The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method

Shankar Kumar,[1] Djamal Bouzida,[2] Robert H. Swendsen,[2] Peter A. Kollman,[3] and John M. Rosenberg[1]*

[1]*Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260*
[2]*Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213*
[3]*Department of Pharmaceutical Chemistry, University of California-San Francisco, San Francisco, California 94143*

Received 13 February 1992; accepted 28 April 1992

The Weighted Histogram Analysis Method (WHAM), an extension of Ferrenberg and Swendsen's Multiple Histogram Technique, has been applied for the first time on complex biomolecular Hamiltonians. The method is presented here as an extension of the Umbrella Sampling method for free-energy and Potential of Mean Force calculations. This algorithm possesses the following advantages over methods that are currently employed: (1) It provides a built-in estimate of sampling errors thereby yielding objective estimates of the optimal location and length of additional simulations needed to achieve a desired level of precision; (2) it yields the "best" value of free energies by taking into account all the simulations so as to minimize the statistical errors; (3) in addition to optimizing the links between simulations, it also allows multiple overlaps of probability distributions for obtaining better estimates of the free-energy differences. By recasting the Ferrenberg-Swendsen Multiple Histogram equations in a form suitable for molecular mechanics type Hamiltonians, we have demonstrated the feasibility and robustness of this method by applying it to a test problem of the generation of the Potential of Mean Force profile of the pseudorotation phase angle of the sugar ring in deoxyadenosine.

INTRODUCTION

Several methods have been used to calculate the changes in the free energies between interacting molecules and to investigate relative stabilities of the different conformational states of a given molecule with respect to a conformation coordinate of interest. Such calculations are especially important in providing valuable insight into the role of structure-function relationships in biomolecular interactions and in providing a rational basis for the design and modeling of new drugs. However, free-energy calculations for large molecules are computationally demanding, because the entropy that depends on the extent of the phase space of the molecular system cannot generally be extracted from a simple ensemble average of some property of the given system. Hence, new methods for fast, efficient, and accurate determination of free-energy differences are needed. An increase in efficiency can be achieved in two ways: (1) by improving the efficiency of the simulational method itself and (2) by maximizing the amount of information obtained from either Monte Carlo (MC) or Molecular Dynamics (MD) simulations. This article deals with (1) the Single Histogram (SH) method and (2) the Extended Ferrenberg-Swendsen (WHAM) algorithm, which belong to the latter category; the WHAM equations developed here are extensions of the Multiple Histogram equations developed by Ferrenberg and Swendsen.[1-3] The SH and WHAM methods are applicable for both (constant temperature) MD and MC simulations. Methods for increasing the efficiency of the simulational protocol have been discussed elsewhere.[4-6]

We will first describe the nature of the problems that can be treated by these methods. This will be followed by a brief description of the SH and WHAM equations that can be used for biomolecular systems. An outline of the derivation of the WHAM equations will be given in the Appendix. Finally, we will apply these methods to generate the Potential of Mean Force (PMF) profile of the pseudorotation phase angle of the sugar ring in deoxyadenosine with the objective being to demonstrate the feasibility and robustness of the WHAM algorithm when applied to biomolecular systems.

*Author to whom all correspondence should be addressed.

CCC 0192-8651/92/081011-11$04.00

BASIC STRUCTURE OF THE PROBLEM

The problem of calculating free energies can be broadly divided into two classes for computational purposes: (1) those involving the generation of a PMF profile along a coordinate and (2) those involving the computation of free-energy differences as a given molecular system is modified from a standard initial state to a final state. The latter are special cases of the former class. The approaches that have been commonly used so far in the solution of these problems are Free Energy Perturbation (FEP), and Umbrella Sampling methods.[7-15]* In both the Umbrella Sampling and the FEP methods the Hamiltonian $\hat{H}_0(x)$ is replaced by a modified potential, $\hat{H}_{\{\lambda\}}$, of the form

$$\hat{H}_{\{\lambda\}}(x) = \hat{H}_0(x) + \sum_{i=1}^{L} \lambda_i \hat{V}_i(x) = \sum_{i=0}^{L} \lambda_i \hat{V}_i(x) \quad (1)$$

with $\lambda_0 = 1$ and $\hat{V}_0(x)$ defined as being identical to $\hat{H}_0$. Circumflexes over the symbols denote functions.† Here the coordinates of the atoms of the molecule are represented by x; the L functions, $\hat{V}_1(x)$, $\hat{V}_2(x)$, . . . , $\hat{V}_L(x)$, are restraining potentials. The restraining potentials are functions of the molecular coordinates x. The λ_i are coupling parameters. The symbol in braces, $\{\lambda\}$, denotes the set of values $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_L$. Thus $\{0\}$ indicates that all the λ_i ($i = 1, 2, \ldots, L$) have been set to zero; unless stated otherwise λ_0 always takes on the value of unity. The restraining potentials are chosen in such a manner that the sampling distribution is shifted along a coordinate of interest such as a reaction coordinate. Multiple restraining potentials are useful for sampling "long" reaction pathways where separate simulations with different coupling parameters $\{\lambda\}$ are carried out to sample different regions of the reaction path. The reaction coordinate (termed ξ here) will be a function of x. By adjusting the values of λ_i in eq. (1) any region of interest along the coordinate ξ can be preferentially sampled. Free energies (or PMF values) can then be obtained after corrections for the restraining potential; relative free energies can also be obtained as a function of the coupling parameters.‡

In the problem discussed here, ξ is the Pseudorotation Phase Angle[17-19] of the sugar ring in the nucleic acid base deoxyadenosine. The Hamiltonian is written as

$$\hat{H}_\lambda(x) = \hat{H}_0(x) + \lambda \sum_{i=0}^{3} [1.0 + \cos(\nu_i - \alpha_i + \pi)] \quad (2)$$

The ν_i in eq. (2) refer to the usual sugar torsion angles and are restrained to the values α_i. Here, $\alpha_0 = 36.14°$, $\alpha_1 = 337.6°$, $\alpha_2 = 0.0°$, and $\alpha_3 = 22.34°$; $\hat{H}_0(x)$ is the AMBER all-atom force field of Kollman and coworkers.[20,21] The Hamiltonian $\hat{H}_\lambda(x)$ of eq. (2) has only one restraining potential with

$$\hat{V}(x) = \sum_{i=0}^{3} [1.0 + \cos(\nu_i - \alpha_i + \pi)] \quad (3)$$

The α_i in eqs. (2) and (3) have been chosen so as to bias the sampling toward the energetically unfavorable region in the vicinity of the $O_{4'}$-exo conformation. The restraint is on the torsion angles that determine the pseudorotation phase angle and is chosen to enhance sampling in the neighborhood of $\xi = 270°$. The pseudorotation phase angle is not a simple function of the coordinates, thus requiring a complicated restraining potential. Simulations can be carried out with the coupling parameter λ set at various values so as to minimize statistical errors.

The Umbrella Sampling and FEP equations for simulations carried out with multiple restraining potentials as given in eq. (1) are given below primarily to explain the notations used here.

The probability density $P_{\{\lambda\},\beta}(\xi)$ obtained from a simulation with the Hamiltonian $\hat{H}_{\{\lambda\}}$ [as in eq. (1)] can be written as

$$P_{\{\lambda\},\beta}(\xi) = \exp[-\beta W_{\{\lambda\},\beta}(\xi)] = \langle \delta[\xi - \hat{\xi}(x)] \rangle_{\{\lambda\},\beta} \quad (4)$$

The angular brackets denote ensemble averages and the subscripts refer to the values of the coupling parameters λ_i and to the parameter β given by

*We are following terminology currently in use in the field of biomolecular simulations when we refer to methods described here as "Umbrella Sampling." These same methods are sometimes referred to as "Multistage Sampling" because of historical distinctions between the original Umbrella Sampling and Multistage Sampling methods.

†Thus $\hat{V}_i(x)$ denotes the function and V_i a particular *value* the function takes; circumflexes will be used only where ambiguities might arise.

‡A recent method for calculating PMFs along "internal coordinates" of interest is due to Tobias and Brooks.[16] In this method a holonomic constraint is used to fix the coordinate (analogous to the SHAKE algorithm) at a series of values at which the relative free energies (or PMFs) are calculated. This method is well suited to simple reaction coordinates such as a hydrogen bonding distance. However, it is not clear how to apply this method to situations where the coordinate of interest is a complicated function of internal coordinates as in the case of the pseudorotation phase angle that is discussed here; applying constraints to many internal coordinates could lead to improper sampling of conformational space.

$\beta = 1/k_BT$ where k_B is the Boltzmann constant and T is the temperature. $W_{\{\lambda\},\beta}(\xi)$ is the PMF associated with ξ when the simulation is carried out with the coupling parameters $\{\lambda\}$ at temperature T.

If $P_{\{0\},\beta}(\xi)$ is the probability density obtained from an unbiased sampling, i.e., with all the λ_i (except λ_0 which is equal to one) set to zero, then

$$P_{\{0\},\beta}(\xi) = \exp[-\beta W_{\{0\},\beta}(\xi)] \tag{5}$$

or

$$P_{\{0\},\beta}(\xi) = \frac{Z_{\{\lambda\},\beta}}{Z_{\{0\},\beta}} \times \left\langle \delta[\xi - \hat{\xi}(x)] \prod_{i=1}^{L} \exp[\beta\lambda_i \hat{V}_i(x)] \right\rangle_{\{\lambda\},\beta} \tag{6}$$

where Z is the partition function. If we restrict the restraining potential $\hat{V}_i(x)$ to be functions of the coordinate ξ only—that is if

$$\hat{V}_i(x) \equiv \hat{V}_i[\hat{\xi}(x)] \tag{7}$$

then $P_{\{\lambda\},\beta}(\xi)$ will be related to $P_{\{0\},\beta}(\xi)$ by

$$P_{\{\lambda\},\beta}(\xi) = D(\{\lambda\}, \beta) P_{\{0\},\beta}(\xi) \exp\left[-\sum_{k=1}^{L} \lambda_k \beta \hat{V}_k(\xi)\right] \tag{8}$$

and $W_{\{\lambda\},\beta}$ is related to $W_{\{0\},\beta}$ by[8,9]

$$W_{\{0\},\beta}(\xi) = -\sum_{j=1}^{L} \lambda_j \hat{V}_j(\xi) + W_{\{\lambda\},\beta}(\xi) + C(\{\lambda\}, \beta) \tag{9}$$

where the functions $D(\{\lambda\}, \beta)$ and $C(\{\lambda\}, \beta)$ are given by

$$D(\{\lambda\}, \beta) = \frac{Z_{\{0\},\beta}}{Z_{\{\lambda\},\beta}},$$
$$C(\{\lambda\}, \beta) = \beta^{-1} \ln D \tag{10}$$

The equations given above can be extended to situations where the parameter β is also varied. Equation (9) is the form that has been used most often in estimating free-energy differences and generating PMF profiles by using the Umbrella Sampling method. The method can also be used to calculate free-energy differences as a function of any coupling parameter λ_k. The Umbrella Sampling method, for instance, can be used to estimate the free energy of binding between receptor and ligand molecules, when the binding takes place along a suitable path of approach (or a reaction coordinate). By choosing a different set of λ_i for each simulation such that successive simulations sample overlapping regions along ξ, the function $C(\{\lambda\}, \beta)$ in eq. (9) can be determined so as to make $W_{\{0\},\beta}(\xi)$ agree in the regions of overlap.[12-14]

The standard FEP equations can be readily generalized to the case of multiple restraining potentials as follows:

$$\frac{\partial W_{\{\lambda\},\beta}}{\partial \lambda_k} = \frac{\langle \hat{V}_k \delta[\xi - \hat{\xi}(x)] \rangle_{\{\lambda\},\beta}}{\langle \delta[\xi - \hat{\xi}(x)] \rangle_{\{\lambda\},\beta}} - \langle \hat{V}_k \rangle_{\{\lambda\},\beta} \tag{11}$$

The FEP methods are generally used in situations where the Hamiltonian is changed in small steps so that a given molecule can be "mutated" to a desired end state gradually. By calculating the free-energy changes that occur at each step and by finally summing these free-energy changes, the total free-energy change can be obtained. As a typical example, consider a Hamiltonian of the form

$$\hat{H}_\lambda = (1 - \lambda)\hat{H}_i + \lambda \hat{H}_f = \hat{H}_i + \lambda(\hat{H}_f - \hat{H}_i) \tag{12}$$

where $\hat{H}_i$ and $\hat{H}_f$ could be the Hamiltonian for a "wild-type" and mutated biomolecule; here, λ is a coupling parameter and by varying λ slowly from 0 to 1 the system can be taken from its initial state to its desired end state. Equation (12) is a special case of eq. (1). For the special case of eq. (12) the discretized forms of the FEP equations for the free energy A are

$$\beta[A(\lambda = 1) - A(\lambda = 0)] = -\sum_{i=1}^{i=n} \ln\langle \exp(-\beta[\hat{H}_{\lambda_{i+1}} - \hat{H}_{\lambda_i}]) \rangle_{\lambda_i} \tag{13}$$

and

$$A(\lambda = 1) - A(\lambda = 0) = \sum_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial \hat{H}_\lambda}{\partial \lambda} \right\rangle_\lambda \Delta\lambda = \sum_{\lambda=0}^{\lambda=1} \langle \hat{H}_f - \hat{H}_i \rangle_\lambda \Delta\lambda \tag{14}$$

n in eq. (13) is the number of intervals between $\lambda = 0$ and $\lambda = 1$ over which the summation is carried out. Equations (13) and (14) are the basic FEP equations. Sometimes the implementation of eq. (13) has been referred to as the "Windowing" method and that of eq. (14) as the "Integration" method in the literature. The FEP equations do not have an in-built estimate of errors which makes it difficult for estimating statistical errors in the results. The WHAM algorithm does provide for objective estimation of statistical errors [see eq. (22)].

To summarize: PMF profiles and free-energy differences have been calculated thus far generally by using Umbrella Sampling techniques that use eqs. (7), (8), and (9) and by using FEP methods that utilize eqs. (12), (13), and (14).

Normal Mode analyses[7-22] have also been used in the investigation of the relative stabilities of different conformational states of a molecule.[23] However, conformational states of a biomolecule are characterized by transitions across several energy minima and therefore Normal Mode methods can-

not give a reasonable estimate of the entropy of the biomolecule.

The basic problem then is this: What happens to the free energy as some parameter (or set of parameters) is varied? We present below the Single Histogram and Multiple Histogram equations, which we can use to study the behavior of the free energy as some parameter—either a "coupling" parameter λ or the temperature T—is changed. The WHAM equations presented here are essentially those of Ferrenberg and Swendsen,[1-3] but have been extended to the case of molecular mechanics potentials that characterize biomolecules and can readily be applied to situations where free energies and PMFs are needed as a function of the coupling parameter(s) λ_i and/or the temperature T.

We have tested the SH and WHAM equations on the problem of generating the PMF profile of the pseudorotation phase angle of the sugar ring in deoxyadenosine, the main purpose of this study being to test the feasibility and robustness of the histogram equations when applied to molecular mechanics type potentials that characterize biomolecules. Although the AMBER "All-Atom" force-field of Kollman and coworkers was used in this study the efficiency of the method should not depend upon the particular Hamiltonian that is being used. Applications of these methods to larger systems are in progress.*

SINGLE AND MULTIPLE HISTOGRAM METHODS

The partition function $Z_{\{\lambda\},\beta}$ of a system whose Hamiltonian is given by eq. (1) is

$$Z_{\{\lambda\},\beta} = \sum_{\{V\},\xi} \Omega(\{V\}, \xi) \prod_{i=0}^{L} e^{-\lambda_i \beta V_i} \qquad (15)$$

where $\Omega(\{V\}, \xi)$ is a generalized density of states given by

$$\Omega(\{V\}, \xi) = \int dx\, \delta[\xi - \hat{\xi}(x)] \prod_{i=0}^{L} \delta[V_i - \hat{V}_i(x)] \qquad (16)$$

$\Omega(\{V\}, \xi)$ is independent of $\{\lambda\}$ and β. The SH and WHAM methods can be applied when the partition function is of the form given in eqs. (15) and (16).

An outline of the derivation of the SH and WHAM equations is given in the Appendix. In this section, we will first describe how to obtain PMFs and probability densities from a single simulation using SH equations before generalizing to the case of multiple simulations.

Single Histogram Equations

The first description of the SH equations dates back to 1959 and is due to Salsburg, Jacobsen, Fickett, and Wood.[24] We will present the "operational" form of the SH equations as applied to biomolecular systems here. Using these equations, the objective generally is to generate the PMF profile of the coordinate ξ from a single simulation (and hence the term "Single Histogram"). Let us suppose that a simulation was carried out at temperature $T_1 = 1/k_B\beta_1$ with λ_0 set to one and with the restraining potentials appropriately weighted by the coupling parameters $\lambda_1, \lambda_2, \ldots, \lambda_L$ (to enhance sampling in high energy regions). The quantity of interest is then the probability $\tilde{P}_{\beta_2}(\xi)$ that the coordinate ξ would take if a simulation were done with $\lambda_0 = 1$ and all the other coupling parameters set to zero at a temperature $T_2 = 1/k_B\beta_2$. Generally, $T_1 > T_2$ so as to enhance conformational sampling in high energy regions along ξ. By taking the logarithm of the probabilities PMF profiles can be generated. The data is put into "bins" to generate histograms and the "operational" form of the SH equations becomes

$$\tilde{P}_{\beta_2}[\xi \in (\xi_m, \xi_{m+1})] = \frac{\sum_{j=1}^{\eta(m)} \exp\left[(\beta_1 - \beta_2)\tilde{V}_{0,j}^{(m)} + \sum_{i=1}^{L} \lambda_i \beta_1 \tilde{V}_{i,j}^{(m)}\right]}{\sum_{k=1}^{B} \sum_{j=1}^{\eta(k)} \exp\left[(\beta_1 - \beta_2)\tilde{V}_{0,j}^{(k)} + \sum_{i=1}^{L} \lambda_i \beta_1 \tilde{V}_{i,j}^{(k)}\right]} \qquad (17)$$

where the expression now gives the probability that ξ has the value between ξ_m and ξ_{m+1}—the mth bin—at the temperature T_2. $\tilde{V}_{i,j}^{(k)}$ is the value that the restraining potential V_i takes at the jth snapshot of the kth bin. $\eta(k)$ is the total number of data points that the simulation yielded in the kth bin; it is just the value taken on by the histogram at the bin numbered k. B is the total number of bins that the data has been divided into.

Equation (17) can also be expressed in terms of $N_{\{\lambda\},\beta_1}(\{V\}, \xi)$ where $N_{\{\lambda\},\beta_1}(\{V\}, \xi)$ is the value taken by the histogram at $\{V\}$ and ξ during the simulation at temperature $T_1 = 1/k_B\beta_1$ and with the coupling parameters set to $\{\lambda\}$. Again, $\tilde{P}_{\beta_2}(\xi)$ refers to the probability of occurrence of the coordinate ξ during a simulation performed at temperature T_2 with no restraints. In terms of $N_{\{\lambda\},\beta_1}(\{V\}, \xi)$ we have†

*For an interesting account of the history of Histogram techniques see Ferrenberg's thesis.[1]

†Summation over $\{V\}$ as in eq. (18) denotes summation over the possible values of $V_1, V_2, \ldots, V_L$. Similar remarks apply to summation over $\{\lambda\}$.

$$\tilde{P}_{\beta_2}(\xi) = \frac{\sum_{\{V\}} N_{\{\lambda\},\beta_1}(\{V\}, \xi) \exp\left[(\beta_1 - \beta_2)V_0 + \sum_{i=1}^{L} \lambda_i \beta_1 V_i\right]}{\sum_{\{V\},\xi} N_{\{\lambda\},\beta_1}(\{V\}, \xi) \exp\left[(\beta_1 - \beta_2)V_0 + \sum_{i=1}^{L} \lambda_i \beta_1 V_i\right]} \quad (18)$$

WHAM Equations

The WHAM equations are a natural generalization of the SH equations. Simulations are carried out with various sets of coupling parameters to enhance conformational sampling. PMFs are then calculated for the case when a simulation is done with the desired set of coupling parameters at a specified temperature. We will state the main results first and an outline of the derivation of the WHAM equations will be presented in the next section.

Consider R simulations with the ith simulation being carried out at temperature $T_i = 1/k_B\beta_i$ with the coupling parameters in eq. (1) set to $\{\lambda\}_i$*; also, let the number of snapshots taken from the ith simulation be n_i. Then the (unnormalized) probability histogram $P_{\{\lambda\},\beta}(\{V\}, \xi)$ is given by[1-3]†

$$P_{\{\lambda\},\beta}(\{V\}, \xi) = \frac{\sum_{k=1}^{R} N_k(\{V\}, \xi) \exp\left(-\beta \sum_{j=0}^{L} \lambda_j V_j\right)}{\sum_{m=1}^{R} n_m \exp\left(f_m - \beta_m \sum_{j=0}^{L} \lambda_{j,m} V_j\right)} \quad (19)$$

and

$$\exp(-f_j) = \sum_{\{V\},\xi} P_{\{\lambda\}_j,\beta_j}(\{V\}, \xi) \quad (20)$$

where $N_i(\{V\}, \xi)$ is the value taken by the histogram at $\{V\}$ and ξ during the ith simulation, and f_j is the (dimensionless) free energy of the system described by the Hamiltonian of eq. (1) with coupling parameters $\{\lambda\}_j$; $f_j = \beta_j A_j$ where A_j is identical to the (Helmholtz) free energy of the system during the jth simulation. Equations (19) and (20) were derived by minimizing the errors (see Appendix) in the overlapping probability distributions.[1-3] By iterating eqs. (19) and (20) the f_i and, therefore, the free energies, can be determined self-consistently. For the case of a single simulation the WHAM equations reduce to the SH equations except for a normalization factor.

We can compute the f_i directly from the data (to reduce computational errors) by using the following expression:

$$\exp(-f_i) = \sum_{k=1}^{R} \sum_{t=1}^{n_k} \frac{\exp\left[-\beta_i \sum_{j=0}^{L} \lambda_{j,i} V_{j,t}^{(k)}\right]}{\sum_{m=1}^{R} n_m \exp\left[f_m - \beta_m \sum_{j=0}^{L} \lambda_{j,m} V_{j,t}^{(k)}\right]} \quad (21)$$

In this expression $V_{j,t}^{(l)}$ is the value that the restraining potential V_j takes at the tth snapshot of the lth simulation.

One can start with an arbitrary (but not too unreasonable) set of values for the f_i; a good starting point would be to set all the f_i to zero initially. Convergence was generally very fast for the problem discussed here with the number of iterations being less than 10 and no special care was needed for the initial assignment of values for the f_i to accelerate convergence. However, it is quite possible that free-energy calculations for some systems could benefit from acceleration techniques; the interested reader is referred to Ferrenberg's thesis.[1]

The relative error, $\delta\Omega(\{V\}, \xi)/\Omega(\{V\}, \xi)$, (see Appendix) can be shown to be[1-3]

$$\frac{\delta\Omega(\{V\}, \xi)}{\Omega(\{V\}, \xi)} = \left[g^{-1} \sum_{i=1}^{R} N_i(\{V\}, \xi)\right]^{-1/2} \quad (22)$$

Thus, by knowing where the $\delta\Omega/\Omega$ are high more simulations can be done with the appropriate value of the coupling parameters thus reducing the error by increasing the statistics obtained from the simulations. An overall factor g, with $g = 1 + 2\tau$ where τ is an integrated correlation time[25] for the simulations has been included in the equation; since only the relative magnitudes of the quantity $\delta\Omega(\{V\}, \xi)/\Omega(\{V\}, \xi)$ are of interest the quantity g may safely be omitted.

APPLICATION OF THE HISTOGRAM EQUATIONS

We will now demonstrate the use of Single and Multiple Histogram equations by applying them to estimate the PMF of the Pseudorotation Phase Angle of the sugar ring in deoxyadenosine. While this system is small its Hamiltonian contains most of

*$\{\lambda\}_k$ refers to the value of the coupling parameters during the kth simulation, that is, $\{\lambda\}_k$ denotes the set $\{\lambda_1, \lambda_2, \ldots, \lambda_L\}_k$ which is identical to $\{\lambda_{1,k}, \lambda_{2,k}, \ldots, \lambda_{L,k}\}$.

†In the original formulation by Ferrenberg and Swendsen the equations for the probability distributions contained factors, g_i, that depended upon the integrated correlation times of a simulation; these have been omitted here in eq. (19). For biomolecular systems these factors are approximately equal for each simulation and therefore cancel out of eq. (19). In fact, for biomolecular systems, we have ascertained that these factors make negligible difference to the results even if they differed by factors of 9 or 10. The g_i, however, should not be neglected when phase transitions are involved (see the Appendix and next section).

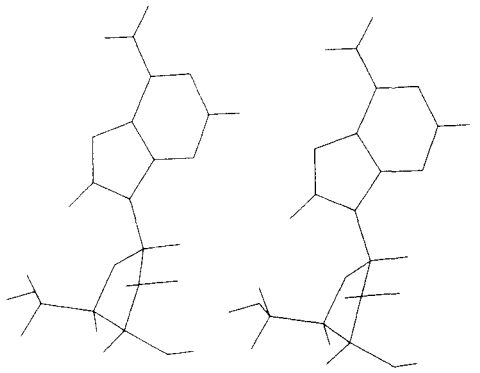

Figure 1. Stereoview of deoxyadenosine.

the complexity of larger molecular systems and thus presents a good test case for the WHAM method.

The system (Fig. 1) consists of 31 atoms; its Hamiltonian is given in eq. (2) with $\hat{H}_0(x)$ being the AMBER "All-Atom" force-field. The restraining potential $\hat{V}_1(x)$ is necessary, for without it very poor statistics are obtained for the pseudorotation phase angle around the $O_{4'}$-exo region ($\xi = 270°$) (Fig. 2). Figure 3 shows the corresponding histogram from a simulation carried out at 298 K with $\lambda = 1.4$; the sampling in the $O_{4'}$-exo region is seen to be better than when there was no restraining potential.

Data from different MD simulations were taken. One simulation was carried out at 250 K; the rest were done at either 298 or 350 K. To eliminate the high frequency bond vibrations, bond lengths were constrained to the values in the AMBER[20] database using SHAKE.[26] All the simulations were done with the restraint given in eq. (2) but with a different value of the coupling parameter λ [see eq. (1)]. The starting coordinates of the molecule were obtained from the AMBER database. Prior to the MD phase of each simulation, the molecular structure was relaxed using the method of Conjugate Gradients[27,28] to an energy gradient of the order of 10^{-2} kcal/Å^2 mol. The MD updates were done with the AMBER program using the leap-frog algorithm[35]; temperature was maintained constant by coupling the system to a heat bath as proposed by Berendsen et al.[30] A distance-dependent dielectric function[21] was used in this study. The details of the MD runs are summarized in Table I.

RESULTS AND DISCUSSION

The first test of the WHAM equations was to see whether or not the calculated free energies were independent of the arbitrarily assigned initial values for the f_j. Since the correlation times were about the same in all the MD runs each g_i (see second footnote on p. 1015 and the Appendix) was set equal to one. Convergence was very fast and was achieved in less than 20 iterations irrespective of the initial values of f_j. The free energies obtained from different starting values of f_i are identical. It can be seen that the "All-Atom" force field of Kollman and coworkers gives a barrier of about 2.5 kcal/mol for a $C_{2'}$-endo ($\xi \approx 144°$) to $C_{3'}$-endo ($\xi \approx 36°$) transition via the $O_{4'}$-exo ($\xi \approx 270°$) region; about 0.5 kcal/mol for the $C_{2'}$-endo → $C_{3'}$-endo

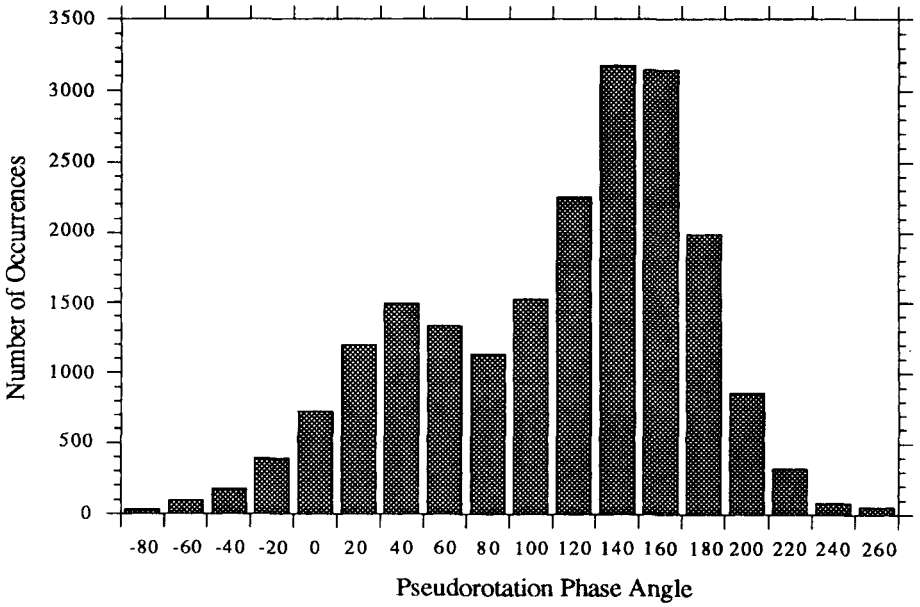

Figure 2. Histogram of the pseudorotation phase angle from simulation 1 (see Table I).

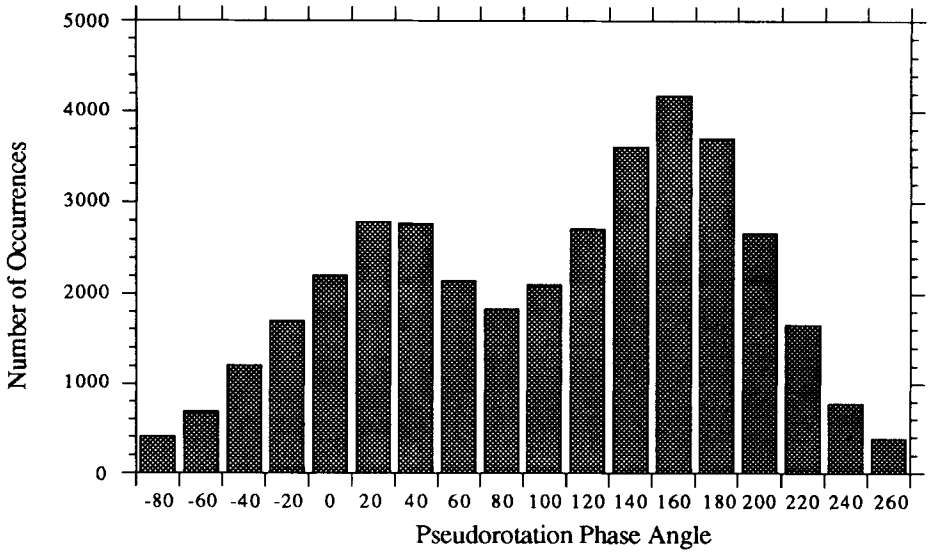

Figure 3. Histogram of the pseudorotation phase angle with $\lambda = 1.4$.

transition via the $O_{4'}$-endo ($\xi \approx 90°$) region; and almost no barrier for the $C_{3'}$-endo → $C_{2'}$-endo transition via the $O_{4'}$-endo region (see Fig. 4).

Due to the difficulty in measuring correlation times accurately it is important that the free-energy differences do not depend strongly on the relative magnitudes of g_i. The values of g_i were now varied over a wide range (from 1–10) (keeping the initial values of f_i equal to zero) to ascertain that the calculated free energies would not differ greatly from the values determined with the g_i set to one. In spite of the wide variation in the ratios of the g_i, we found that the discrepancies in the free energies were negligible. For example, when the WHAM calculations were carried out on simulation 2–7 with $\{g_2 = 3.6, g_3 = 1.8, g_4 = 1.0, g_5 = 1.0, g_6 = 1.0, g_7 = 1.0\}$ and with $\{g_2 = 10.0, g_3 = 1.8, g_4 = 1.0, g_5 = 1.0, g_6 = 1.0, g_7 = 1.0\}$ the maximum discrepancy in the relative values of the free energies f_i was less than 2%. Therefore, differences in the PMF profiles were also negligible. This aspect of the method makes it particularly suitable for free-energy calculations using the Hamiltonian of eq. (1) even if the correlation times tend to vary with the coupling parameter λ; correlation times can be easily determined to within a factor of 2 or 3 (and certainly to within a factor of 10.0). The g_i reflect the weights assigned to each of the histograms; under conditions of biomolecular simulations where phase transitions do not occur, the ratio of the g_i should not differ significantly from one. The last column of Table I gives the approximate correlation time g_i for the simulations.

Qualitative behavior of systems under investigation can also be obtained from histogram techniques. We have used the WHAM equations to obtain the PMF of the pseudorotation phase angle at temperatures 350 and 250 K (Fig. 5); these PMFs were calculated for the case when λ is zero. It can

Table I. Summary of the simulations.

Simulation no.	n	λ_0	λ_1	T(K)	g
1[a]	20,000	1.00	0.00	298	4.0
2	20,000	1.00	0.20	298	4.0
3	20,000	1.00	0.40	298	1.7
4	20,500	1.00	0.50	298	1.1
5	45,000	1.00	1.00	350	1.0
6	48,750	1.00	1.20	350	1.0
7	52,500	1.00	1.40	298	1.1
8	45,000	1.0	1.40	250	1.5
9[a]	60,000	1.0	0.00	298	1.7
10	60,000	1.0	0.50	298	1.7

[a]The difference in the correlation times between runs 1 and 9 is due to the difference in the time step used in the Verlet algorithm[33] and to the difference in the archival rates of the snapshots.

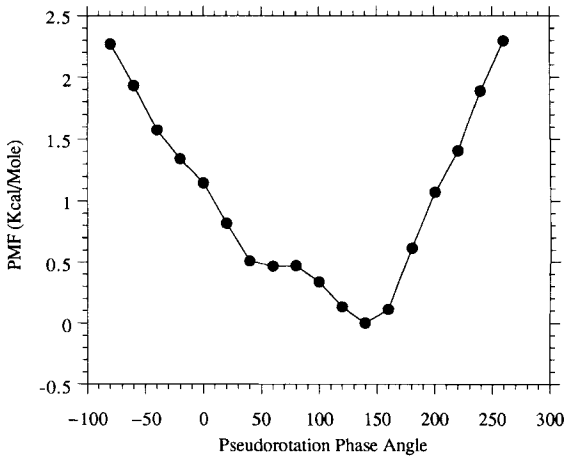

Figure 4. PMF of ξ at 298 K from all simulations.

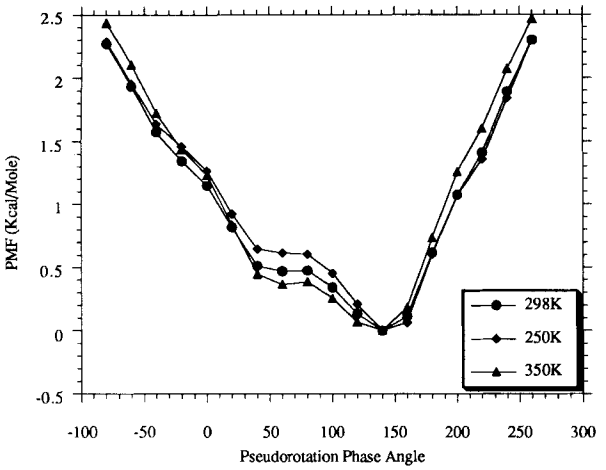

Figure 5. PMF profiles of the pseudorotation phase angle at 250, 298, and 350 K.

be seen that the PMF varies much more with temperature between $\xi = 18°$ and $\xi = 144°$ than around $\xi = -90°$ (the $O_{4'}$-exo region). This suggests that entropy contributes more to the PMF in the former region than the latter. That is expected since the energetics of the $O_{4'}$-exo ($\xi = -90°$) conformation is dominated by steric clash between the $C_{5'}$ hydrogen atoms and the base.[17] Kollman and coworkers[21] found by energy minimization that the difference in energy between the $C_{3'}$-endo and $C_{2'}$-endo region to be ~0.6 kcal/mol. The difference in the PMF values obtained in this work is seen to tend toward this value as the temperature is lowered (Fig. 5). Kollman et al. also report an energy barrier of about 3.7 kcal/mol for the $C_{2'}$-endo to $C_{3'}$-endo transition via the $O_{4'}$-exo region. While the results of Kollman and coworkers are from energy minimizations keeping the sugar puckering amplitude[17-19] fixed, the results obtained here include entropic effects also. This is the main cause for the apparent discrepancies between the two results.

The PMF profile of the pseudorotation phase angle ξ depends on the size, nature, etc. of the molecule comprising the sugar ring. For instance, the $C_{2'}$-endo → $C_{3'}$-endo transition via the $O_{4'}$-exo region for sugar rings in the dodecamer CGCGAATTCGCG is greater than the barrier reported here by about 2 kcal/mol.[31] These will be reported in a future communication. Comparative studies between the WHAM, FEP, and Umbrella Sampling methods will also be reported in a future communication.

Initially only four simulations (1-4 in Table I) were carried out. However, to decrease the relative errors $\delta\Omega/\Omega$ [see eq. (22)] in the "outlying" bins ($\xi \approx 270°$) of the histograms six more simulations (5-10 in Table I) were carried out with increased values for the coupling constant λ. The error propagation from individual bins to the final PMF is not straightforward; however, one can look for errors by breaking all or some of the simulation runs into multiple runs and carrying out the WHAM calculations. We carried out a variety of such calculations and the resulting PMF of ξ was always found to be in agreement with Figure 4.

GENERAL COMPARISONS

As stated in the previous section we will report quantitative comparisons between the FEP, Umbrella Sampling, and WHAM techniques; nevertheless, with the experience to date on the WHAM method certain general comparisons between the methods can be made and will be outlined in this section.

The WHAM method is an extension of the Umbrella Sampling method but it has a number of advantages over the conventional Umbrella Sampling method. The WHAM method, in addition to optimizing the links between simulations, also allows multiple overlaps of probability distributions for obtaining better estimates of the free-energy differences. The older method of obtaining a single distribution function by requiring that the probability distributions agree at some point in the overlap region will fail to yield unique free-energies if three or more distributions are involved in the overlap region.[32] This algorithm provides a built-in estimate of errors that give investigators objective estimates of the optimal location and length of additional simulations to improve the accuracy of their results. With only two simulations, the WHAM method is still better than the conventional Umbrella Sampling, and actually reduces to Bennett's optimal solution for this special case.[33]

Umbrella Sampling methods that rely on eq. (9) cannot use the most general form of the restraining potential (and it is this special form with all but one of the λ_i set to zero that has generally been used so far by researchers). The WHAM method, however, can be used with the most general form of the restraining potential given in eq. (1); it lends itself particularly well to situations where the potential energy and/or the restraining potentials cannot be expressed as a direct function of the parameter(s) of interest. One of the limitations of the Umbrella Sampling method is in the determination of the value that the function $C(\{\lambda\}, \beta)$ so as to make $W_{\{0\},\beta}(\xi)$ agree in the regions of overlap[12-14]; the accuracy of $C(\{\lambda\}, \beta)$ is limited by the statistical errors in the distributions that are "stitched" together. To achieve the same level of accuracy conventional Umbrella Sampling would require much longer simulations than the WHAM method presented here. The WHAM method overcomes this

difficulty by taking into account all the simulations that produce overlapping distributions.

The calculation of free energies and the PMF of reaction or conformation coordinates using the FEP or the conventional Umbrella Sampling methods are computationally expensive. This is a consequence of the convergence problem associated with these computational techniques where many simulations have to be carried out as the Hamiltonian is gradually changed to propel the system along a certain coordinate. When using the FEP method $\Delta\lambda$ in eq. (14) or $(\lambda_{i+1} - \lambda_i)$ in eq. (13) has to be made small to assure convergence and to control the errors of discretization; moreover, errors propagate when connecting distributions at each step. The WHAM method is not a discretization. It uses multiple overlaps that do not have to be as close together as they have to be if the FEP method is used. The WHAM method links the different simulations through the overlapping histograms in an optimal manner. The FEP equations do not have a built-in estimate of errors, which makes it difficult for estimating statistical errors in the results, while the WHAM algorithm does provide for objective estimation of statistical errors [see eq. (22)].

The WHAM equations can also be readily used to generate PMFs and free energies as a function of the coupling parameter(s) λ_i and/or the temperature. This is useful as simulations can be carried out at a range of temperatures to improve conformational sampling and the results extrapolated (or interpolated) to the desired temperature.

This work was supported by grants from the National Institute of General Medicine, NIH (GM25671), the division of Advanced Scientific Computing of the National Science Foundation (ASC-9015310), and the Pittsburgh Supercomputing Center (DMB 90026P). The authors thank Yong Duan of the University of Pittsburgh for helpful suggestions and discussions.

APPENDIX: DERIVATION OF THE WHAM EQUATIONS

Consider R constant temperature simulations with the ith simulation being carried out at temperature T_i and with coupling parameters $\{\lambda\}_i$. Let the number of snapshots of the system taken from the ith simulation be n_i. The objective of the WHAM equations is then to obtain the best estimates of the probability density $P_{\{\lambda\},\beta}(\{V\}, \xi)$ at some $\{\lambda\}$ and β. The WHAM equations also yield the R free energies—$A_1, A_2, \ldots, A_R$—of the system associated with the R simulations.

An estimation of the generalized density of states from the kth simulation, $\Omega_k(\{V\}, \xi)$, can be written as

$$\Omega_k(\{V\}, \xi) = N_k(\{V\}, \xi) \exp\left[\left(\sum_{i=0}^{L} \beta_k \lambda_{i,k} V_i\right) - f_k\right] \quad (k = 1, 2, \ldots, R) \quad (23)$$

where $N_{\{\lambda\}_k,\beta_k}(\{V\}, \xi)$ has been shortened to $N_k(\{V\}, \xi)$ and $f_k = \beta_k A_k$. There will be R such estimates. The best value for the density of states, $\Omega(\{V\}, \xi)$ is written as a weighted sum of the R estimates $\Omega_i(\{V\}, \xi)$ $(i = 1, 2, \ldots, R)$, that is

$$\Omega(\{V\}, \xi) = \sum_{j=1}^{R} \omega_j(\{V\})\Omega_j(\{V\}, \xi) \quad (24)$$

subject to the condition

$$\sum_{j=1}^{R} \omega_j(\{V\}) = 1 \quad (25)$$

The set of ω_i that yield the best estimate of $\Omega(\{V\}, \xi)$ is derived by minimizing the statistical error, $\delta^2\Omega(\{V\}, \xi)$, in the best estimate of $\Omega(\{V\}, \xi)$. If the restraining potentials V_i are functions of the coordinate ξ, then the weights ω_i will depend on ξ through the restraining potentials. Now, the error, $\delta^2\Omega(\{V\}, \xi)$, arises out of the errors, $\delta^2\Omega_1(\{V\}, \xi)$, $\delta^2\Omega_2(\{V\}, \xi)$, $\delta^2\Omega_3(\{V\}, \xi), \ldots, \delta^2\Omega_R(\{V\}, \xi)$, in the R estimates $\Omega_1(\{V\}, \xi)$, $\Omega_2(\{V\}, \xi)$, $\Omega_3(\{V\}, \xi), \ldots, \Omega_R(\{V\}, \xi)$, which in turn depend upon the errors in the histograms, $\delta^2 N_1(\{V\}, \xi)$, $\delta^2 N_2(\{V\}, \xi)$, $\delta^2 N_3(\{V\}, \xi), \ldots, \delta^2 N_R(\{V\}, \xi)$. Equations (26) and (27) summarize this:

$$\delta^2\Omega(\{V\}, \xi) = \sum_{j=1}^{R} \omega_j^2(\{V\})\delta^2\Omega_j(\{V\}, \xi) \quad (26)$$

and

$$\delta^2\Omega_k(\{V\}, \xi) = n_k^{-2} \exp\left[\left(2\sum_{i=0}^{L} \beta_k \lambda_{i,k} V_i\right) - 2f_k\right] \times \delta^2 N_k(\{V\}, \xi) \quad (k = 1, 2, \ldots, R) \quad (27)$$

Following Ferrenberg and Swendsen,[1-3] the error in $N_i(\{V\}, \xi)$ is written as

$$\delta^2 N_i(\{V\}, \xi\} = g_i \overline{N_i(\{V\}, \xi\}} \quad (i = 1, 2, \ldots, R) \quad (28)$$

where the bar indicates the expectation value with respect to all simulations of length n_i and $g_i = 1 + 2\tau_i$ where τ_i is the integrated correlation time of the ith simulation.[29] It should be noted that for biomolecular systems the g_i $(i = 1, 2, \ldots, R)$ are roughly equal to each other and hence cancel each other out of the WHAM equations.

We now make an estimate of the $\overline{N_i(\{V\}, \xi\}}$ as follows:

$$\overline{N_i(\{V\}, \xi)} = n_i\Omega(\{V\}, \xi) \exp\left(f_i - \beta_i \sum_{j=0}^{L} \lambda_{j,i} V_j\right) \quad (i = 1, 2, \ldots, R) \quad (29)$$

From eqs. (26), (27), (28), and (29) we can obtain an expression for $\delta^2\Omega(\{V\}, \xi)$. The error is then minimized by setting the partial derivatives $\partial[\delta^2\Omega(\{V\}, \xi)]/\partial\omega_i$ ($i = 1, 2, \ldots, R$) equal to zero subject to eq. (25). From the resulting expression the WHAM equations

$$P_{\{\lambda\},\beta}(\{V\}, \xi) = \frac{\sum_{k=1}^{R} g_k^{-1} N_k(\{V\}, \xi) \exp\left(-\beta \sum_{j=0}^{L} \lambda_j V_j\right)}{\sum_{m=1}^{R} n_m g_m^{-1} \exp\left(f_m - \beta_m \sum_{j=0}^{L} \lambda_{j,m} V_j\right)} \quad (30)$$

and

$$\exp(-f_j) = \sum_{\{V\},\xi} P_{\{\lambda\}_j,\beta_j}(\{V\}, \xi) \quad (31)$$

can be derived. The density of state $\Omega(\{V\}, \xi)$ can also be determined by setting $\partial[\delta^2\Omega(\{V\}, \xi)]/\partial\omega_i$ to zero and is given by

$$\Omega(\{V\}, \xi) = \frac{\sum_{k=1}^{R} g_k^{-1} N_k(\{V\}, \xi)}{\sum_{m=1}^{R} n_m g_m^{-1} \exp\left(f_m - \beta_m \sum_{j=0}^{L} \lambda_{j,m} V_j\right)} \quad (32)$$

By inserting the expression for $\Omega(\{V\}, \xi)$ into the expression for $\delta^2\Omega(\{V\}, \xi)$ the relative error in $\delta\Omega/\Omega$ can be determined to be

$$\frac{\partial\Omega(\{V\}, \xi)}{\Omega(\{V\}, \xi)} = \left[\sum_{k=1}^{R} g_k^{-1} N_k(\{V\}, \xi)\right]^{-1/2} \quad (33)$$

When the restraining potential is a function of the coordinate ξ only the dimensionality of the histograms reduces from $L + 2$ to 2 and eqs. (30) and (31) simplify to

$$P_{\{\lambda\},\beta}(V_0, \xi) = \frac{\sum_{k=1}^{R} g_k^{-1} N_k(V_0, \xi) \exp\left[-\beta\lambda_0 V_0 - \beta \sum_{j=1}^{L} \lambda_j \hat{V}_j(\xi)\right]}{\sum_{m=1}^{R} n_m g_m^{-1} \exp\left[f_m - \beta_m\lambda_0 V_0 - \beta_m \sum_{j=1}^{L} \lambda_{j,m} \hat{V}_j(\xi)\right]} \quad (34)$$

and

$$\exp(-f_j) = \sum_{V_0,\xi} P_{\{\lambda\}_j,\beta_j}(V_0, \xi) \quad (35)$$

The WHAM equations can easily be generalized to situations where the objective is to generate multidimensional PMF profiles of multiple reaction coordinates.

References

1. A.M. Ferrenberg, PhD Thesis, Carnegie Mellon University, Pittsburgh, PA, 1989.
2. A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.*, **61**, 2635 (1988).
3. A.M. Ferrenberg and R.H. Swendsen, *Phys. Rev. Lett.*, **63**, 1195 (1989).
4. D. Bouzida, S. Kumar, and R.H. Swendsen, in *Computer Simulation Studies in Condensed Matter Physics III, Springer Proceedings in Physics 53*, D.P. Landau, K.K. Mon, and H.-B. Schuttler, Eds., Springer-Verlag, Berlin, 1991.
5. R.H. Swendsen, D. Bouzida, and S. Kumar, Almost Markov Processes in Monte Carlo Simulation of Biological Molecules, Technical Report 91-121-NAMS-27, Carnegie Mellon University, Pittsburgh, PA, 1991.
6. D. Bouzida, S. Kumar, and R.H. Swendsen, *Phys. Rev. A* (to appear).
7. D.A. McQuarrie, *Statistical Mechanics*, Harper & Row, New York, 1976.
8. J.A. McCammon and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1989.
9. C.L. Brooks, M. Karplus, and M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics, Advances in Chemical Physics*, vol. **LXXI**, John Wiley & Sons, New York, 1989.
10. P.A. Bash, U.C. Singh, R. Langridge, and P.A. Kollman, *Science*, **236**, 564 (1987).
11. S.H. Northrup, M.R. Pear, C.-Y. Lee, J.A. McCammon, and J. Karplus, *Proc. Natl. Acad. Sci. USA*, **79**, 4035 (1982).
12. C. Pangali, M. Rao, and B.J. Berne, *J. Chem. Phys.*, **71**, 2975 (1979).
13. G.N. Patey and J.P. Valleau, *J. Chem. Phys.*, **63**, 2334 (1975).
14. J.P. Valleau and D.N. Card, *J. Chem. Phys.*, **57**, 5457 (1972).
15. M. Mezei, P.K. Mehrotra, and D.L. Beveridge, *J. Am. Chem. Soc.*, **107**, 2239 (1985).
16. D.J. Tobias, C.L. Brooks III, *Chem. Phys. Lett.*, **142**, 472 (1987); D.J. Tobias, C.L. Brooks III, *J. Chem. Phys.*, **89**, 5115 (1988).
17. W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984.
18. C. Altona and M. Sundaralingam, *Tetrahedron*, **24**, 13 (1968).
19. D. Cremer and J.A. Pople, *J. Am. Chem. Soc.*, **97**, 1354 (1975).
20. P.K. Weiner and P.A. Kollman, *J. Comp. Chem.*, **2**, 287 (1980).
21. S.J. Weiner, P.A. Kollman, D.T. Nguyen, and D.A. Case, *J. Comp. Chem.*, **7**, 230 (1986).
22. H. Goldstein, *Classical Mechanics*, Addison-Wesley, Reading, MA, 1980.
23. K.K. Irikura, B. Tidor, B.R. Brooks, and M. Karplus, *Science*, **229**, 571 (1985).
24. Z.W. Salsburg, J.D. Jacobson, W. Fickett, and W.W. Wood, *J. Chem. Phys.*, **30**, 65 (1959).
25. H. Muller-Krumbhaar and K. Binder, *J. Stat. Phys.*, **8**, 1 (1973).
26. J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen, *J. Comp. Phys.*, **23**, 327 (1977).
27. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, New York, 1986.
28. D.H.J. Mackay, A.J. Cross, and A.T. Hagler, in *Prediction of Protein Structure and the Principles of Protein Conformation*, G.D. Fasman, Ed., Plenum Press, New York, 1989.
29. R.W. Hockney and J.W. Eastwood, *Computer Simu-*

lation Using Particles, Adam Hilger, Bristol, UK, 1988.
30. H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak, *J. Chem. Phys.*, **81,** 3684 (1984).
31. S. Kumar, PhD thesis, University of Pittsburgh, Pittsburgh, PA, 1990.
32. G.M. Torrie and J.P. Valleau, *Chem. Phys. Lett.*, **28,** 578 (1974).
33. C.H. Bennett, *J. Comp. Phys.*, **22,** 245 (1976).