998 American Institute of Physics. [S0021-9606(98)50244-9]

I. INTRODUCTION

Computational study of peptides and proteins provides insights into their structure and dynamics that are difficult to obtain using direct experimental methods. From a computational viewpoint, however, these glass forming materials are quite difficult to study, essentially because their energy landscapes contain many different minima, so that any computational scheme for computing actual thermal averages must be so designed that it can sample all of the appropriate energy surface for any given experiment.

In recent months much attention has focused on the application of nonphysical sampling distributions to estimate thermodynamic averages of model in vacuo polypeptide and protein systems. These non-Boltzmann sampling distributions are designed to avoid the shortcomings of conventional Metropolis Monte Carlo (MC) (Ref. 6) and molecular dynamics (MD) simulations of systems exhibiting frustration in which phase space bottlenecks inhibit thorough exploration of the entire conformational space. Such considerations are not limited to peptides or polymers, but arise in studies of other glass formers. In the interest of computational ease, many studies of the conformations of biologically important molecules reduce the complexity of the physical system by constraining some of the degrees of freedom of the system, such as the bond lengths or bond angles in the molecule. The results of calculations that utilize such constraints correspond to statistical averages over a reduced hyper-surface of phase space rather than over the full space. It is hoped that the potential energy surface of the constrained system closely resembles the full potential energy surface so that the important regions of phase space are weighted properly and averages calculated correctly. Below a high temperature regime, bond lengths fluctuate relatively little due to the nature of the potential that describes vibrations, with fairly stiff harmonic spring constants. Furthermore, many structural properties, such as the radius of gyration, are only weak functions of the bond lengths. It therefore seems reasonable that constraining the bond lengths in a MC calculation to values near their equilibrium averages leads to only minor changes in the predicted values of many average structural quantities. It is not so clear, however, that bond angles can similarly constrained without causing profound differences in the model.

In this paper we utilize “umbrella” sampling and multiple Markov chain methods to calculate physical properties of short polypeptide chains over a wide range of temperature and avoid difficulties of numerical convergence due to quasi-ergodicity. We discuss a variety of trial MC moves, designed specifically to optimize the exploration of conformations in polypeptide chains which should be extendable to larger, more biologically significant systems. We use these techniques to examine the agreement of results for Monte Carlo calculations with and without different types of constraints for average structural properties of the peptide chain as a function of temperature.

In Sec. II we describe the model for the physical system and discuss the methodology of the MC simulation. In the following section we examine the results for the model system and discuss issues pertaining to sampling efficiency and accuracy. Finally we summarize the results of the study in Sec. IV and mention a few ideas currently under investigation to improve the rate of exploration of phase space and to extend the methods used here to larger systems.

II. MODEL AND METHOD

A. Potential energy

Our test systems are five and ten residue glycine chains in which the methylene groups are represented in the united atom approach and the interactions among the atoms in the chain are represented classically by the “Charmm”-like potentials used in the simulation software package PROSIS. The potential energy function utilized is the sum of non-bonded interactions $V_{sb}$, namely, electrostatic interactions and a 12-6 Lennard-Jones term,  

$$ V_{sb} = \sum_{ij} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{K_i}{r_{ij}} \right), $$

where the prime on the summation indicates that the summation includes all pairs of atoms separated by four or more bonds, and the bonded interactions $V_b$. 


*Current address: Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Toronto, Ontario M5S 3H6, Canada.
\[ V = \sum_{i=1}^{N-1} K^b_i (b_i - b_i^0)^2 + \sum_j K^\theta_j (\theta_j - \theta_j^0)^2 \]
\[ + \sum_j V^\text{for}_j (\phi_j). \]  

(2)

The bonded interactions include bond length, bond angle, and dihedral interactions, which have the form
\[ V^\text{for}_j (\phi_j) = K^\text{for}_j (\phi_j - \phi_j^0)^2 \]
for improper dihedral angles and
\[ V^\text{for}_j (\phi_j) = \sum_{n=0}^{3} K^{\text{for} n}_{j, n} \cos(n \phi_j) \]
for proper dihedral angles. In this implementation of the potential energy function there are no explicit hydrogen bonding terms as hydrogen bond interactions are incorporated via the electrostatic interactions with the dielectric constant \( K_d = 332.0638 \) (kcal Å/mole).

**B. Generalized coordinate system**

In this work, we define a local coordinate system for each atom in the chain in which the position of atom \( i \) for \( i = 1, \ldots, N \) is specified in terms of a bond length and two angles \( \theta_i \) and \( \phi_i \). The coordinate transform from Cartesian to the generalized coordinates \( \mathbf{q}^N \) is
\[ \mathbf{r}_i = \mathbf{r}_i (\mathbf{r}_0, \alpha, \beta, \gamma, \mathbf{b}, \theta, \phi), \]
where \( \alpha, \beta, \gamma \) are Euler angles specifying the orientation of the plane defined by the first three atoms, \( \mathbf{r}_0 \) is the Cartesian coordinate of the first atom, and \( \mathbf{b}, \theta, \phi \) represent sets of \( N-1 \) bond lengths, \( N-2 \) bond angles, and \( N-3 \) dihedral angles of the chain. In this generalized coordinate system the bonded terms \( V^b \) in expression (2) and the imposition of the bond angle and bond length constraints are simplified expressed. The position vector \( \mathbf{r}'_j \) in the local bond reference frame \( i \) with \( i < j \) is calculated via
\[ r'_j = \mathbf{r}_j + \delta \mathbf{r}'_j, \]
where \( \delta \mathbf{r}'_j = \mathbf{q}_{\text{op}}^{-1} \delta \mathbf{q}_{\text{op}} \mathbf{q}_{\text{op}}^{-1} \),
where \( \delta \mathbf{r}'_j = \mathbf{r}'_j - \mathbf{r}_i \), \( \mathbf{q}_{\text{op}} \) is the quaternion operator
\[ \mathbf{q}_{\text{op}} = \cos(\Delta \phi/2) + \sin(\Delta \phi/2) \mathbf{u}, \]
and the * notation indicates the quaternion product. For a discussion of the properties of quaternions and their application in molecular physics, see Refs. 18 and 19.) Similarly, bond angle changes can be represented by rotations along a unit vector perpendicular to the plane of the atoms composing the bond angle. According the quaternion algebra, basic moves consisting of simultaneous bond angle and dihedral rotations are represented by the operator
\[ \mathbf{q}_{\text{op}} = \mathbf{q}_b \mathbf{q}_d, \]
where \( \mathbf{q}_b \) and \( \mathbf{q}_d \) are the quaternion operators for isolated bond angle and dihedral rotations, respectively.

The average of an observable \( A(\mathbf{r}^N) \) can be written as an integral over the \( 3N \) spatial degrees of freedom
\[ \bar{A} = \langle A(\mathbf{r}^N) \rangle = \int d\mathbf{r}^N A(\mathbf{r}^N) \rho(\mathbf{r}^N), \]  

(14)
where $\rho(\mathbf{r}^N)$ is the canonical distribution function. In our system of $3N$ generalized coordinates, (14) is

$$\bar{A} = \langle A(q^N) \rangle = \int dq^N J(q^N) A(q^N) \rho(q^N),$$

(15)

where $J(q^N)$ is the Jacobian determinant of the transformation from Cartesian to generalized coordinates in (5). For an isolated polypeptide chain system observables $A(q^N)$ and the distribution function $\rho(q^N)$ are independent of $r_0$, $\alpha$, $\beta$, and $\gamma$. These degrees of freedom can be integrated over, and the expression for $\bar{A}$ complete with the Jacobian determinant becomes

$$\bar{A} = \int db_1 \, db_2 \, d(\cos \theta_3) \times \prod_{i=3}^{N-1} db_i \, d(\cos \theta_i) \, d\phi_i \, b_i^2 b_j^2 b_k^2 A(b, \theta, \phi) \rho(b, \theta, \phi).$$

(16)

Equation (16) is a $3N-6$ dimensional integral that we will calculate numerically.

In modeling large biological systems it has become customary to constrain the bond length and bond angle degrees of freedom to specific values in order to reduce the computational work required to evaluate integrals such as (16). The constraints are justified in Monte Carlo calculations provided that the distribution function $\rho$ has sharply defined narrow peaks around the average values $\langle b_i \rangle$ and $\langle \theta_i \rangle$ at the temperature of interest. In this case the use of constraints corresponds to replacing the full distribution function $\rho$ by a conditional distribution $\rho^c$, which is obtained from $\rho$ by setting the constrained degrees of freedom to their constrained values. If $A(q^N)$ is independent of the some degrees of freedom such as the bond lengths, then the integrals over these degrees of freedom can be carried out and the equilibrium average of $A$ can be written as an average of $A(q^N)$ over the effective reduced distribution function $\rho^{\text{red}}$. There is no guarantee, however, that the conditional distribution function $\rho^c$ will resemble the reduced distribution $\rho^{\text{red}}$. At low temperatures it is generally true that all the bond lengths are close to their equilibrium values due to the fact that the harmonic bond-stretching potentials have relatively large spring constants $K_i^b$ (see Eq. (2)). The bond angle constraint approximation, on the other hand, is more questionable since the spring constants $K_i^{ab}$ parameterizing the bond angle potential are usually an order of magnitude smaller than $K_i^b$.14

There are two different models for examining the constrained averages mentioned above: the flexible model and the rigid model.20 The models differ in the fundamental Hamiltonian on which the classical equilibrium statistics is based. The flexible model assumes that the full potential and the constraints are enforced by taking the limit of infinite spring constants for the bond lengths and/or bond angle potentials. In the limit of infinite spring constants, the harmonic terms in the bonded potential (2) lead to delta functions which constrain the bond length and/or bond angle degrees of freedom to their equilibrium values and

$$\bar{A}^c = \int d\phi \, A^c(\phi) \rho^c(\phi),$$

(17)

where $A^c(\phi)$ is $A(q^N)$ with constrained degrees of freedom.

The rigid model, on the other hand, assumes a Hamiltonian containing only the unconstrained degrees of freedom and is of the form

$$H_i = \sum_{j=1}^{N-1} g_{ij}^{-1}(\phi) P(\phi)_j + V(\phi),$$

(18)

where $g_{ij}(\phi)$ is an angle dependent tensor which couples the momenta conjugate to the generalized coordinates $\phi$. The average of a quantity $X(\phi)$ depending only on the generalized coordinates $\phi$ is therefore

$$\bar{X} = \int d\phi \, \sqrt{\det g(\phi)} \exp\{-\beta V(\phi)\} A(\phi)$$

$$\int d\phi \, \sqrt{\det g(\phi)} \exp\{-\beta V(\phi)\},$$

(19)

where the factor $\sqrt{\det g(\phi)}$ depends on $\phi$ in a complicated way. MD simulations cannot deal with the diverging frequencies that occur in the flexible model in the limit of infinite spring constants and are therefore based upon the rigid model. It is therefore necessary in MD simulations of constrained systems to calculate the “Fixman” potential21

$$V_f(\phi) = K_T T \, \frac{1}{2} \log \det g(\phi)$$

(20)

and propagate the equations of motion based upon the total potential $V(\phi) + V_f(\phi)$. For example, it has been demonstrated that the angular distribution function $\rho(\phi)$ calculated from a constrained MD simulation of a freely-rotating three bond chain is nonuniform unless the Fixman potential is utilized,22 regardless of whether the holonomic constraints are dealt with by the matrix method23 or by the SHAKE algorithm.24

Since the diverging frequencies encountered in the limit of infinite spring constants do not pose a problem in a Monte Carlo calculation, we base our calculations of structural properties of polypeptide chains with and without constraints on the flexible model and thereby avoid the need to calculate the corrective Fixman potential. Constrained averages, such as in Eq. (17), are calculated by generating a Markov chain which has the limiting distribution $\rho^c(\phi)$.

C. Sampling method

Variations of the sampling method, which is based upon thermodynamic scaling techniques,1 used in this work are described in detail elsewhere.2-5,25 In the standard Metropolis Monte Carlo method the Markovian sequence of configurations is generated with a limiting canonical distribution $P_G(\Gamma)$. In the Monte Carlo method of evaluating integrals, one constructs the chain of configurations sequentially by creating a trial configuration which depends on the current state of the system. The trial configuration is accepted as the next configuration in the chain of states or rejected according to an acceptance criterion. If the trial configuration is rejected, the previous state is accepted as the next state in the
chain. In the simplest implementation, a trial state $\Gamma'$ is generated based on the current configuration $\Gamma$ and is accepted with probability $\min(1, T_m)$, where

$$T_m = \exp\{-\beta(V(\Gamma') - V(\Gamma))\}. \quad (21)$$

The proper distribution of states is guaranteed by the condition of detailed balance

$$\Pi(\Gamma' \rightarrow \Gamma) = \Pi(\Gamma \rightarrow \Gamma') T_m, \quad (22)$$

where $\Pi(\Gamma \rightarrow \Gamma')$ is the transition probability of going from state $\Gamma$ to state $\Gamma'$. It is easy to demonstrate that calculations of thermodynamic and average structural properties of the polypeptide chain based on (21) converge slowly at room temperature due to quasi-ergodic behavior of the sampling. In such calculations a subset of configurations appears in the Markovian chain of states too frequently because the sampling migrates very slowly (if at all) out of low potential energy, possibly meta-stable regions of configurational space. Similarly, molecular dynamics simulations do not sample the full configurational space of in vacuo polypeptides efficiently. One possible means of circumventing the computational inefficiency of Metropolis MC is to use “entropy-sampling” MC techniques (ESMC). In ESMC a Markovian chain of states is generated in which the states are distributed not with Boltzmann probability but with probability

$$P(\Gamma) = \exp\left\{-\frac{\mathcal{S}[V(\Gamma)]}{K_b}\right\}. \quad (23)$$

The particular choice of the sampling distribution (23) is designed so that all potential energies $V$ over some targeted range of energy appear with equal probability

$$P(V) = \mathcal{N}(V) \exp\left\{-\frac{\mathcal{S}[V(\Gamma)]}{K_b}\right\} = \text{constant}, \quad (24)$$

where $\mathcal{N}(V)$ is the spectral density (number of configurations with potential energy $V$). Since all energies are equally probable over the targeted range of energy, the simulation is confined in meta-stable regions of phase space for only relatively short sequences of configurations in the Markovian chain of states and ergodic sampling is much more probable provided the simulation is long enough and trial MC states are generated efficiently. Since Eq. (24) implies $\mathcal{S}[V(\Gamma)] \sim K_b \log \mathcal{N}(V), \mathcal{S}[V]$ is identified as the entropy of the system with potential energy $V$. In ESMC a new trial configuration $\Gamma'$ is accepted as the next state in the simulation sequence of states with probability $\min(1, T_a)$, with

$$T_a = \frac{P(\Gamma')}{P(\Gamma)} = \exp\left\{-\frac{\mathcal{S}[V(\Gamma')] - \mathcal{S}[V(\Gamma)]}{K_b}\right\}. \quad (25)$$

This MC method, which is a special case of general thermodynamic scaling MC methods, has also been called “multi-canonical” MC in the literature. One of the major advantages of ESMC is that the structural properties of interest can be obtained from one long simulation run by “re-weighting” the configurations. The canonical average $\langle A(r^n) \rangle_c = \bar{A}(T)$ at temperature $T$ of any configurational quantity $A(r^n)$ can be expressed as the ratio of two simulation averages

$$\bar{A}(T) = \frac{\langle A(r^n) \exp[\beta(T \mathcal{S}[V(r^n)] - V(r^n))] \rangle }{\langle \exp[\beta(T \mathcal{S}[V(r^n)] - V(r^n))] \rangle}. \quad (26)$$

In (26), the notation $\langle \cdots \rangle$ represents an average over the configurations in the simulation sequence. In evaluating such averages, however, care must be taken to insure that the sampling distribution adequately covers the appropriate regions of phase space for the property and physical conditions (such as temperature). Generally, this implies that the sampling distribution has a temperature range over which the calculated results are accurate, and therefore much attention must focus on reliable methods of estimating simulation errors.

In the ESMC finding the appropriate sampling distribution for the targeted physical conditions is usually a non-trivial task, as it generally is in thermodynamic scaling MC. For the simple polypeptide chain system, however, a straightforward algorithm can be implemented which solves for $\mathcal{S}[V]$ self-consistently. The targeted energy range of the system is broken up into discrete increments $\Delta V$. Given an initial guess for the value of $\mathcal{S}[V]$ on the mesh, a simulation is carried out and a histogram $H(V)$ is constructed for the number of conformations from the simulation run which fall in each of the energy bins. Following the run the estimate for $\mathcal{S}[V]$ is improved by

$$S^{\text{new}}[V] = \begin{cases} S^{\text{old}}[V] + K_b(\log H(V) - \log H_{\text{av}}) & \text{for } H_{\text{max}} \geq H(V) \geq H_{\text{min}} \\ S^{\text{old}}[V] + K_b(\log H_{\text{min}} - \log H_{\text{av}}) & \text{for } H(V) < H_{\text{min}} \\ S^{\text{old}}[V] + K_b(\log H_{\text{max}} - \log H_{\text{av}}) & \text{otherwise}, \end{cases} \quad (27)$$

where $H_{\text{min}}$ and $H_{\text{max}}$ are adjustable minimum and maximum values for the histogram $H(V)$ designed to speed up convergence of the algorithm, and $H_{\text{av}}$ is the average number of visits to each bin during an iteration. Iterating the procedure eventually will yield a flat histogram $H(V) \approx H_{\text{av}}$, at which point the proper weight function $\mathcal{S}[V]$ has been found.

A simple example of the application of ESMC nicely illustrates the strengths and weaknesses of the method. Consider the double well potential of the form (units of kcal/mol)

$$V(x) = 16 + ((x + 1)^2 - 4)((x - 1)^2 - 3.8) \quad (28)$$

shown in Fig. 1. At $T = 2500$ K two regions of configura-
tional phase space (x coordinate) are populated (see Fig. 2). At that temperature or below, a Metropolis MC simulation in which a trial \(x_t\) is generated from the current configuration \(x_c\) using a random number \(x_r\) via \(x_t = x_c + (x_r - 0.5) \Delta x\) will never sample both regions of phase space provided the maximum displacement \(\Delta x\) is small (\(\Delta x < 2.0\)), whereas the non-Boltzmann sampling distribution (23) is able to explore all regions of phase space below an energy cut-off provided the cut-off is high enough. The probability distribution is calculated correctly by the ESMC method for any temperature where \(P_B(V)\) falls within the targeted energy range, as is shown in Fig. 2. However if the cut-off is too low for the value of \(\Delta x\) chosen, the ESMC sampling is also quasi-ergodic. Care must be taken in applying the ESMC method to insure that all pertinent potential energy barriers can be overcome and that the energy cut-off is high enough for the ways in which the trial moves are constructed.

For this simple one-dimensional potential, the ESMC is less efficient at any given temperature than Metropolis MC with a large value of \(\Delta x\) due to the fact that, in a ESMC simulation, the system spends a good deal of time exploring regions of phase space that are irrelevant at that temperature. A histogram plot of the x coordinate range of states visited during the simulation shows that the simulation spends a lot of time in high energy regions of the potential surface and relatively little time sampling near the minima of the double well (Fig. 3). This implies the statistics for low temperature canonical averages is likely to be worse than that at higher temperatures. However, even though the sampling method might be an inefficient means of calculating average quantities like \(\bar{x}\) at a specific temperature, \(\bar{x}\) can be calculated accurately over a wide range of temperature (see Fig. 4).

D. Statistical uncertainty

The estimation of statistical uncertainty of canonical averages from ESMC runs is more complicated than that in standard Boltzmann-sampling MC because the canonical averages \(\langle A \rangle_c\) of an observable \(A(x^N)\) are ratios of simulation averages of other quantities of the form

\[
\langle A \rangle_c = \frac{\bar{f}}{\bar{g}},
\]

where \(\bar{f}\) and \(\bar{g}\) are averages over the simulation run. However the variance \(\sigma^2_a\) of the canonical average \(\langle A \rangle_c\) is connected to the variances of \(\bar{f}\) and \(\bar{g}\) via

\[
\sigma^2_a = \langle A \rangle_c^2 \left( \frac{\sigma_f^2}{\bar{f}^2} + \frac{\sigma_g^2}{\bar{g}^2} - 2 \frac{\text{cov}(\bar{f}, \bar{g})}{\bar{f} \bar{g}} \right),
\]
FIG. 4. A plot of the average configurational quantity \( \bar{x} \) as a function of temperature. At low temperatures \( \bar{x} \) approaches the minimum of the double well potential (near \( x = -2.0 \)) and at high temperatures \( \bar{x} \) approaches 0 due to the near-symmetry of the potential \( V(x) \).

where \( \text{cov}(\bar{f}, \bar{g}) \) is the covariance of \( \bar{f} \) and \( \bar{g} \). The quantities \( \sigma_f^2 \) and \( \sigma_g^2 \) can be estimated in a straightforward fashion by applying the “blocking” transform method\(^{29}\). In this method the block variation \( \gamma_f^n \) of \( \bar{f} \) is calculated

\[
\gamma_f^n = \frac{1}{2^n} \sum_{i=1}^{2^n} (f_i - \bar{f})^2,
\]

where \( n \) is the size of the block sub-averages \( f_i \). \( \gamma_f^n \) is calculated for the series of block sizes \( n = 2, 4, \ldots, 2^i \) and the quantity \( \gamma_f^0 = \gamma_f^1/n \) is monitored. Generally \( \gamma_f^0 \) increases and eventually reaches a plateau (with statistical noise) as the size of the block sub-averages \( n \) increases if the data sets are large enough. Since \( \sigma_f^2 \approx \gamma_f^0 \) and \( \lim_{n \to \infty} \gamma_f^0 \) converges to a fixed point, the true statistical error can be obtained easily from this re-normalization group method. In Fig. 5 the estimate for the square of the standard deviation \( \sigma_f^2 \) is shown as a function of block size.

E. Generation of trial moves

One of the most difficult aspects of a challenging computational MC problem is the creation of techniques to generate new trial configurations to test with the acceptance criterion. The manner in which new configurations are generated has a large impact on the rate of phase-space exploration as a function of CPU time, as was observed in the double-well model. If new trial conformations are very different from the current conformation of the system, the potential energy differences are typically large and the acceptance ratio of trial conformations is low. On the other hand if the successive conformations are very similar and the acceptance ratio is high, then the sequences of states in the simulation Markovian chain are strongly correlated and the sampling may be quasi-ergodic, particularly if there are a large number of deep minima in the potential energy surface of the system. The model peptide system examined here falls into this class of systems due to the effects of strong hydrogen bonding which allow for many energetically stable dissimilar conformations of the chain.

Due to the strong restoring forces of the bond length and bond angle potential it is reasonable to expect a fairly simple one-minimum effective potential corresponding to these degrees of freedom. The same is true of the peptide dihedral angles \( \omega \) which tend to be strongly peaked around \( \omega \approx \pi \) due to the partial conjugation of the peptide bond. The other backbone dihedral angles, however, have effective potentials which depend strongly on their local environment, and vary significantly from \( [-\pi, \pi] \) during a simulation. These considerations are important when designing techniques to generate new trial configurations.

1. Nonbiased sampling

The most straightforward way\(^{29}\) of generating a new trial conformation of the peptide chain is to choose a particular atom of the chain and to change its generalized coordinates \( q_i \) by a displacement \( dq_i \), selected uniformly from a flat distribution in \( [-\Delta q_{\text{max}}, \Delta q_{\text{max}}] \).

\[
q_i' = q_i + dq_i.
\]

The maximum displacements \( \Delta q_{\text{max}} \) are chosen to optimize the phase space exploration. For our model system, based on the relative magnitudes of the spring constants in the bonded potential (2), we set

\[
\frac{\Delta b_{\text{max}}}{b_0} < \frac{\Delta \cos \theta_{\text{max}}}{\pi} \approx \frac{\Delta \omega_{\text{max}}}{2\pi} \ll \frac{\Delta \phi_{\text{max}}}{2\pi},
\]

where \( \Delta b_{\text{max}}, \Delta \cos \theta_{\text{max}}, \Delta \omega_{\text{max}}, \) and \( \Delta \phi_{\text{max}} \) correspond to the maximum displacements for the bond length, cosine of the bond angle, improper and peptide dihedral and the “soft” Ramachandran\(^{30}\) backbone dihedral degrees of freedom, respectively. In the ESMC method, a trial configuration \( \Gamma' \) is accepted with probability \( \min(1, T) \), where
The factor involving the bond lengths $b_i$ appearing in (34) is due to the Jacobian determinant for the transformation from Cartesian to generalized coordinates as in Eq. (16). To allow for major changes in conformation via hopping over the barrier walls of locally stable configurations, we set $\Delta \phi_{\text{max}} = \pi$. Due to our choice of coordinate system and the connected nature of the chain molecule, changing the coordinate of backbone atom $i$ along the chain necessitates recalculating the Cartesian coordinates of atoms $i+1$ through $N$. Thus a small local coordinate change can have quite drastic structural effects on the conformation of the overall chain. The local coordinate moves typically break hydrogen bonds when the initial configuration is stable and can create energetically unfavorable, overlapping conformations. The cost in energy of such a move consequently leads to low acceptance ratios and slow sampling in low energy regions of phase space. On the other hand, in high energy regions, the simple unbiased algorithm efficiently generates uncorrelated configurations of the system.

2. Ramachandran configurational bias updating

In order to improve the sampling in the important low energy regions of phase space we bias the soft dihedral angles toward preferred values adopted by low energy structures. At low energies, the two backbone dihedral angles $\phi$ and $\psi$ in each residue in proteins and polypeptides are strongly correlated and sharply peaked around characteristic residue values. The strong preference for certain values of $\phi-\psi$ in low temperature polypeptides is reflected in histogram plots of the $\phi-\psi$ dihedral angles, known as Ramachandran maps. To incorporate the dihedral angle preferences into a MC algorithm, the important values for the Ramachandran dihedrals are identified by carrying out several conjugate gradient potential energy minimizations to locate different low energy, stable conformations. Each conformation is classified by its backbone dihedrals and important sets of dihedral angles $\{\phi, \psi\}$ are identified. This information is used to construct normalized Gaussian distributions of width $\alpha$.

$$
T = \exp \left\{ \frac{S(\Gamma) - S(\Gamma')}{{K_b}} \prod_{i=1}^{N-1} \frac{b_i}{b_i'} \right\}.
$$

(34)

from which the new trial values of dihedral angles $\phi_i'$ and $\psi_i'$ are drawn. In Eq. (35), the sum over index $j$ extends over the entire set of preferred angles and $\beta_{cb}$ is an adjustable parameter. The Ramachandran distribution $\rho_1(\phi, \psi)$ can be incorporated into a configurational bias algorithm as follows: First a residue of the chain is chosen at random to be the starting residue for a regeneration process. A trial chain is constructed by regrowing the chain residue by residue, starting at the initial residue $i=1$ and ending at the last residue of the chain $i=n$. At each step in the regrowth process, trial sets of $k$ pairs of dihedrals $\{(\phi_l, \psi_l); l=1, \ldots, k\}$ are drawn with probability $p_l(\phi_l, \psi_l).$ Each set of angles for residue $i$ is assigned a probability of selection

$$
p^i_l = \exp \{-\beta_{cb} U_i(\phi_l, \psi_l)\} \exp \{\beta_{cb} V_{\text{ex}}(\phi_l, \psi_l)\},
$$

(36)

with

$$
W^i_l = \sum_{i=1}^k \exp \{-\beta_{cb} U_i(\phi_l, \psi_l)\} \exp \{\beta_{cb} V_{\text{ex}}(\phi_l, \psi_l)\},
$$

(37)

where $U_i(\phi_l, \psi_l)$ is the biasing interaction potential of residue $i$ with the rest of the chain preceding it. Once the $i$th residue angles $\{(\phi^{(i)}, \psi^{(i)})\}$ have been chosen, the process is repeated for the $(i+1)$th residue. After generating the trial chain, the new conformation $\Gamma_2$ is accepted with probability $\text{min}(1, T)$, with the parameter $T$ given by

$$
T = \exp \left\{ \frac{S(\Gamma_1) - S(\Gamma_2)}{{K_b}} \right\} \times \exp \{\beta_{cb} (V'(\Gamma_2) - V'(\Gamma_1))\} \frac{W_2}{W_1},
$$

(38)

where

$$
V'(\Gamma) = \sum_{i=1}^n (U_i(\phi^{(i)}, \psi^{(i)}) - V_{\text{ex}}(\phi^{(i)}, \psi^{(i)}))
$$

(39)

and

$$
W_2 = \prod_{i=1}^n W^i_l
$$

(40)

is the Rosenbluth weight for the trial configuration $\Gamma_2$. The Rosenbluth weight $W_1$ for the original configuration is calculated similarly to $W_2$ except that at each residue $k-1$ pairs of dihedrals are drawn from $\rho_1$ with the first set $(\phi_0, \psi_0)$ being the actual angles of the original chain. It is also possible to vary other degrees of freedom, such as the bond angles, while constructing the trial chain $\Gamma_2$.

We have found that this adapted configurational bias algorithm, in which the dihedral angles are biased towards values observed in low temperature conformations, is about 3 times more efficient than conventional configurational biasing methods for polypeptides at inducing significant configurational changes in low energy structures. Since the Ramachandran angles are often strongly dependent on the conformation of subsequent regions of the chain, it is often useful to base the biasing potential $U_i(\phi, \psi)$ upon a scaled version of the full potential to control the degree to which certain interactions and local environments dictate the choice of backbone dihedrals.

3. Multiple Markov chains

The idea behind biased Monte Carlo updates of drawing trial states from a known distribution can be utilized to greatly accelerate conformational sampling. In biased MC, a new trial conformation $\Gamma_2$ is drawn from a known distribution $\rho_1(\Gamma_2)$ and accepted with probability $\text{min}(1, T)$, where
\[ T = \frac{P(\Gamma_2) \rho_i(\Gamma_1)}{P(\Gamma_1) \rho_i(\Gamma_2)}. \]

Of course one means of generating states according to a distribution is via a Monte Carlo algorithm. Geyer\(^{11}\) has recently described a method in which multiple Markov chains are evolved in parallel for a specified number of steps. After the independent evolution of the chains, swaps of configurations are attempted between a pair of chains with the acceptance parameter given by \((41)\). The configurations are mixed if the distributions \(P(\Gamma)\) and \(\rho_i(\Gamma)\) have significant overlap in the energy regions appropriate for the conformations \(\Gamma_1\) and \(\Gamma_2\). If the set of distributions is chosen to span from high energy regions, where the simulation correlation time is short, to low energy regions, where configurational sampling is slow, canonical equilibrium averages can be calculated from the Markov chains by weighting averages over individual chains.\(^{12}\) The distributions \(\rho_i\) centered around high energy conformations (i.e., high temperatures) act effectively like a heat bath which randomizes the low temperature Markov chains. To maximize the swapping of conformations, it is helpful to utilize artificial "umbrella" distribution functions rather than a large number of Boltzmann distributions at different temperatures. The shape of the artificial distributions can be adjusted iteratively using very short MC runs with an updating algorithm like \((27)\) to control the rate of configurational sampling. Unlike the pure umbrella-sampling prescription, relatively little effort or computational time is required to construct the artificial distribution functions used in the final calculation. Another major advantage of the multiple Markov chain method over any single chain method is the ease and efficiency of parallelizing computer code to run on multiple CPU (or clustered) computers.

4. Hybrid MC

Both the biased and unbiased updating algorithms elaborated above suffer the drawback of creating drastic non-local changes in the peptide chain conformation. The hybrid MC method\(^{34}\) uses time-reversible molecular dynamics to generate new trial configurations which are accepted or rejected in accordance with a modified acceptance criterion ensuring detailed balance. The algorithm has the advantage that a select group, or all, of the coordinates are changed collectively and simultaneously. The collective changing of coordinates is often extremely important in cases where phase-space bottlenecks lead to quasi-ergodic sampling. Although hybrid MC was originally designed to sample a canonical distribution with Boltzmann weighting, it is straightforward to extend the method to more general weight functions. In hybrid MC, fictitious momenta conjugate to the spatial coordinates (Cartesian or general) are introduced and time-reversible MD integration is performed for a specified length of time, using some time-step \(\delta t_{\text{int}}\), starting from the current configuration with momenta randomly drawn from a Maxwellian distribution. The change in the total effective Hamiltonian \(\Delta H\) is calculated and the new trial configuration accepted with probability \(\min(1, \exp(-\Delta H/\hbar^2))\). The extension of this algorithm to ES/MC is made by naming \(S[V]/\hbar^2\) the "effective" potential.\(^{35}\) The instantaneous forces used in the time propagation are therefore merely a rescaling of the standard force \(-\nabla_i V\),

\[ F_i = \frac{-1}{K_i} \frac{dS}{dV} \nabla_i V(r^N). \]

The utility of hybrid MC lies in the great flexibility of designing trial moves. An arbitrary set of system coordinates can be changed simultaneously under any driving potential desired provided that the "effective" Hamiltonian is used in the acceptance criterion. For the peptide system the important coordinates to change are clearly the Ramachandran dihedral angles and perhaps some bond angles to facilitate the motions of conformational change. Bond lengths and peptide dihedral angles have strong restoring forces and remain relatively constant throughout most conformational changes at low temperatures and are therefore best left constrained during the dynamical propagation. We represent the subset of generalized coordinates \(q^*\) of interest as \(q^s\), and postulate a model driving Hamiltonian for the coordinates \(q^s\),

\[ \tilde{H} = \sum_k \frac{\Pi_k^2}{2I_k} + S[V(q^s)], \]

where \(\Pi_k\) is the fictitious momentum coordinate conjugate to the generalized coordinate \(q^s_k\), and \(I_k\) are the effective moments of inertia. Since the momenta \(\Pi_k\) are uncoupled by design in \((43)\), the resulting equations of motion are simple,

\[ \frac{d\Pi_k}{dt} = -\frac{1}{K_k} \frac{dS}{dV} \frac{\partial V}{\partial q^s_k}, \]

\[ \frac{dq^s_k}{dt} = \frac{\Pi_k}{I_k}. \]

The equations of motion \((44)\) are integrated numerically using a leap-frog algorithm to ensure detailed balance. If coordinate \(q^s_k\) is altered the Cartesian coordinate positions of all subsequent atoms of the chain must be recalculated. This implies that the effective moment of inertia tensor should decrease along the sites of the chain as an end of the chain is approached. In their MC studies of dense polymer systems, Forrest and Suter\(^{36}\) found for a polybead chain of \(N\) sites that setting

\[ I_k \sim (N + 1 - k)^{2.5} \]

maximizes the exploration of configurational space. The generalized forces in \((44)\) are computed using the facts that the derivatives of Cartesian coordinates with respect to the generalized coordinates are given by

\[
\begin{align*}
\frac{\partial r_j}{\partial \phi_i} &= \begin{cases} 
\mathbf{u}_i \times (r_j - r_i) & \text{if } i < j \\
0 & \text{otherwise}
\end{cases} \\
\frac{\partial r_j}{\partial \theta_i} &= \begin{cases} 
\frac{\mathbf{u}_i \times \mathbf{u}_{i-1}}{|\mathbf{u}_i \times \mathbf{u}_{i-1}|} \times (r_j - r_i) & \text{if } i < j \\
0 & \text{otherwise,}
\end{cases}
\end{align*}
\]

where \(\mathbf{u}_i\) is the unit vector along the backbone connecting backbone atom \(i\) to the previous backbone atom. At each
time step in the leapfrog propagation, the generalized forces are calculated for the configuration and the derivative $dS/dV$ is estimated using finite difference methods.

5. Concerted rotations

One of the major drawbacks of using the Flory local coordinate system in a MC calculation is the difficulty of generating local structural changes by varying the locally defined generalized coordinate—a simple rotation around backbone dihedral $\phi_i$ changes the Cartesian coordinates from $i + 1$ to the end of the chain and can often result in high energy overlapping configurations. It is therefore interesting to consider elementary MC moves that modify internal degrees of freedom in a coordinated way so that only local regions of the chain are altered and the problem of thrashing the long chain section following a rotated dihedral is mitigated.

The geometrical problem of finding the correct combination of rotations of a series of dihedral angles which result in local chain rearrangement has been solved for general polymer beads. In the polymer bead chain six consecutive dihedrals numbered (locally) $\{\phi_1, \ldots, \phi_6\}$ are altered while keeping all other atoms fixed except for those numbered locally $\{1, \ldots, 4\}$. All atoms before 1 and after 4 are unchanged provided all torsional angles before $\phi_1$ are fixed and the positions of atoms $\{5,6,7\}$ and $\phi_2$ are unchanged. The geometrical constraints are enforced by requiring that the Cartesian coordinate vector of atom 5, the unit bond vector $\mathbf{u}_6$ connecting atoms 5 and 6, and $\gamma_6$, the Euler angle specifying the orientation of the bond vector $\mathbf{u}_7$ around the axis defined by $\mathbf{u}_6$, remain unchanged. Six constraint equations for the six unknowns result from the geometrical restrictions for the remaining degrees of freedom. Due to the definition of the local coordinate frames of reference, the geometrical problem can be reduced to finding the zeros of a nonlinear function of $\phi_1$ only.

$$F_5(\phi_1) = 0. \quad (47)$$

There are four different branches of $F_5$ to search for roots. A zero of a branch of $F_5(\phi_1)$ provides a solution set $\{\phi_1, \ldots, \phi_6\}$ of dihedrals which correspond to angles which result in crank shaftlike local structural changes. Generally speaking, the topology of the $F_5 - \phi_1$ space is fairly complex, where each branch of $F_5(\phi_1)$ is defined for a range of $\phi_1$. Near the edges of each range the branches merge to form closed curves in $F_5 - \phi_1$. Thus, excepting the case where a branch has a maximum or minimum at $F_5(\phi_1) = 0$, there are an even number of solution sets to the geometrical problem.

Changing six consecutive backbone dihedrals is extremely energetically costly in the case of a peptide chain due to the stiff spring constants associated with peptide bonds. For this reason, we generalize the concerted rotation method to allow for rigid rotation of the C–N bond of each residue. In the Appendix it is demonstrated that the atoms of two consecutive residues can be altered in a cooperative fashion leaving the peptide dihedrals intact. As in the standard concerted rotation method, the solution sets of angles $\{\phi_1, \phi_2, \phi_3\}$ correspond to zeros of a nonlinear, multibranched function of the angle $\phi_1$ (see Fig. 6). Once $\phi_1$ is specified, the other dihedral angles can be solved for successively and the new Cartesian coordinate positions of the atoms in the altered residues calculated.

To implement the concerted rotation solution of the geometrical constraint problem in a MC algorithm, it is essential to note that the dihedral angles $\phi_i$ are no longer independent variables. In order to solve the geometrical problem we transform six dihedral angles to six coordinates which reflect the constraints, $\{r_j, u_k, \gamma_k\}$. To account for the coordinate transformation, a Jacobian determinant relating the volume elements in the two coordinate frames must be included,

$$d\phi_1 d\psi_1 d\phi_2 d\psi_2 d\phi_3 d\psi_3 = J \frac{\phi_1, \psi_1, \phi_2, \psi_2, \phi_3, \psi_3}{r_j, u_k, \gamma_k} dr_j du_k d\gamma_k. \quad (48)$$

This Jacobian is straightforward to calculate using (46). As a test of the Jacobian, a MC simulation was performed on a simplified freely-rotating polypeptide chain. It was found that only when the Jacobian in (48) is included in the MC acceptance criterion are the correct uniform dihedral distributions obtained (see Fig. 7).

The most time-consuming part of the concerted rotation MC method is the search for the roots of the polynomial. Thus the most efficient implementation of the concerted residue rotation method is likely to be to weight all the solutions obtained by the correct weight factor and to choose the trial move based on that weight factor. For example, suppose there are $n$ solutions in the solution set $\{\Gamma_1, \ldots, \Gamma_n\}$. The new trial conformation $\Gamma_j$ is chosen out of the solution set with probability

$$\alpha = \frac{\exp \left( - \frac{S(\Gamma_j)}{K_b} \right) J(\Gamma_j)}{\sum_{i=1}^{n} \exp \left( - \frac{S(\Gamma_i)}{K_b} \right) J(\Gamma_i)}. \quad (49)$$

FIG. 6. A plot of $F(\phi_1)$ vs the first torsional angle $\phi_1$. Each point where $F(\phi_1)$ crosses the x axis represents a solution set of angles $\{\phi_1, \phi_2, \phi_3\}$ for a concerted rotation.
Note that the probability of selecting a particular solution out of the set of solutions includes the Jacobian factor. If this factor is neglected the MC transition probabilities do not obey detailed balance and the configurations will not be generated with the correct limiting distribution.

A MC algorithm can also be designed in which some of the other “hard” degrees of freedom $q^i$, such as the bond lengths or bond angles, are changed by some uniformly distributed $\Delta q^i$. In this case the solution sets must be calculated for the initial “hard” coordinates as well as the destination “hard” coordinates. Suppose there are $n$ solutions for the initial coordinate set and $n'$ solutions with the destination coordinate set. In order to satisfy the condition of detailed balance, a conformation $\Gamma'_i$ chosen from the $n'$ possible destination solutions with probability

$$
\alpha' = \frac{\exp\left(\frac{-S(\Gamma'_i)}{K_b}\right) J(\Gamma'_i)}{\sum_{i=1}^{n} \exp\left(\frac{-S(\Gamma'_i)}{K_b}\right) J(\Gamma'_i)}
$$

is accepted with probability $\min(1,C'/C)$, with

$$
C' = \sum_{j=1}^{n} \exp\left(\frac{-S(\Gamma'_j)}{K_b}\right) J(\Gamma'_j) \prod_{i=1}^{N-1} (b_i^j)^2,
$$

$$
C = \sum_{j=1}^{n} \exp\left(\frac{-S(\Gamma_j)}{K_b}\right) J(\Gamma_j) \prod_{i=1}^{N-1} (b_i^j)^2.
$$

Since the root searches of the constraint equation are computationally expensive, both algorithms are relatively inefficient compared to other methods of generating trial moves for short chains. However, since the concerted rotation moves cause local structural changes by design, the algorithm becomes competitive for long chains and dense, entangled systems. Furthermore, $n$ distinct local regions can be changed simultaneously and the resulting number of possible solutions in the total solution set to choose from increases faster than $2^n$. The computational work required scales linearly with $n$ since each local region is independent from the others. Such qualities suggest the algorithm will benefit greatly from parallelizations and be of greater utility on larger systems.

### III. RESULTS AND DISCUSSION

#### A. Simulation specifics

The updating algorithms elaborated in the previous section were applied in “umbrella”-sampling MC studies of two test systems, gly$_5$ and gly$_{10}$. The initial step in the MC studies of these systems was to carry out molecular mechanics or simulated annealing calculations to locate the potential energy minimum $V_{\text{min}}$ for the model systems. After locating $V_{\text{min}}$, short ESMC runs were done to solve for the weight function parameter $S[\gamma]$ self-consistently over the targeted energy range $[V_{\text{min}}, V_{\text{max}}]$ according to Eq. (27). It is important to set $V_{\text{max}}$ high enough so that the conformations in the initial simulation chain of states are uncorrelated given the particular choice of techniques for generating trial conformations. It is also important, particularly for short peptide chains, to include MC moves in which peptide torsional angles are flipped by $\pm \pi$ so that the entire conformational space of the model system is explored. Short glycine peptide chains have a significant number of conformations populated at high temperatures in which some of the peptide dihedral angles are cis, unlike longer chains in which all peptide dihedrals are trans.

For the gly$_5$ and gly$_{10}$ model systems the iteration process was started with the initial choice of the form $S[\gamma] = a_0 \rho \log(\Delta V) + a_1 \rho \Delta V$, where $\Delta V = V - V_{\text{min}}$, $n$ is the number of degrees of freedom of the chain, and $a_0$ and $a_1$ are system dependent constants. This particular choice of initial weight function is motivated by the fact that the sampling in the ESMC method is essentially “umbrella sampling” in which the overlapping distributions composing the weight function are Boltzmann distributions at different inverse temperatures $\beta$. Thus

$$
\exp\left(\frac{-S[\gamma]}{K_b}\right) = W(\Delta V; \beta_{\text{min}})
$$

$$
= \left(\int_{\beta_{\text{min}}}^{\infty} d\beta \exp(-\beta \Delta V)\right)^n
$$

$$
= \left(\frac{\exp(-\beta_{\text{min}} \Delta V)}{\Delta V}\right)^n,
$$

which implies that

$$
S[\gamma] \approx a_0 n \log \Delta V + a_1 n \Delta V,
$$

where $a_0$ and $a_1$ are system dependent constants which must be solved for iteratively. The exact form of $S[\gamma]$, which is

![Free Torsional Model](image-url)
The steep drop off of the region of the global energy minimum is approached. Since reflecting a rapid decrease in the density of states as the weight function drops logarithmically to zero, it is customary looking at distributions of angular degrees of freedom. The number of significant conformations at a given temperature can be identified by systematically looking at distributions of angular degrees of freedom. However, the weight function drops logarithmically to zero, and this can be mitigated by the use of an energy-based sampling scheme, such as the Reverse Jump Metropolis (RJM) or the Metropolis-Hastings (MH) algorithm.

In the previous section we mentioned that it has become customary to constrain certain degrees of freedom to reduce the computational work involved in calculating average structural properties from simulations. Such simplifying assumptions are justified when constrained averages $\bar{A}^c$ calculated for a structural property $A(q^N)$ differ little from the true average $\bar{A}$. That is, $\bar{A}^c$ differs little from $\bar{A}$.

In Fig. 8 we show the probability $P(V) = N(V) \times \exp(-\beta V)$ of potential energy $V$ at temperatures $T_{\text{min}}$ and $T_{\text{max}}$. In this figure, the energy histogram for the simulation run is plotted over a range of temperatures, where $T_{\text{min}} = 200$ and $T_{\text{max}} = 1000$. Since both distributions fall within the targeted energy range over which the simulation samples, average structural properties of the model system can be calculated extremely accurately over the temperature range $[T_{\text{min}}, T_{\text{max}}]$. In Fig. 9 the square of the radius of gyration $R_{\text{gyr}}^2$ is plotted over a range of temperatures, where

$$R_{\text{gyr}}^2 = \frac{1}{M} \sum_{i=1}^{N} M_i (R_i - R_{\text{cm}})^2.$$

In (54), $M_i$ is the mass of atom (or unified atom) $i$, $M = \sum M_i$, and $R_{\text{cm}}$ is the center of mass of the system. Since the radius of gyration is a good indicator of relative compactness of a molecule, it is clear from Fig. 9 that glycine adopts a fairly compact, globular conformation on average at low temperatures. Near $T = 400$ K the average radius of gyration starts to increase dramatically, indicating more extended average conformations. At low temperatures, little variation is seen in $R_{\text{gyr}}$ which indicates that the low temperature distributions are sharply peaked around the average radius of gyration. In the vicinity of the temperature region near $T = 400$ K there is a rapid broadening in the distributions of $R_{\text{gyr}}^2$, indicating that as temperature increases, more and more conformations with appreciably different compactness contribute to the observed structural averages.

### B. Effect of constraints

In the previous section we mentioned that it has become customary to constrain certain degrees of freedom to reduce the computational work involved in calculating average structural properties from simulations. Such simplifying assumptions are justified when constrained averages $\bar{A}^c$ calculated for a structural property $A(q^N)$ differ little from the true average $\bar{A}$. That is,
observables

A

angle and bond length constraints.

the predicted constrained averages

properties for our polypeptide systems accurately, we compare

ESMC simulations calculate the averages of structural prop-

with constrained degrees of freedom. Since we believe the

model polypeptide system, minor quantitative differences be-

radius of gyration calculated with and without bond angle

angles are typically stretched. The angle constraints also re-

of several different

observables \(A(r^N)\) to the full averages \(A\) subject to bond

angle and bond length constraints.

The average \(R_{\text{gyr}}^2\) and its variation calculated for the gly

system subject to bond length constraints agree within statistical uncertainty with the unconstrained averages over the entire temperature range. However constraining the bond angle degrees of freedom in addition to the bond lengths leads to significant differences between calculated values of the radius of gyration and its variation in the constrained and unconstrained systems over most of the temperature range. In the simulation studies of the constrained system the bond angles are fixed to values observed in the minimum energy conformation, the meta-stable conformation. If the bond angles are fixed to the values observed in the minimum energy conformation, the meta-stable conformation is de-stabilized to a potential energy value \(-250\) kcal/mol. Slight adjustments of the other unconstrained degrees of freedom can lower the energy to around \(-260\) kcal/mol while maintaining the overall structure. Thus it appears the bond angle constraints greatly alter the relative stability of important low-energy meta-stable conformations and thereby affect the statistical weight of the conformations. In general, such meta-stable wells have a small density of states compared to the minimum energy well and make little contribution to many statistical averages, but some observables, such as the average radius of gyration, are affected.

One might expect that averages of observables which are independent of the constrained degrees of freedom will have less dramatic differences. In Fig. 12 plots of distribution functions of the central backbone torsional angle of the gly

system under different constraints at several different temperatures are shown. The multi-modal peaks in the conformational distributions indicate that more than one conformation contributes to structural averages at the specified temperatures. Clearly there are differences between the angular distributions calculated from simulations run with and without bond angle constraints at both temperatures plotted. At \(T = 500\) K (upper plot of Fig. 12), the angular distribution functions of the system with bond angle constraints are more sharply peaked around angles corresponding to low energy structures compared to the distributions calculated for the unconstrained system. At this temperature, the average con-

\[
\bar{A}^c = \int dX A^c(X) \rho^c(X) \approx \bar{A} = \int dX A(r^N) \rho(r^N),
\]

where \(X\) are the unconstrained degrees of freedom, \(\rho^c(X)\) is the conditional distribution of \(\rho(r^N)\), and \(A^c(X)\) is \(A(r^N)\) with constrained degrees of freedom. Since we believe the ESMC simulations calculate the averages of structural properties for our polypeptide systems accurately, we compare the predicted constrained averages \(\bar{A}^c\) of several different observables \(A(r^N)\) to the full averages \(\bar{A}\) subject to bond angle and bond length constraints.

Although there are clear differences between the average radius of gyration calculated with and without bond angle constraints, the qualitative behavior of the averages as a function of temperature is similar. Given the crudity of the model polypeptide system, minor quantitative differences be-

between constrained and unconstrained averages of \(R_{\text{gyr}}^2\) are relatively unimportant for short peptide chains. However for larger systems the lack of flexibility of the peptide chain with frozen bond angles leads to qualitatively as well as quantitatively different behavior. In Fig. 11 two low energy con-

formations of gly\(_{10}\) are shown. The cork-screw shaped conformation has a minimum energy of approximately \(-310\) kcal/mol in the PROSIS force field, whereas the meta-stable conformation in Figure 11 has a minimum energy of \(-295\) kcal/mol. The meta-stable conformation is a sensitive function of the bond angles, however, and small variations in these angles lead to large changes in the potential energy of the conformation. If the bond angles are fixed to the values observed in the minimum energy conformation, the meta-stable conformation is de-stabilized to a potential energy value \(-250\) kcal/mol. Slight adjustments of the other unconstrained degrees of freedom can lower the energy to around \(-260\) kcal/mol while maintaining the overall structure. Thus it appears the bond angle constraints greatly alter the relative stability of important low-energy meta-stable conformations and thereby affect the statistical weight of the conformations. In general, such meta-stable wells have a small density of states compared to the minimum energy well and make little contribution to many statistical averages, but some observables, such as the average radius of gyration, are affected.

One might expect that averages of observables which are independent of the constrained degrees of freedom will have less dramatic differences. In Fig. 12 plots of distribution functions of the central backbone torsional angle of the gly\(_5\) system under different constraints at several different temperatures are shown. The multi-modal peaks in the conformational distributions indicate that more than one conformation contributes to structural averages at the specified temperatures. Clearly there are differences between the angular distributions calculated from simulations run with and without bond angle constraints at both temperatures plotted. At \(T = 500\) K (upper plot of Fig. 12), the angular distribution functions of the system with bond angle constraints are more sharply peaked around angles corresponding to low energy structures compared to the distributions calculated for the unconstrained system. At this temperature, the average con-
formation in the unconstrained system is less compact than that observed in the system with angle constraints. At low temperatures ($T = 200 \text{ K}$), the distribution functions for the unconstrained system show small peaks near $\phi = 3\pi/4$, $5\pi/4$ which are not observed in the angle-constrained system. The small peaks correspond to conformations which do not contribute to structural averages at this temperature in the constrained model system.

C. Sampling efficiency

It is necessary to devote more than usual attention to sampling efficiency due to the great difficulty of thoroughly sampling the configurational space of the polypeptide systems. A good measure of the sampling efficiency of the simulations can be obtained by examining the decay of (Monte Carlo) time dependent correlation functions since the number of MC steps should scale roughly with the CPU time. The auto-correlation functions of several different observables of the Gly5 system are shown in Fig. 13. All the auto-correlation functions exhibit rapid exponential decay, indicating that only short segments of the simulation chain of states are correlated with one another. For the larger Gly10 system, however, the correlation time of the conformations is somewhat longer. As the peptide chain length increases, the depth of the potential energy well corresponding to a locally stable low energy conformation becomes greater and leads to trapping. In the peptide systems all low energy conformations are stabilized by many hydrogen bonds which must be broken to escape from the local energy wells. Such trapping behavior suggests that the weight function $S[V]$ should be modified for longer chains.

One means of modifying the weight function to obviate inefficient sampling due to hydrogen bond stabilization is to selectively scale the terms in the potential energy of the model Hamiltonian which account for hydrogen bonding.

$$V_{\alpha \beta}(\alpha) = \sum_{i,j} \left( A_{ij} \frac{1}{r_{ij}^6} - B_{ij} \frac{1}{r_{ij}^6} + \alpha_{ij} (r_{ij}^N) \frac{K_d}{r_{ij}} \right).$$

The scaling parameter $\alpha_{ij}(r_N)$ can be constant or can depend on the conformation of the system in a complicated fashion. Simulations are carried out in the standard fashion with the modified Hamiltonian with the stipulation that, for each recorded MC conformation, the unscaled ($\alpha_{ij} = 1$) potential energy as well as the value of the weight function for the conformation must be recorded so that the re-weighting procedure can produce the correct equilibrium averages via Eq. (26).

The simplest useful modification of the standard ESMC method is to set $\alpha_{ij}(r_N) = c$ with $c < 1$, which effectively weakens all the electrostatic and hydrogen bonding interactions in the system. Simulations of the Gly10 system with the weight function based upon the modified Hamiltonian with $\alpha_{ij} = 0.5$ no longer spend many consecutive MC steps in a particular conformational well, as an inspection of the simulation time series of $R_{gyr}^2$ shows (upper plot of Fig. 14). The modified ESMC procedure reduces the system correlation time by an order of magnitude (see Fig. 15), and is likely to lead to greater improvements in efficiency in larger systems.

The simple Hamiltonian scaling ESMC mentioned above works well for the Gly10 system, but does not selectively scale interaction terms based on structural information and is therefore not optimally designed for studies of peptide conformation. Since typically only extremely compact conformations exhibit long simulation auto-correlation times (see Fig. 14), a hydrogen bonding scaling parameter $\alpha_{ij}(r_N)$ which depends explicitly on structural properties such as the radius of gyration would likely lead to more efficient sampling, particularly for longer chains. Such sampling methods are currently under investigation.
FIG. 14. A slice of the time series of the observable $R^2_{gyr}$. The upper plot is from a ESMC run with the scaled Hamiltonian parameter $a=0.5$, and the lower plot is from a simulation run with the full Hamiltonian ($a=1$).

IV. SUMMARY

In this article we have shown that MC simulations based on non-Boltzmann sampling are well-suited to address fundamental questions of polypeptide structure. By applying such MC methods to short peptide systems, we have shown that constraining the bond length degrees of freedom introduces little qualitative or quantitative change in the predicted thermodynamic averages of structural properties of the peptide chain. However, some average structural properties calculated from MC simulations in which bond angle degrees of freedom are frozen in addition to bond length constraints differ appreciably from full system averages. While the qualitative differences of averages of short peptide systems may seem relatively unimportant, there is no guarantee that the constraints will not lead to significant qualitative and quantitative differences in larger systems. In any event, bond angle constraints do not lead to a great reduction in computational effort in MC simulations of this sort and should therefore be avoided to protect against unphysically biasing the calculations.

The standard ESMC algorithm is inefficient for larger polypeptide systems, due to strong hydrogen bond stabilization effects in low energy conformations. Scaling the hydrogen bonding terms in the Hamiltonian effectively weakens the stability of these conformations and leads to a tenfold decrease in the simulations correlation times for the gly10 system. As the number of residues increases, it becomes more and more difficult to find a weight function $S[V]$ suitable for the targeted energy range. For large polypeptide or protein systems, the multiple Markov chain method avoids the computationally intensive task of iteratively solving for the weight function. For these systems, multiple Markov chain MC techniques, based on weight functions that incorporate detailed structural information in the Hamiltonian and hydrogen bond scaling parameters, in combination with hybrid MC algorithms with selectively constrained segments, are likely to yield optimal results.37

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jiří Kolafa for assistance with the code development and the use of his molecular dynamics and mechanics simulation package PROSIS. One of the authors (J.S.) would like to thank Robert Harley for useful comments on code optimizations. The research described herein was supported in part by the DOD/MVSLI Program, and partly by the DOE/LBL Program on Advanced Batteries.

APPENDIX

In this appendix we demonstrate that the geometrical problem of finding the combination of backbone dihedral angles that locally rotate two residues of the peptide chain while leaving the peptide dihedrals and the rest of the chain intact can be reduced to finding the roots of a function of only one variable $\phi_1$. Our discussion closely follows that in the Appendix of Ref. 17 which elaborates an analogous method of solution in greater detail.

The constraint conditions are that the vectors

$$r_i^{(1)} = (T^{lab}_i)^T \cdot (r_i - r_0) = l_i + T_1 \cdot l_2 + \cdots + T_1 \cdot T_2 \cdot \cdots T_6 \cdot l_7,$$

$$u_i^{(1)} = T_1 \cdot T_2 \cdot \cdots T_6 \cdot l_7 \cdot e_1,$$

with $e_1 = (1,0,0)^T$, are constant as the dihedral angles $\phi_1$, $\phi_2$, $\phi_4$, $\phi_5$, $\phi_7$ and $\phi_8$. Angles $\phi_3$ and $\phi_6$ are peptide dihedrals which remain fixed. The first step is to relate $\phi_2$ to $\phi_1$ noting that

$$t = T_1^T \cdot (r^{(1)}_1 - l_1)$$

(A2)

depends only on $\phi_1$. Defining

FIG. 15. Angular auto-correlation functions of an interior backbone torsional angle (fifth residue) of a 10 residue glycine system calculated from simulation runs with and without electrostatic scaling parameters in the system Hamiltonian.
Once the first 1–6 dihedrals of the chain are calculated, the remaining dihedrals of atoms 1–4 can be calculated. Manipulating Eq. (A4) yields an expression which allows us to solve for \( \phi_2 \) in terms of \( \phi_1 \) via

\[
\alpha \cos \phi_2 + \beta \sin \phi_2 = t^2 + q_2^2 - q_1^2
\]

(5)

with

\[
\alpha = 2t_5(q_{1x} \cos \theta_2 - q_{1y} \sin \theta_2)
\]

\[
\beta = 2t_5q_{1z} + 2t_6q_{1x} \sin \theta_2 - q_{1y} \cos \theta_2
\]

(6)

where \( a_x \) refers to the x component of vector \( a \) = \( \{a_x, a_y, a_z\} \). Using \( \phi_1 \), \( \phi_2 \), and \( \phi_4 \), the Cartesian positions of atoms 1–4 can be calculated.

Once \( \phi_3 \) is known as a function of \( \phi_1 \), we can solve for \( \phi_4 \) in terms of \( \phi_1, \phi_2, \) and \( \phi_3 \) in exactly the same fashion by applying the redefinitions

\[
\mathbf{r}_1 = \mathbf{T}_3 \cdot \mathbf{r}_2
\]

\[
\mathbf{q}_1 = \mathbf{T}_4 \cdot \mathbf{q}_2
\]

(7)

in Eq. (5). The Cartesian coordinate position of atom 5 can now be calculated.

Finally, \( \phi_5 \) can be obtained from the equation

\[
t = \mathbf{T}_3^{lab} \cdot (\mathbf{r}_7 - \mathbf{r}_6) = \mathbf{I}_5 + \mathbf{q}_2
\]

(8)

with

\[
\mathbf{t} = \mathbf{T}_4^{lab} \cdot (\mathbf{r}_7 - \mathbf{r}_6) = \mathbf{I}_5 + \mathbf{q}_2
\]

\[
\mathbf{q}_2 = \mathbf{I}_5 + \mathbf{q}_2
\]

(9)

\[
\alpha = q_2 \sin \theta_5 - q_2 \cos \theta_5
\]

\[
\beta = q_2
\]

Once \( \phi_1, \ldots, \phi_5 \) are specified and the new atomic positions 1–6 of the chain are calculated, the remaining dihedrals \( \phi_7 \) and \( \phi_8 \) can be determined from the geometry of the chain. From Eq. (A1) the root equation is therefore

\[
F(\phi_1) = \left[ \mathbf{u}_1^{(1)} \right]^{T} \cdot \mathbf{T}_1 \mathbf{T}_2 \cdots \mathbf{T}_6 - \cos \theta_2 = 0
\]

(A10)

which depends only on \( \phi_1 \) since specification of \( \phi_1 \) allows us to solve for the remaining angles.

References