Method for Efficient Computation of the Density of States in Water-Explicit Biopolymer Simulations on a Lattice†

Bryan A. Patel,‡ Pablo G. Debenedetti,*,‡ and Frank H. Stillinger§

Department of Chemical Engineering, and Department of Chemistry, Princeton University, Princeton, New Jersey 08544

Received: August 2, 2007; In Final Form: October 3, 2007

We present a method for fast computation of the density of states of binary systems. The contributions of each of the components to the density of states can be separated based on the conditional independence of the individual components' degrees of freedom. The conditions establishing independence are the degrees of freedom of the interfacial region between the two components. The separate contributions of the components to the density of states can then be calculated using the Wang-Landau algorithm [Wang, F.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050]. We apply this method to a 2D lattice model of a hydrophobic homopolymer in water that exhibits protein-like cold, pressure, and thermal unfolding. The separate computation of the protein and water density of states contributions is faster and more accurate than the combined simulation of both components and allows for the investigation of larger systems.

1. Introduction

During the past 50 years, computer simulations have become an essential tool for the investigation of the dynamic and equilibrium properties of complex systems. Monte Carlo (MC) methods are widely used to investigate equilibrium behavior, and numerous techniques have been developed to study systems for which the traditional Metropolis algorithm is inadequate.[1] Ergodic sampling is difficult to attain for simulations of high-density or low-temperature systems, which are likely to become trapped in local potential energy minima for a large number of simulation steps. This problem is especially pronounced in simulations of proteins, where the comparatively few native state configurations are separated by a large energy gap from the ensemble of denatured configurations.[2]

Several advanced MC methods have been applied to protein simulations that were developed to promote ergodic sampling by improving the exploration of phase space. These include configurational bias trial moves,[3−6] pruned-enriched Rosenbluth methods,[7−8] parallel tempering,[9,10] multicanonical methods,[11,12] and the Wang−Landau (WL) method.[13−15] The multicanonical and WL methods belong to a class of techniques called flat-histogram methods that are designed to achieve a broad sampling of phase space and to directly calculate free energies. They attempt to produce a uniform distribution of a macroscopic property, such as the potential energy, by sampling each microstate of the system with a probability inversely proportional to the density of states (DOS) of the corresponding energy level. The DOS is not known initially but is instead determined in the course of the simulation, either explicitly in the WL scheme, or implicitly in multicanonical methods. Knowledge of the density of states, Ω, then allows the calculation of the thermodynamics of the system.

The WL method operates by performing a random walk in some system property, often the energy, to sample a large region of phase space and provide an estimate for the DOS through successive refinement at every simulation step.[13,14] The method was originally developed for lattice systems[13] but has been successfully extended to continuum systems.[16,17] The WL method is frequently used to study proteins both on a lattice[15,18] and in continuum space[19−21] because of its ability to efficiently sample a wide range of configurations and energies. It has also been applied to perform random walks in non-thermodynamic variables, quantities other than the energy, volume, or number of particles. These applications include calculation of the density of states as a function of reaction coordinates[22,23] and the end-to-end distance of a polyelectrolyte chain.[24] This flexibility of the method is utilized in the approach presented here to separate the calculations of the protein and water contributions to the density of states in a lattice model.

Separation of the protein and water DOS calculations dramatically reduces the computation time and increases the speed of a simulation of a previously developed lattice model for proteins in explicit water.[25] This approach reproduces the experimentally observed phenomena of cold-, pressure-, and heat-induced protein unfolding using a model of a hydrophobic homopolymer in water. The earlier study showed that a physical treatment of the entropic and enthalpic properties of hydrophobic hydration in a simple protein model was sufficient to recover the qualitative shape of the protein phase diagram. However, the complexity of simulating both protein and water, even in a reduced two-dimensional lattice model, limited the sizes of proteins to be simulated to 20 monomers or fewer.

The structure of the paper is as follows. In section 2, we begin with a review of the lattice model for protein and water used in ref 25 and then discuss the technique for separating the contributions of water and the protein to the DOS. The WL algorithm is reviewed briefly, and implementation features specific to the protein and water model are discussed. In section 3, we then compare the speed and accuracy of the new and the

† Part of the "Giacinto Scoles Festschrift".

* Corresponding author. Address: Department of Chemical Engineering, Princeton University, A419 Engineering Quadrangle, Princeton, NJ 08544. Tel.: (609) 258−5480. Fax: (609) 258−0211.

‡ Department of Chemical Engineering.

§ Department of Chemistry.

10.1021/jp0761970 CCC: $37.00
Published on Web 11/09/2007

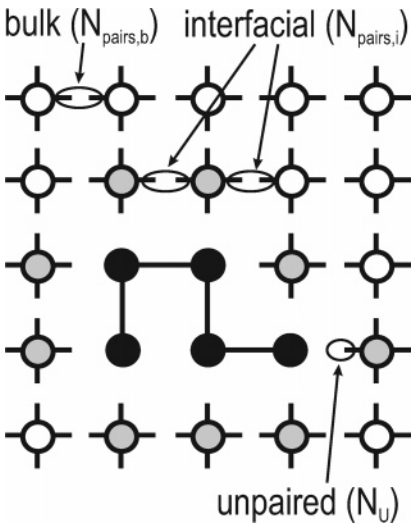

Figure 1. Model protein and water. The black circles represent protein monomers, and the lines connecting them are covalent bonds. The white and gray circles are bulk and interfacial water molecules, respectively. The four arms on each water molecule are the hydrogen bonding arms. The variable $N_{pairs,b}$ counts the number of bulk bonding arm pairs involving two bulk water molecules. Two examples are shown of interfacial bonding arm pairs, which involve at least one interfacial water molecule. $N_{pairs,i}$ is the total number of these interfacial bonding arms. N_u counts the number of unpaired bonding arms associated with a protein nearest neighbor site. Note that this is a depiction of a portion of the simulation box, and in practice a larger box is used to prevent the protein from interacting with itself across the periodic boundary.

original methods. The main conclusions and suggestions for further applications of the method are presented in section 4.

2. Methods

2.1. Model. This model was developed to probe the various properties of protein pressure-temperature stability through the use of a simplified set of interactions representing water−water hydrogen bonding and the hydration of hydrophobic solutes.[25] The protein and water molecules sit on a 2-D lattice, where every site is occupied either by a protein monomer or a water molecule, as shown schematically in Figure 1. The protein is modeled as a chain of attached monomers where adjacent monomers on the protein occupy nearest-neighbor sites on the lattice. Each monomer on the protein is hydrophobic, and the protein has no self-interaction aside from excluded-volume interactions. Its only interaction with the water is through its effect on hydrogen bonding, described below. We are currently studying an extension of this simple model that includes different types of monomers along the protein's backbone (hydrophobic, polar), but here we restrict our discussion to hydrophobic homopolymers. Although we use the term protein throughout, it should be understood that in reality this is a minimal model, albeit one that exhibits complex, protein-like phase behavior.

Water molecules have four hydrogen bonding arms, and each arm can interact with a neighboring lattice site. The variable σ_{ij} denotes the orientation of a bonding arm on water molecule i associated with adjacent site j. A bonding arm can have one of q orientations, and therefore, σ_{ij} can have values between 1 and q. Each of the bonding arms on the same water molecule assumes orientations independently of the other three. A hydrogen bond forms between two neighboring water molecules i and j when their bonding arms are properly oriented, satisfying the condition $|\sigma_{ij} - \sigma_{ji}| \leq \lambda$. λ represents a tolerance for hydrogen bonding or the size of the range of acceptable bonding arm pair orientations. This range differs between bulk water (λ_b) and interfacial water (λ_h). Here interfacial water refers to a situation where either of the hydrogen-bonding water molecules is adjacent to one or more protein monomers. Bulk water molecules are colored white in Figure 1, and interfacial water molecules are colored gray. A smaller range of hydrogen bonding orientations for interfacial water molecules ($\lambda_h < \lambda_b$) constitutes an entropic penalty for hydrogen bonding around the protein monomers. This penalty originates from observations that water molecules forming hydrogen bonds around hydrophobic solutes sample fewer configurations and thus have reduced entropy compared to the bulk.[26,27]

The model also incorporates an energetic distinction between bulk and interfacial hydrogen bonds. Bulk hydrogen bonds form with a strength J, whereas interfacial hydrogen bonds form with strength $J + J_H$. Generally, we use $J_H > 0$ since there is an enthalpic bonus for interfacial hydrogen bonding, originating from the lower enthalpy configurations sampled by solvation shell hydrogen bonds around hydrophobic solutes.[27,28] The complete Hamiltonian of the model is then

$$\mathscr{H} = -JN_{HB} - J_H N_{HB,i} \tag{1}$$

where N_{HB} is the total number of hydrogen bonds (both bulk and interfacial) and $N_{HB,i}$ is the number of interfacial hydrogen bonds.

The lattice is treated as compressible, and the total volume expands uniformly by a value Δv upon the formation of a hydrogen bond. This effect reproduces in the model the lower local density associated with hydrogen bond formation, which is important for recovering water's unusual thermodynamics, such as its density anomalies. An expression for the system volume is then

$$V = V_0 + N_{HB}\Delta v \tag{2}$$

where V_0 is the system volume without hydrogen bonding. $V_0 = v_0 N_{sites}$, where v_0 is the volume per lattice site and N_{sites} is the number of lattice sites. The compressible lattice model of water with independent hydrogen bonds was originally developed by Sastry et al. to study the thermodynamics of supercooled water.[29]

2.2. General Simulation Method. The basis of the present method lies in the observation that the properties of independent subsystems are separable.[30] Given two subsystems 1 and 2 that are weakly interacting (i.e., the subsystems interact sufficiently to maintain thermal equilibrium but not enough that intermolecular interactions are taken into account), the combined energy of the system, E_t, can be written

$$E_t = E_1 + E_2 \tag{3}$$

where E_1 and E_2 are the energies of subsystems 1 and 2, respectively. This additive relation holds true for properties such as the entropy and free energy.[30] The density of states for the combined system Ω_t has the property

$$\Omega_t = \Omega_1 \Omega_2 \tag{4}$$

where Ω_1 and Ω_2 are the DOS's of subsystems 1 and 2, respectively.

Although the above relations hold for any set of independent subsystems, the overwhelming majority of binary models of interest involve strongly interacting subsystems. However, the complexity of a simulation of a binary system could be reduced if the calculations of the properties of the components could be

separated, allowing for independent simulations of each species. Some systems, such as the model presented here, are well-suited for separation because the short-range nature of their intermolecular interactions limits the size of the interacting region between the components. As we discuss below, the separate calculations of protein and water properties can then be performed analytically or through simpler simulations without the need to evaluate the interaction potential.

Separate simulations are possible for conditionally independent subsystems. The mathematical definition of conditional independence states that events A and B are conditionally independent given that event C has occurred, when

$$p(A \cap B|C) = p(A|C)p(B|C) \tag{5}$$

$p(A \cap B|C)$ is the joint probability of observing events A and B given that event C has occurred, $p(A|C)$ is the conditional probability of event A given event C, and $p(B|C)$ is the conditional probability of event B given event C.[31]

The principle of conditional independence can be applied to statistical mechanics through extension of eqs 3 and 4. A binary system where components 1 and 2 interact directly can be described by some potential

$$U(\xi_1, \xi_2) = U_1(\xi_1) + U_2(\xi_2) + U_i(\xi_1, \xi_2) \tag{6}$$

where U is the total potential energy, which is a function of ξ_1 and ξ_2, the degrees of freedom of the two components. U_1 and U_2 are the individual energies of components 1 and 2, dependent only on the degrees of freedom of each component. U_i is the potential describing the interaction between the two components and is a function of both components' degrees of freedom. In principle, term U_i in eq 6 is only a function of a subset of the degrees of freedom of components 1 and 2, those specifying the interface (ξ_i) between the two species. Consider the case of a protein molecule in water, where the protein−water interaction is limited to the first few solvation shell layers of water. The degrees of freedom of bulk water far away from the protein are irrelevant for the computation of the protein−water interaction energy and can be ignored. This type of simplification has been applied in some simulations of protein−water systems where only the first few solvation shell layers of water are treated explicitly and the bulk solvent is represented as a continuum dielectric.[32,33]

Given an appropriate set of the interfacial degrees of freedom, the properties of the two components can be computed on the basis of their internal degrees of freedom separately. In the case of the protein and water example, the protein and water properties are conditionally independent for a given microscopic state of the interface. For conditionally independent subsystems, eq 4 can be rewritten as

$$\Omega_t(\xi_1, \xi_2; \xi_i) = \Omega_1(\xi_1; \xi_i)\Omega_2(\xi_2; \xi_i) \tag{7}$$

where Ω_t is the density of states of the total system. The notation $\Omega_1(\xi_1; \xi_i)$ denotes the density of states of component 1 as a function of its degrees of freedom given a specific set of interfacial degrees of freedom. This equation allows us to relate the simpler quantities of the individual component DOS to the total DOS.

The challenge in implementing the method is to find appropriate interfacial degrees of freedom that can establish independence of the two components. The details and complexity of these degrees of freedom will vary based on the interaction potential and model of interest, but we can identify some guiding principles for applying this method. The method is best suited for lattice models, where the interactions are usually local and limited to nearest- or next-nearest-neighbor sites. The method also works well for solvation studies with a small concentration of solute in a large continuum of solvent. These situations restrict the number of interfacial degrees of freedom that must be tracked in the simulation and offer the greatest opportunity for improved computational efficiency.

2.3. Implementation. Ω is estimated using the WL method,[14] a flat histogram algorithm that iteratively refines an estimate for the DOS. A conventional WL simulation performs a random walk in energy (U) in the range of attainable energies with probability proportional to the reciprocal of the density of states $1/\Omega(U)$. The density of states is not known a priori but is determined in the course of the simulation. The simulation begins in a randomly generated configuration with the DOS estimator set at $\Omega(U) = 1$ for all energy levels. Trial moves from an old configuration (o) to a new configuration (n) at energy levels U_o and U_n, respectively, are accepted with probability

$$p_{acc}(o \rightarrow n) = \min\left[1, \frac{\Omega(U_o)}{\Omega(U_n)}\right] \tag{8}$$

Every time a state with energy U is visited during the simulation, the corresponding bin in the density of states estimate is updated by multiplying the current value by a modification factor f, i.e., $\Omega(U) \rightarrow \Omega(U)f$. The modification factor is usually initialized at $f_0 = e^1 \approx 2.71828$ to allow for efficient sampling of all possible energy levels. During the simulation, a tally of the frequency of visits to each energy level is updated in the form a histogram, $h(U)$, i.e., $h(U) \rightarrow h(U) + 1$. To ensure an even sampling of energy levels, the simulation continues until $h(U)$ is considered sufficiently flat. The random walk should converge to be a perfectly flat histogram after an infinite amount of time, where $h(U)$ has the same value for each energy U, because states with energy U have a degeneracy $\Omega(U)$ but are visited with probability $1/\Omega(U)$. Instead, we allow the simulation to continue until $h(U)$ at each energy level is greater than some percentage of the average value $\langle h(U)\rangle$. When this condition is satisfied, the modification factor is reduced to $f_{new} = \sqrt{f_{old}}$, in order to refine the precision of the density of states estimation process. The energy histogram $h(U)$ is then reset to zero and a new iteration started. The process continues until the histogram is again sufficiently flat and the modification factor is reduced accordingly. This procedure is repeated until f approaches unity to within some designated tolerance.

In the original simulations of the protein and water model,[25] we adapted the WL method and performed a random walk in the two variables in our system Hamiltonian: N_{HB} and $N_{HB,i}$. Any accessible state specified by these variables corresponds to a specific system energy and volume given by eqs 1 and 2. The result of these simulations is the density of states, $\Omega(N_{HB}, N_{HB,i})$, which can then be converted to $\Omega(U, V)$, since U and V are determined once N_{HB} and $N_{HB,i}$ are known. The advantage of performing a random walk in N_{HB} and $N_{HB,i}$, as opposed to U and V, is that the parameters J, J_H, and Δv need not be assigned values during the simulation; this allows us to gather the system thermodynamics from a single simulation for any parameter set.

The new separated simulation method divides the calculation into two parts. The protein contribution to the density of states, Ω_p, is calculated first using a WL simulation of the protein in vacuo. The water contribution to the density of states, Ω_w, is

then computed analytically, as described below. The method takes advantage of the fact that the interactions between the protein and water extend only to the first hydration shell. The interfacial degrees of freedom, ξ_i, that establish conditional independence are the properties of first hydration shell water molecules. These are the number of interfacial bonding arm pairs, $N_{\text{pairs,i}}$, and the number of unpaired bonding arms associated with a nearest-neighbor protein monomer, N_u.

A WL simulation of the protein in vacuo is performed to calculate $\Omega_p(N_{\text{pairs,i}}, N_u)$. This simulation essentially determines the degeneracy of protein configurations that produce a set of solvation shell conditions if the water molecules were present. These variables contain all of the relevant information about the effect of the protein on the water structure and entropy, allowing for the separate computation of the water density of states. Note that $N_{\text{pairs,i}}$ is a measure of the number of interfacial bonding arm pairs, regardless of whether or not a hydrogen bond is formed. Thus, if the first hydration shell were fully hydrogen-bonded, the number of interfacial hydrogen bonds would be equal to the number of interfacial hydrogen-bonding pairs, or $N_{\text{HB,i}} = N_{\text{pairs,i}}$.

Because the orientations of the hydrogen bonding arms on a single water molecule fluctuate independently of each other, each bonding arm pair can be treated separately and independently. This allows for the exact computation of the water orientational density of states, $\Omega_{w,o}$. According to the hydrogen-bonding criteria described above, there are three types of hydrogen bonding arms on water molecules. Examples of each of these three classes of bonding arms are shown in Figure 1. The first type are those bonding arms associated with a hydrogen bonding arm on an adjacent water molecule where both water molecules are not in the first solvation shell and are subject to the hydrogen-bonding criteria of bulk water. The total number of these bulk bonding arm pairs is denoted by the variable $N_{\text{pairs,b}}$, and these pairs can be subdivided into those that have formed hydrogen bonds and those that have not, $N_{\text{HB,b}}$ and $N_{\text{NHB,b}}$, respectively. The second type of bonding arms are those associated with a neighboring water molecule's hydrogen bonding arm where either or both of the water molecules are in the first solvation shell and are subject to the interfacial hydrogen-bonding criteria. The number of interfacial hydrogen bonding and non-hydrogen-bonding pairs are given by the variables $N_{\text{HB,i}}$ and $N_{\text{NHB,i}}$, respectively. The third type are those unpaired bonding arms associated with a protein nearest-neighbor molecule, N_u.

Analytical expressions exist for the density of states of each of these three types of bonding arms. The derivation of the orientational density of states is provided in the Appendix. There, it is shown that the orientational density of states of bulk bonding arm pairs ($\Omega_{w,b}$) for specific values of $N_{\text{HB,b}}$ and $N_{\text{NHB,b}}$ is given by the relation

$$\Omega_{w,b}(N_{\text{HB,b}}, N_{\text{NHB,b}}) = \frac{(N_{\text{pairs,b}})!}{N_{\text{HB,b}}!N_{\text{NHB,b}}!} q^{N_{\text{pairs,b}}}(2\lambda_b + 1)^{N_{\text{HB,b}}}(q - 2\lambda_b - 1)^{N_{\text{NHB,b}}} \quad (9)$$

A similar expression for the orientational density of states of the interfacial bonding arms as a function of $N_{\text{HB,i}}$ and $N_{\text{NHB,i}}$ is given by the formula

$$\Omega_{w,i}(N_{\text{HB,i}}, N_{\text{NHB,i}}) = \frac{(N_{\text{pairs,i}})!}{N_{\text{HB,i}}!N_{\text{NHB,i}}!} q^{N_{\text{pairs,i}}}(2\lambda_h + 1)^{N_{\text{HB,i}}}(q - 2\lambda_h - 1)^{N_{\text{NHB,i}}} \quad (10)$$

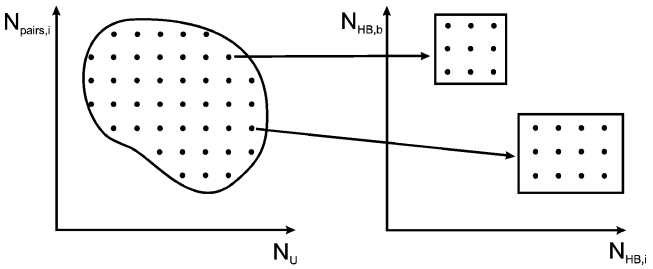

Figure 2. Schematic representation of protein (left) and water (right) states. A point in the protein plane is defined by the variables $N_{\text{pairs,i}}$ and N_u and corresponds to many possible protein microstates. The enclosed region of points on the left represents the limited range of possible values of $N_{\text{pairs,i}}$ and N_u for a given protein size. The irregular shape of the enclosed region on the left shows that not all combinations of $N_{\text{pairs,i}}$ and N_u are possible. Each point in the protein plane is associated with many points in the water plane, with two examples indicated by arrows. Possible water states within these subdomains are defined by the variables $N_{\text{HB,b}}$ and $N_{\text{HB,i}}$ and shown by points on the right. The rectangular shape of these subdomains reflects the fact that all combinations of $N_{\text{HB,b}}$ and $N_{\text{HB,i}}$ within the upper and lower bounds are possible. For visual clarity, the water subdomains are not shown as overlapping, although they do in practice.

Finally, the orientational density of states of the unpaired bonding arms is simply

$$\Omega_{w,u}(N_u) = q^{N_u} \quad (11)$$

For each combination of the variables $N_{\text{pairs,i}}$ and N_u, there are many possible water orientational states, as illustrated in Figure 2. The left panel corresponds to the protein, with the points inside of the enclosed shape representing possible combinations of the variables $N_{\text{pairs,i}}$ and N_u. Each of these points is associated with $\Omega_p(N_{\text{pairs,i}}, N_u)$ protein microstates, and in turn with many possible water orientational states, defined by the variables $N_{\text{HB,b}}$ and $N_{\text{HB,i}}$.

The value of $N_{\text{pairs,i}}$ places an upper bound on the value of $N_{\text{HB,i}}$ in eq 10, since $N_{\text{pairs,i}} = N_{\text{HB,i}} + N_{\text{NHB,i}}$. Thus, if $N_{\text{pairs,i}} = 20$, there are 21 possible interfacial bonding arm states in $\Omega_{w,i}$ since ($N_{\text{HB,i}}$, $N_{\text{NHB,i}}$) can take on values (0,20), (1,19), ..., (19,1), (20,0). For a given value of ($N_{\text{pairs,i}}$, N_u) and system size (i.e., N_w), $N_{\text{pairs,b}}$ can then be calculated from

$$4N_w = 2N_{\text{pairs,b}} + 2N_{\text{pairs,i}} + N_u \quad (12)$$

The value of $N_{\text{pairs,b}}$ places an upper bound on the value of $N_{\text{HB,b}}$ in eq 9, since $N_{\text{pairs,b}} = N_{\text{HB,b}} + N_{\text{NHB,b}}$. Thus, if $N_{\text{pairs,b}} = 50$, there are 51 possible bulk bonding arm states in $\Omega_{w,b}$ defined by ($N_{\text{HB,b}}$, $N_{\text{NHB,b}}$).

The values of $N_{\text{HB,b}}$ and $N_{\text{HB,i}}$ can vary independently of each other because the orientations of bonding arms on an individual water molecule fluctuate independently. The total number of possible water orientational states for a given ($N_{\text{pairs,i}}$, N_u) point is then the product of the number of bulk bonding arm states times the number of interfacial bonding arm states, or $51 \times 21 = 1071$ for the sample case discussed here. This is also shown in Figure 2 by the rectangular shapes of the regions corresponding to possible water orientational states.

With the upper and lower bounds of the variables that define water's DOS determined, the water orientational DOS can be calculated using eqs 9−11. The complete water orientational DOS is a product of these three variables, given by

$$\Omega_{w,o}(N_{\text{HB,b}}, N_{\text{NHB,b}}, N_{\text{HB,i}}, N_{\text{NHB,i}}, N_u) = \Omega_{w,b}(N_{\text{HB,b}}, N_{\text{NHB,b}})\Omega_{w,i}(N_{\text{HB,i}}, N_{\text{NHB,i}})\Omega_{w,u}(N_u) \quad (13)$$

We then compute the total density of states corresponding to each complete specification of water's hydrogen bonding state from the following relation:

$$\Omega_t(N_{HB,b}, N_{NHB,b}, N_{HB,i}, N_{NHB,i}, N_u) = \Omega_p(N_{pairs,i}, N_u) \Omega_{w,o}(N_{HB,b}, N_{NHB,b}, N_{HB,i}, N_{NHB,i}, N_u)\Omega_{w,c} \quad (14)$$

where $\Omega_{w,c}$ is the water configurational DOS. $\Omega_{w,c}$ for a fully occupied lattice where the protein already occupies space is merely the number of different ways of arranging N_w water molecules on N_w lattice sites, or $N_w!$.

Applying eq 14 yields the total density of states for each possible specification of water's hydrogen-bonding state associated with one value of ($N_{pairs,i}$, N_u). Repeating the procedure for all possible ($N_{pairs,i}$, N_u) combinations (see Figure 2) yields the total density of states. To recover the density of states as a function of the independent variables used in the original nonseparable computation, $\Omega(N_{HB}, N_{HB,i})$, we simply sum Ω_t over all possible combinations of the variables determining Ω_t (see eq 14) consistent with the choice (N_{HB}, $N_{HB,i}$).

Extracting protein properties from the simulation data requires converting the density of states into more useful metrics. Since there are fluctuations in both internal energy and volume in the simulation, we reweigh the density of states in the isobaric−isothermal ensemble. Given a pressure P and temperature T, the probability of a state, j, specified by N_{HB} and $N_{HB,i}$ is

$$p_j(P, T) = \frac{\Omega(N_{HB}, N_{HB,i})e^{-\beta U(N_{HB},N_{HB,i})-\beta PV(N_{HB})}}{\Delta(P, T)} \quad (15)$$

where $\beta = 1/k_BT$ and Δ is the isobaric−isothermal partition function. The protein native state is defined as the set of system states where the protein is maximally compact, when it has formed the most possible nearest-neighbor protein−protein contacts. Summing the probabilities of these compact states gives the probability that the protein is folded

$$p_f(P, T) = \sum_{\text{compact states}} p_{N_{HB},N_{HB,i}}(P, T) \quad (16)$$

The change in free energy upon unfolding, ΔG, can then be calculated from the folding probability by using the equilibrium relation from the two-state model of protein folding

$$\Delta G(P, T) = G_{unfolded} - G_{folded} = RT \ln\left[\frac{1 - p_f(P, T)}{p_f(P, T)}\right] \quad (17)$$

The transition between the folded and unfolded states occurs when $\Delta G(P, T) = 0$ or, equivalently, when $p_f(P, T) = 0.5$.

3. Results

To test the accuracy of the proposed method, we compare the density of states generated by both approaches [i.e., nonseparable DOS calculations[25] and separable DOS calculations (this work)] to the exact DOS for small homopolymers. We perform 10 independent runs of homopolymers of size $N_{Monomers} = 4-8$ for each method, calculating the dimensionless entropy, $S = \ln \Omega$, rather than the density of states itself. All simulations were performed with the flat histogram requirement that the bin with the fewest entries in the histogram of visited states has a value of at least 80% of the average number of visits to a binned state, before proceeding to the next iteration. The simulations continued until $\ln f < 10^{-7}$.

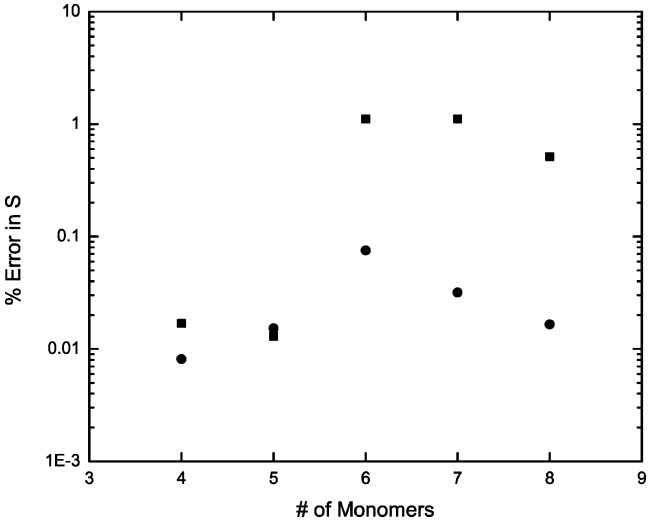

Figure 3. Relative error of the dimensionless total entropy, $S = \ln \Omega_t$, calculated from 10 WL simulations of the combined protein and water system (squares) and 10 simulations of the separated protein in vacuo system with exact enumeration of water orientations (circles). The error is computed relative to the exact protein and water DOS, enumerated by hand.

To determine the exact DOS, the protein component, Ω_p, was enumerated by hand for each homopolymer size, and the water component was calculated using eqs 9−11. We followed the analysis of Shell et al.[34] and calculated the root-mean-square deviation of the corresponding total entropy microstates between each of the two simulation methods and the exact calculation using the relation

$$\text{RMSD} = \sqrt{\langle (S_{exact} - S_{simulation})^2 \rangle} \quad (18)$$

where the notation $\langle\rangle$ indicates an average over all microstates. The percent root-mean-square deviation is then calculated by normalizing the root-mean-square deviation by the range of entropies observed in the exact enumeration, $S_{max} - S_{min}$. This quantity is a measure of both the systematic and statistical error in the simulation methods. In the remainder of the discussion, we refer to this quantity as the percent error or relative error.

The results of these calculations are presented in Figure 3. The two methods show comparable relative error for the smallest homopolymers, the 4- and 5-mers. However, the relative error of the combined DOS method increases dramatically for the larger homopolymers while the separable DOS method shows no significant change. The jump in the error of the combined method between the 5- and 6-mers reflects an increase in the complexity of phase space. The number of configurations available to a 4- or 5-mer on a square lattice is limited to either maximally compact or mostly unfolded states with a small number of translational trial moves separating them. When $N_{Monomers} \geq 6$, a number of partially folded intermediate configurations become possible which are visited less frequently during a random walk, yielding less accurate estimates of the degeneracy of those states. The number and complexity of these partially folded intermediate configurations, and therefore the complexity of phase space, can vary between monomer sizes. This is reflected in the drop in the relative error of both methods from the 7-mer to 8-mer. Even in the case of the 8-mer, the separable DOS method is more than an order of magnitude more accurate than the combined DOS method.

There are two reasons for the improved accuracy of the new method. First, only Ω_p is calculated by the WL simulation,

whereas Ω_w is calculated exactly. Thus, any error remaining in the new method comes from imperfect sampling of protein configurations. The dynamics of the protein simulation are such that the most accurate estimates for its configurational degeneracy are provided for more extended conformations, whereas less accurate estimates are obtained for compact and nearly compact configurations. These rare conformations are less likely to be visited in a random walk and long, high-precision WL simulations are needed to calculate their degeneracy to the same accuracy as extended conformations. When translating the protein configurational DOS into the total density of states, these extended conformations are associated with a larger number of microstates than the compact protein conformations. The minimal error of the less compact configurations is thus emphasized and the error in the total DOS is reduced in the new method.

The in vacuo protein simulations also improve the sampling of protein configurations over that obtained in combined protein and water simulations. The previous method required simulation of the protein chain in a fully occupied lattice of water, where each trial move requires displacement of a water molecule. Although the WL acceptance criteria improves the acceptance of these local trial moves over conventional Boltzmann sampling, the dynamics of the random walk are sluggish enough that the simulation requires long tunneling times between visits to rare configurations. Bachmann and Janke observed a similar difficulty in obtaining good sampling of rare configurations of lattice proteins from WL simulations using local trial moves.[35] In contrast, the new method can explore phase space more effectively because the protein trial moves in the absence of water are accepted more frequently, reducing the number of simulation steps required between visits to these rare configurations. For example, a local translational move involving two monomers is accepted approximately 2% of the time in a protein and water simulation of a 6-mer, whereas it is accepted 15% of the time in a protein in vacuo simulation. Better sampling of low-degeneracy configurations both improves their accuracy and reduces the overall simulation time.

As a further verification of the method, we compare the phase diagrams of a 17-mer calculated from simulations using the previous (nonseparable) and the present (separable DOS) method in Figure 4. The lines shown in the figure demarcate the ranges of dimensionless temperature and pressure where the maximally compact state is stable and indicate where an unfolding transition occurs. Figure 4 shows good agreement between the two methods at low and high temperatures, with some deviation at intermediate temperatures near the point of maximum pressure stability. The difference between the two simulations arises from the improved accuracy in calculation of the cold-denatured state degeneracy with the new method. The high-temperature transition of the protein from its folded conformation to an ensemble of unfolded conformations upon thermal denaturation is essentially the same for both methods, indicating that the simulation estimates for the entropy of these states is very close in both calculations. At temperatures approaching $T = 0$, the two methods show similar predictions for the phase diagram. At $T = 0$, the transition between the folded and cold-denatured state is determined solely by the respective enthalpies, and errors in the simulation estimates have no effect on the transition pressure. However, the different predictions for the cold denaturation portion of the phase diagram result from differences in the prediction of the entropy of the cold-denatured state.

A major advantage of the separate simulation of the protein and water is the dramatic improvement in the speed of the

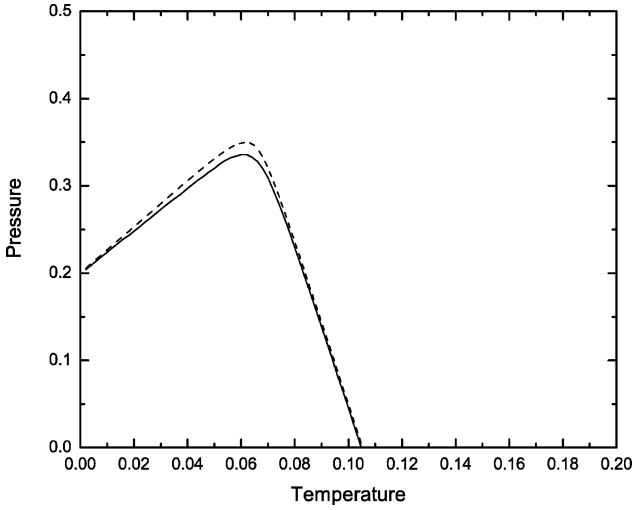

Figure 4. Phase diagram of a 20-mer homopolymer protein calculated using the combined protein and water simulation DOS (solid line) and the separated protein in vacuo simulation (dashed line). Temperature and pressure are presented in dimensionless units, and the model parameters used were $J_H/J = 0.2$, $\Delta v/v_0 = 0.348$, $\lambda_h = 0$, $\lambda_b = 1$, and $q = 30$.

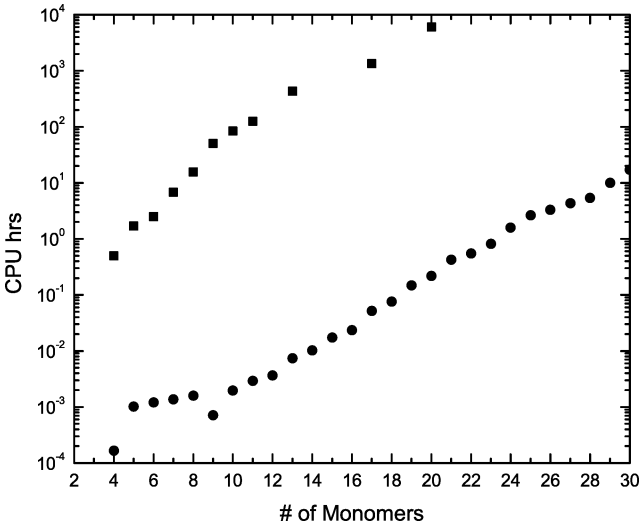

Figure 5. Runtimes for proteins of various sizes for the combined protein and water simulation (squares) and the separated protein in vacuo simulation (circles). The dip in runtime for the 4- and 9-mers relative to the other protein chain lengths is due to the 2-D nature of the lattice, in which these small chains form compact configurations of perfect squares with fewer rare unfolding intermediates. This simplifies the computation of the density of states for these proteins.

simulations. Figure 5 compares the simulation running time with increasing protein size for the previous nonseparable simulation method against the present (separable degrees of freedom) approach. The calculation of the water density of states requires negligible computation time (less than 1 s) and is not included in the runtime shown for the present method. It is clear from Figure 5 that the separate simulation is more than 4 orders of magnitude faster than the combined simulation, largely due to the very pronounced reduction in the size and complexity of the phase space sampled by the new method.

Figure 6 shows the growth in the number of microstates with protein size in the DOS of the previous and present methods. The removal of the bulk water degrees of freedom reduces the number of microstates by more than 2 orders of magnitude, from almost 4000 to 9 in the case of a 6-mer. Furthermore, the size of phase space increases faster with system size for the

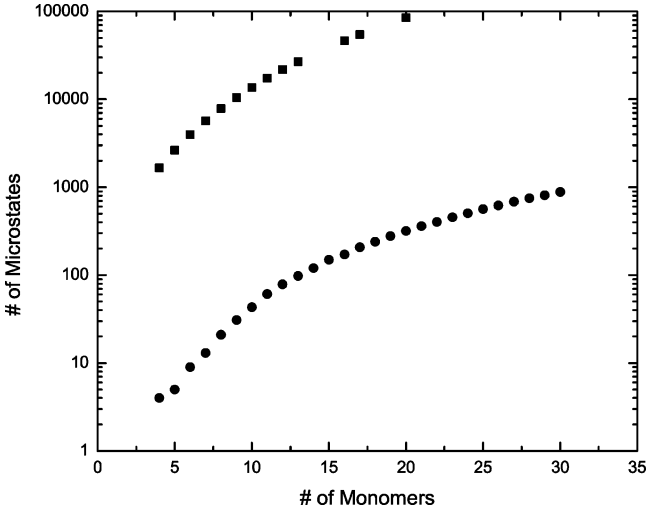

Figure 6. Growth in the number of accessible microstates in phase space with increasing protein chain length for the combined protein and water simulation (squares) and the separated protein in vacuo simulation (circles).

combined simulation. The number of microstates in the combined protein and water simulation is approximately a product of the sizes of the DOS variables N_{HB} and $N_{HB,i}$. N_{HB} is a bulk property and increases as $N_{Monomers}^2$, whereas $N_{HB,i}$ is an interfacial property and therefore increases linearly with $N_{Monomers}$. The number of microstates in the combined simulation DOS increases as $N_{Monomers}^3$. In contrast, the protein configurational DOS is a function of two interfacial properties, $N_{pairs,i}$ and N_u, both of which scale with $N_{Monomers}$. Thus, the number of microstates in the protein in vacuo simulation increases as $N_{Monomers}^2$.

4. Conclusions

The method presented here is a fast and accurate way to simulate binary interacting systems as separate one-component systems. Separate calculation of the protein and water DOSs reduced the error in the computation of Ω, improving the predictions of the protein phase diagram. The method also drastically reduced the time needed for simulation, allowing for larger systems to be to studied. Although our investigation was restricted to serial simulations on one processor and proteins of 30 monomers or less in size, the method can be readily extended to parallel computation for the simulation of proteins larger than 30 monomers. As computational power grows, the method will be scalable to even larger systems with the next generation of computers.

The practical application of the method is limited to systems where the interactions of the two components can be separated, satisfying the requirement of conditional independence. The approach is most suitable for binary lattice models, which generally have local interactions that can be enumerated easily for the set of interfacial conditions. It is also well suited for studies of solvation at infinite dilution, where the interfacial region is limited in size and the degrees of freedom are limited in number.

There are many possible applications of the method, including larger and more complex simulations of a mixture of lattice polymers with structured monomers than have been previously possible.[36] Another possibility is to extend a recent study that examined the effect of a single-site ionic solute on the energy landscape of a dipolar solvent on a 9×9 lattice.[37] Studies of polyelectrolytes on a larger lattice could be performed using the approach presented here. The authors are currently applying the present method to a modified version of the water-explicit protein model that incorporates hydrophobic and polar protein monomers. The method is also readily applicable to three-dimensional lattice models.

Acknowledgment. We thank Scott Shell for many helpful and constructive discussions. P.G.D. gratefully acknowledges the support of the National Science Foundation (Collaborative Research in Chemistry Grant No. CHE0404699) and the U.S. Department of Energy, Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences, Grant No. DE-FG02-87ER13714.

Appendix

The water orientational density of states, $\Omega_{w,o}$ can be divided into three components: the orientational density of states of the bulk hydrogen bonding arm pairs, $\Omega_{w,b}$; the orientational density of states of the interfacial hydrogen bonding arm pairs, $\Omega_{w,i}$; and the orientational density of states of the unpaired hydrogen bonding arms, $\Omega_{w,u}$. These quantities enumerate the number of orientations available to each of these types of bonding arms. Because each of the bonding arms on a water molecule can vary independently of each other, each bonding arm pair can be treated separately and independently in the computation of $\Omega_{w,o}$. Thus, $\Omega_{w,o}$ is specified by the relation

$$\Omega_{w,o} = \Omega_{w,b}\Omega_{w,i}\Omega_{w,u} \tag{19}$$

The number of orientations possible to the bulk bonding arm pairs is a function of the number of bulk bonding arms that are properly oriented for hydrogen bonding, $N_{HB,b}$, and the remainder that do not form hydrogen bonds, $N_{NHB,b}$. The degeneracy of bulk orientations for a microstate specified by $N_{HB,b}$ and $N_{NHB,b}$ is given by

$$\Omega_{w,b}(N_{HB,b}, N_{NHB,b}) = \frac{(N_{pairs,b})!}{N_{HB,b}!N_{NHB,b}!} q^{N_{pairs,b}}(2\lambda_b + 1)^{N_{HB,b}}(q - 2\lambda_b - 1)^{N_{NHB,b}} \tag{20}$$

The first term in the product on the right-hand side of eq 20 is the number of ways of distributing $N_{HB,b}$ bulk hydrogen bonds among $N_{pairs,b}$ bulk hydrogen bonding arm pairs, where $N_{pairs,b} = N_{HB,b} + N_{NHB,b}$. The second term on the right-hand side is the number of orientations available to the first bonding arm in each pair, which can assume any of q possible orientations whether it is hydrogen-bonded or not. The third term is the number of orientations then available to the second bonding arm in each of the bonding arm pairs, given that the first bonding arm orientation is already specified. The bulk hydrogen-bonding criterion states that a hydrogen bond is formed when the orientations satisfy the relation $|\sigma_{ij} - \sigma_{ji}| \leq \lambda_b$. Thus, for $\lambda_b = 1$, if $\sigma_{ij} = 9$, σ_{ji} has 3 available orientations that form hydrogen bonds (i.e., 8, 9, or 10). The final term on the right-hand side is the number of orientations available to the second bonding arm in each of the nonbonding pairs, given that the first bonding arm orientation is already specified. This is just the remainder of orientations $q - (2\lambda_b + 1)$ that would not form hydrogen bonds, according to the criterion discussed above.

The degeneracy of interfacial bonding arm pair orientations, $\Omega_{w,i}$ shows the same form as eq 20, given by

$$\Omega_{w,i}(N_{HB,i}, N_{NHB,i}) = \frac{(N_{pairs,i})!}{N_{HB,i}!N_{NHB,i}!} q^{N_{pairs,i}}(2\lambda_h + 1)^{N_{HB,i}}(q - 2\lambda_h - 1)^{N_{NHB,i}} \quad (21)$$

where $N_{HB,i}$ is the number of interfacial bonding arms properly oriented for hydrogen bonding and $N_{NHB,i}$ is the number of interfacial bonding arms that do not form hydrogen bonds. The terms in the product on the right-hand side are analogous to those in eq 20 for bulk hydrogen bonding arms.

Finally, there are a number of hydrogen bonding arms which are unpaired, and are associated with protein monomers at nearest-neighbor sites. The protein does not interact directly with the water bonding arms, and thus the unpaired bonding arms can assume any of q orientations. The number of orientations available to the unpaired bonding arms is given by

$$\Omega_{w,u}(N_u) = q^{N_u} \quad (22)$$

References and Notes

(1) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*, 2nd ed.; Academic Press: San Diego, 1996.
(2) Shakhnovich, E. I.; Gutin, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7195−7199.
(3) Rosenbluth, M. N.; Rosenbluth, A. W. *J. Chem. Phys.* **1955**, *23*, 356−359.
(4) Siepmann, J. I.; Frenkel, D. *Mol. Phys.* **1992**, *75*, 59−70.
(5) O'Toole, E. M.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **1993**, *98*, 3185−3190.
(6) O'Toole, E. M.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **1992**, *97*, 8644−8652.
(7) Grassberger, P. *Phys. Rev. E* **1997**, *56*, 3682−3693.
(8) Hsu, H. P.; Mehra, V.; Nadler, W.; Grassberger, P. *J. Chem. Phys.* **2003**, *118*, 444−451.
(9) Marinari, E.; Parisi, G. *Europhy. Lett.* **1992**, *19*, 451−458.
(10) Gront, D.; Kolinski, A.; Skolnick, J. *J. Chem. Phys.* **2001**, *115*, 1569−1574.
(11) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9−12.
(12) Bachmann, M.; Janke, W. *J. Chem. Phys.* **2004**, *120*, 6779−6791.
(13) Wang, F. G.; Landau, D. P. *Phys. Rev. E* **2001**, *64*, 056101.
(14) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050−2053.
(15) Lee, L. W.; Wang, J. S. *Phys. Rev. E* **2001**, *64*, 056112.
(16) Shell, M. S.; Debenedetti, P. G.; Panagiotopoulos, A. Z. *Phys. Rev. E* **2002**, *66*, 056703.
(17) Yan, Q. L.; Faller, R.; de Pablo, J. J. *J. Chem. Phys.* **2002**, *116*, 8745−8749.
(18) Marques, M. I.; Borreguero, J. M.; Stanley, H. E.; Dokholyan, N. V. *Phys. Rev. Lett.* **2003**, *91*, 138103.
(19) Rathore, N.; de Pablo, J. J. *J. Chem. Phys.* **2002**, *116*, 7225−7230.
(20) Rathore, N.; Knotts, T. A.; de Pablo, J. J. *J. Chem. Phys.* **2003**, *118*, 4285−4290.
(21) Vorontsov-Velyaminov, P. N.; Volkov, N. A.; Yurchenko, A. A. *J. Phys. A* **2004**, *37*, 1573−1588.
(22) Calvo, F. *Mol. Phys.* **2002**, *100*, 3421−3427.
(23) Nielsen, B.; Jeppesen, C.; Ipsen, J. H. *J. Biol. Phys.* **2006**, *32*, 465−472.
(24) Khan, M. O.; Chan, D. Y. C. *Macromolecules* **2005**, *38*, 3017−3025.
(25) Patel, B. A.; Debenedetti, P. G.; Stillinger, F. H.; Rossky, P. J. *Biophys. J.* in press.
(26) Frank, H. S.; Evans, M. W. *J. Chem. Phys.* **1945**, *13*, 507−532.
(27) Silverstein, K. A. T.; Haymet, A. D. J.; Dill, K. A. *J. Chem. Phys.* **1999**, *111*, 8000−8009.
(28) Rossky, P. J.; Zichi, D. A. *Faraday Symp. Chem. Soc.* **1982**, *17*, 69−78.
(29) Sastry, S.; Debenedetti, P. G.; Sciortino, F.; Stanley, H. E. *Phys. Rev. E* **1996**, *53*, 6144−6154.
(30) Hill, T. L. *An Introduction to Statistical Thermodynamics*; Dover Publications, Inc.: New York, 1960.
(31) Stirzaker, D. *Elementary probability*; Cambridge University Press: New York, 1994.
(32) Im, W.; Berneche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924−2937
(33) Woo, H. J.; Dinner, A. R.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392−6400.
(34) Shell, M. S.; Debenedetti, P. G.; Panagiotopoulos, A. Z. *J. Chem. Phys.* **2003**, *119*, 9406−9411.
(35) Bachmann, M.; Janke, W. *J. Chem. Phys.* **2004**, *120*, 6779−6791.
(36) Buta, D.; Freed, K. F. *J. Chem. Phys.* **2004**, *120*, 6288−6298.
(37) Suzuki, Y.; Tanimura, Y. *J. Chem. Phys.* **2006**, *124*, 124508.