JCTC Journal of Chemical Theory and Computation

Stability of the Free and Bound Microstates of a Mobile Loop of α-Amylase Obtained from the Absolute Entropy and Free Energy

Srinath Cheluvaraja and Hagai Meirovitch*

Department of Computational Biology, University of Pittsburgh School of Medicine, 3059 BST3, Pittsburgh, Pennsylvania 15260

Received May 16, 2007

Abstract: The hypothetical scanning molecular dynamics (HSMD) method is a relatively new technique for calculating the *absolute* entropy, S, and free energy, F, from a given sample generated by any simulation procedure. Thus, each sample conformation, i, is reconstructed by calculating transition probabilities that their product leads to the probability of i, hence to the entropy. HSMD is an exact method where all interactions are considered, and the only approximation is due to insufficient sampling. In previous studies HSMD (and HS Monte Carlo − HSMC) has been applied very successfully to liquid argon, TIP3P water, self-avoiding walks, and peptides in a α-helix, extended, and hairpin microstates. In this paper HSMD is developed further as applied to the flexible 7-residue surface loop, 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser) of the enzyme porcine pancreatic α-amylase. We are mainly interested in entropy and free energy differences $\Delta S = S_{free} - S_{bound}$ (and $\Delta F = F_{free} - F_{bound}$) between the free and bound microstates of the loop, which are obtained from two *separate* MD samples of these microstates without the need to carry out thermodynamic integration. As for peptides, we find that relatively large systematic errors in S_{free} and S_{bound} (and F_{free} and F_{bound}) are cancelled in ΔS (ΔF) which is thus obtained efficiently with high accuracy, i.e., with a statistical error of 0.1−0.2 kcal/mol (T=300 K) using the AMBER force field and AMBER with the implicit solvation GB/SA. We provide theoretical arguments in support of this cancellation, discuss in detail the problems involved in the computational definition of a microstate in conformational space, suggest potential ways for enhancing efficiency further, and describe the next development where explicit water will replace implicit solvation.

I. Introduction

I.1. The Role of Free Energy in Structural Biology. The theoretical/computational treatment of peptides, proteins, and other biological macromolecules is extremely difficult due to long-range interactions and their rugged potential energy surface, $E(\mathbf{x})$ ($\mathbf{x}$ is the $3N$-dimensional vector of the Cartesian coordinates of the molecule's N atoms). More specifically, this surface is "decorated" by a tremendous number of localized wells and "wider" ones, defined over regions, Ω_m (called microstates)—each consisting of many localized wells (an example for a microstate is the α-helical region of a peptide, see further discussion in sections II.3, II.11, and II.12). A microstate Ω_m, which typically constitutes only a tiny part of the entire conformational space Ω, can be represented by a sample (trajectory) generated by a *local* molecular dynamics (MD)[1,2] simulation starting from a structure that belongs to Ω_m. MD studies have shown that a molecule will visit a localized well only for a very short time [several femtoseconds (fs)] while staying for a much longer time within a microstate,[3,4] meaning that the microstates are of a greater physical significance than the localized wells.

* Corresponding author phone: (412)648-3338; e-mail: hagaim@pitt.edu.

10.1021/ct700116n CCC: $40.75
Published on Web 12/01/2007

A central aim of computational structural biology is to identify the most stable microstates, i.e., those with the largest *conformational* partition function Z_m (or equivalently with the lowest Helmholtz free energy, F_m)

$$F_m = -k_B T \ln Z_m = -k_B T \ln \int_m \exp[-E(\mathbf{x})/k_B T]d\mathbf{x} \quad (1)$$

where k_B is the Boltzmann constant, T is the absolute temperature, and the integration is carried out over the limited microstate Ω_m, rather than over Ω (for simplicity, we shall denote in most cases a microstate Ω_m by m). Thus, the protein folding problem is the notoriously difficult task of identifying the microstate with the global minimum F_m, which practically might be achieved by two challenging stages: (1) identifying an initial set of microstates with expected high stability (e.g., based on an energetic criterion) and (2) calculating their relative populations, $p_m/p_n = Z_m/Z_n$ [$p_m = \exp[-F_m/k_B T]/Z$, where Z is the (cancelled out) partition function of the entire conformational space, Ω], which leads to minimum F_m

$$p_m/p_n = Z_m/Z_n = \exp - [\Delta F_{mn}/k_B T] \quad (2)$$

where $\Delta F_{mn} = F_m - F_n$.

Calculation of relative populations is also required in problems which are less challenging than protein folding, i.e., in cases of *intermediate flexibility*, where a flexible protein segment (e.g., a side chain or a surface loop), a cyclic peptide, or a ligand bound to an enzyme populates significantly several microstates in thermodynamic equilibrium. It is of interest to know whether the conformational change adopted by a loop (a side chain, ligand, etc.) upon binding has been induced by the other protein (induced fit[5,6]) or alternatively the free loop already interconverts among different microstates where one of them is selected upon binding (selected fit[7]). This analysis requires calculating p_m values, which are also needed for a correct analysis of NMR and X-ray data of flexible macromolecules.[8−11] Calculation of F is essential in many other biological processes. Thus, F determines the binding affinities of protein−protein interactions, it is an important factor in enzymatic reactions, electron transfer, and ion transport through membranes, and it leads to the solubilities of small molecules.

I.2. The Difficulty in Calculating the Free Energy. It should first be pointed out that the *absolute* Helmholtz free energy is $F_m = E_m - TS_m$ where S_m is the absolute entropy. Monte Carlo (MC)[12] and MD[1,2] are dynamic methods, which enable one to generate samples of system configurations, $\mathbf{x}$ distributed according to their Boltzmann probability density, $\rho^B(\mathbf{x})$

$$\rho^B(\mathbf{x}) = \exp[-E(\mathbf{x})/k_B T]/Z_m \quad (3)$$

(Z_m is defined over m or the entire conformational space, Ω). With both methods it is straightforward to estimate ensemble averages of quantities that are measured directly from $\mathbf{x}$, such as $E(\mathbf{x})$. On the other hand, to estimate the entropy (defined up to an additive constant)

$$S_m = -k_B \int_m \rho^B(\mathbf{x}) \ln \rho^B(\mathbf{x}) d\mathbf{x} \quad (4)$$

one has to calculate the practically unknown *value* of $\ln \rho^B(\mathbf{x})$ [$\rho^B(\mathbf{x})$ depends not only on $\mathbf{x}$ but also on the entire microstate through Z_m, where Z_m is extremely difficult to calculate directly from the sample]. Thus, the difficulty in calculating F_m stems from the difficulty in calculating S_m. In most cases, however, one is interested in free energy differences ΔF_{mn}, which are somewhat easier to obtain than F_m and F_n themselves.[13−19]

I.3. Calculation of ΔF_{mn} by the Counting Method and Thermodynamic Integration. As said above, even calculation of relative populations is nontrivial. A straightforward way to estimate $p_n/p_m = \exp - [\Delta F_{mn}/k_B T]$ is by a *counting method*, i.e., from a long MD or MC simulation that "covers" both microstates. Thus, $\Delta F_{mn} = -k_B T \ln[(\#m)/(\#n)]$, where $\#m$ ($\#n$) is the population, i.e., the number of times the molecule visited microstate m (n) during the simulation. However, because of high-energy barriers, the transition between microstates at room temperature might require long times, nanoseconds or more even for side-chain rotamers, meaning that reliable sampling of $\#m$ ($\#n$) might become prohibitive. This problem can be alleviated by applying enhanced sampling techniques such as replica exchange[20] or multicanonical methods;[21,22] however, the conformational search capability of these methods is also limited, and microstates of interest might be visited poorly or will not be visited at all. The common analysis is based on projecting MD (MC) trajectories onto a small number of coordinates using principal component analysis or calculating the populations along one or two physically significant reaction coordinates.[23,24]

Differences, ΔF_{mn}, are commonly calculated by thermodynamic integration (TI) over physical quantities such as the energy, temperature, and the specific heat[25,26] as well as nonphysical parameters[13−19,27−34] (free energy perturbation methods and umbrella and histogram analysis methods[35−37] are also included in this category, see ref 19 and references cited therein). This is a robust and highly versatile approach, which is used successfully for calculating the difference in the free energy of binding of two ligands to the active site of an enzyme. However, if the structural variance of m and n is large, then the integration from m to n becomes difficult and in many cases unfeasible. Furthermore, because MC (MD) simulations constitute models for dynamical processes, one would seek to calculate changes in F and S during a relaxation process, by assuming local equilibrium in certain parts along the trajectory; a classic example is simulation of protein folding.[38] Such information cannot be obtained by TI, and it is thus desirable to develop methods that estimate S and F directly from a given trajectory.

I.4. Calculation of the Absolute Entropy. The problems in calculating ΔF_{mn} mentioned above could be remedied to a large extent by developing methods for calculating the *absolute* F_m from a given sample. This would enable one to carry out (only) two *separate* MD simulations of microstates m and n, calculating directly the absolute F_m and F_n and their difference $\Delta F_{mn} = F_m - F_n$, where the TI process or the long runs needed in the counting method are avoided.

A commonly used approach for estimating the absolute S is based on the harmonic approximation and was introduced

to biomolecules by Gō and Scheraga.[39,40] They obtained $S = -(k_B/2)\ln[\mathrm{Det}(\mathrm{Hessian})]$, where Hessian is the matrix of second derivatives of the force field around an energy minimized structure; the quantum mechanical version was applied later for peptides.[41] An important development has been the introduction of the quasi-harmonic (QH) method by Karplus and Kushick,[42] where the Boltzmann probability density of structures defining a microstate is approximated by a multivariate Gaussian. Thus,

$$S_m^{\mathrm{QH}} = (k_B/2)\{N + \ln[(2\pi)^N \mathrm{Det}(\sigma)]\} \quad (5)$$

where the covariance matrix, σ, is obtained from a local MD (MC) sample, and N is (usually) the number of internal coordinates. Clearly, S^{QH} constitutes an upper bound for S since correlations higher than quadratic are neglected; also, anharmonic contributions are ignored, and QH is not suitable for diffusive systems such as water. While QH has been used extensively during the years, a systematic study of its performance has been carried out only recently by Gilson's group[43] who have found that the performance of QH deteriorates significantly in Cartesian coordinates and when applied to more than one microstate.[19]

The absolute F can also be obtained with TI provided that a reference state R with known F_R is available and an efficient integration path $R \rightarrow m$ can be defined. A classic example is the calculation of F of liquid argon or water by integrating the free energy from an ideal gas reference state. However, for nonhomogeneous systems such integration might not be trivial, and in models of peptides and proteins defining adequate reference states is a problem. Differences in free energy can be obtained by Bennett's method and techniques that are derivatives of Bennett's method (for a more complete discussion about methods for calculating the absolute entropy see ref 19).

I.5. Our Methods for Calculating the Absolute *S*. Another approach for calculating the absolute S (F) has been suggested by Meirovitch and has been implemented in two *approximate* techniques of general applicability (i.e., they are not restricted to harmonic conditions), the local states (LS)[44−46] and the hypothetical scanning (HS)[47−49] methods. With both methods each conformation i of an MC(MD) sample is *reconstructed* step-by-step (from nothing) using transition probabilities (TPs), where their product leads to an approximation for the correct Boltzmann probability (eq 3) from which various free energy functionals (e.g., upper and lower bounds) can be defined. Recently, the deterministic approximate calculation of TP(HS) was replaced by a stochastic calculation carried out by MC(MD) simulations, where *all* interactions are taken into account, and from this respect the method [called HSMC(D)] can be viewed as exact;[50] the only approximation involved is due to insufficient MC(MD) sampling. HSMC(D) has unique features: it provides *rigorous* lower and upper bounds for F, which enable one to determine the accuracy from HSMC(D) results alone without the need to know the correct answer. Furthermore, F can be obtained from a very small sample and even from *any single* conformation. HSMC results, which agree within error bars with TI results, were obtained for liquid argon, TIP3P water,[50,51] self-avoiding walks on a square lattice,[52] and peptides.[53,54] Very recently HSMD has been extended to peptides with side chains simulated by MD.[55] We have found that reliable results for *differences*, ΔS_{mn} and ΔF_{mn}, can be obtained with considerable efficiency, ∼100 times faster (in term of computer time) than with MC. These results obtained for decaglycine and $NH_2(Val)_2(Gly)_6(Val)_2CONH_2$ are very encouraging, suggesting that HSMD might become a highly efficient tool for calculating ΔF_{mn} (our main interest) also for more complex systems such as loops.

I.6. A Mobile Loop in Porcine Pancreatic α-Amylase. In this paper we develop HSMD further by applying it to a flexible surface loop of the enzyme porcine pancreatic α-amylase (PPA). α-Amylases (α-1,4-glucan-4-glucanohydrolases, EC 3.2.1.1) are widespread in all three domains of life: Archaea, Bacteria, and Eucarya. These enzymes catalyze the hydrolysis of internal glycosidic bonds in starch and related poly- and oligosaccharides. α-Amylases play a central role in carbohydrate metabolism of microorganisms, plants, and animals. Furthermore, they are widely used in the food and starch processing industry. Many of the enzymatic studies have been carried out with PPA, which serves as a model system.

PPA is a single polypeptide chain of 496 amino acid residues[56−59] consisting of three structural domains, domain A (residues 1−99, 170−404), domain B (residues 100−169), and domain C (residues 405−496). Domain A adopts a $(\beta/\alpha)_8$ barrel structure and contains the three catalytic residues Asp197, Glu233, and Asp300. Domain B occurs as an excursion from domain A and is the structurally least ordered of the three domains; it contains one calcium-binding site. Domain C forms an all-β structure and seems to be an independent domain with unknown function.[56] The active site and the possible roles of associated ions have been well characterized from the crystal structures of several amylases. A deep cleft in domain A is accepted to be the substrate-binding site.[56−62] An essential chloride ion and a calcium ion are located closer to this V-shaped depression and have been suggested to enhance the catalytic activity.[62−65]

While substantial evidence is available for the role of catalytic residues in amylases, very few studies have been carried out on the role of loops surrounding the active site that interact with the substrate. In the crystal structures of the free protein (PPA I[56] and II[59] which differ by two residues) loop 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser) has larger B-factors than the average B-factors of the atoms in the protein. However, in the crystal structures of PPA I complexed with acarbose[57] and PPA II complexed with V-1532[59] the B-factors of this loop are close to the average value in the protein where the loop has moved toward the active site. The maximum main-chain movement is ∼5 Å at His305, which approaches the inhibitor from the solvent side to make a hydrogen bond with a glucose residue. The outcome of this movement is an apparent closure of the surface edge of the cleft.[57] Subsequently, several hypotheses have been put forward with respect to the function of the mobile loop in α-amylases, such as providing assistance in holding the glucose residues in a favorable orientation during

catalysis,[57] or assisting in the transition state,[66] or inducing a trap-release mechanism of substrate and products.[67]

I.7. Extension of HSMD for Loops. This work constitutes the first step in extending HSMC(D) to surface loops in proteins, for which the above short mobile loop (with its small residues) serves as an ideal system. Two MD simulations, starting from the X-ray structures of the free and complexed PPA II, 1pif and 1pig, respectively,[59] will span the corresponding microstates, and the entropy and free energy will be calculated by HSMD. In this initial study the loop is modeled by the AMBER force field[68] alone (where solvation effects are not considered) and by the AMBER and the highly approximate GB/SA implicit solvent.[69] Therefore, the study is focused mainly on (technical) implementation issues of HSMD rather than on the role of the loop in the enzymatic function of PPA; the latter subject will be discussed in future studies where explicit water will be introduced. Still, the present study might indicate whether the transition of the loop to the bound microstate constitutes a selected fit, i.e., whether this microstate is reachable in the free protein. We also discuss in detail various theoretical aspects of HSMD elaborating in particular on the problematic definition of a microstate.

II. Theory and Methodology

II.1. The Protein and Loop Studied. As was pointed out in section I.6 we study the loop of $N = 7$ residues, 304−310 (Gly-His-Gly-Ala-Gly-Gly-Ser), of PPA in two microstates related to the free and bound loop structures. The starting point is the available crystal structures of PPA II, 1pif and 1pig,[59] respectively. Because the structures of these proteins are almost identical, we have chosen to carry out the calculations with the 1pif structure, where the loop structure of 1pig is attached to the 1pif structure by superimposing the structure of 1pig on that of 1pif (the ligand was discarded); this would enable one to study the stability of the bound microstate of the loop in the structure of the free protein, as discussed in the previous section I.7. PPA is a relatively large protein (496 residues), and it would be computationally unfeasible to include all of its atoms in the calculations. Therefore, we consider only a template of 700 atoms (the same atoms for the bound and free structures) that are close to the loop where the rest of the protein's atoms are ignored. The construction of the template is described in detail in a previous publication.[51]

The loops are modeled in vacuum where the potential energy is defined solely by the AMBER96 force field[68] and in solution where the implicit solvation model, GB/SA,[69] is added to this force field. The His residue is protonated in the free and bound states. These systems are simulated by MD using the package TINKER,[70] where the loop is free to move while the template (of 700 atoms) is kept fixed in its X-ray coordinates and only the loop−loop and loop−template interactions are considered, i.e., they define the potential energy. However, HSMD (as well as LS and QH) is implemented naturally in internal coordinates; therefore, the simulated conformations should be transferred from Cartesians to the dihedral angles φ_i, ψ_i, and ω_i (i=1,N=7) and the bond angles $\theta_{i,l}$ (i=1,N, l=1,3) and the side-chain angles χ and the corresponding bond angles. For the present loop we consider two χ angles one of His and one of Ser, while the contribution of the side chain of Ala is ignored; also, because the side chains are much shorter than the backbone and are not restricted by the loop closure condition, the effect of their bond angles on entropy *differences* is expected to be small and is thus ignored (in the next section we argue that to a good approximation bond stretching can be ignored as well). For convenience, these angles (ordered along the backbone) are denoted by α_k, $k = 1,45 = K$.

II.2. Statistical Mechanics of a Loop in Internal Coordinates. The partition function Z_m (eq 1) of a loop is an integration of $\exp -[E(\mathbf{x})/k_BT]$ with respect to the loop's Cartesian coordinates, $\mathbf{x}$ over a microstate m. The change of the variables of integration from $\mathbf{x}$ to internal coordinates, α_k, $k = 1,K$, makes the integral dependent also on a Jacobian, J, which for a linear chain has been shown to be a simple function of the bond angles and bond lengths independent of the dihedral angles.[39,40,42] This transformation is applied under the assumption that the potentials of the bond lengths ("the hard variables") are strong, and, therefore, their average values can be assigned to J, which to a good approximation can be taken out of the integral (however, see a later discussion in this section). For the same reason one can carry out the integration over the bond lengths (assuming that they are not correlations with the α_k), and the remaining integral becomes a function of the K dihedral and bond angles (α_k)[39,40,42] and a Jacobian that depends only on the bond angles; the same discussion also holds for a loop. The partition function becomes

$$Z'_m = DZ_m = D\int_m \exp\{-E([\alpha_k])/k_BT\}d\alpha_1...d\alpha_K \quad (6)$$

where $[\alpha_k] = [\alpha_1,...\alpha_K]$. D is a product of the integral over the bond lengths and their Jacobian J. The Jacobian $[\Pi_j \sin(\theta_j)]$ of the bond angles, θ_j, that should appear under the integral is omitted for simplicity. We *assume* D to be the same (i.e., constant) for different microstates of the same loop, and, therefore, $\ln D$ cancels and can be ignored in calculations of free energy and entropy *differences*. The Boltzmann probability density corresponding to Z_m (eq 6) is

$$\rho^B([\alpha_k]) = \exp\{-E([\alpha_k])/k_BT\}/Z_m \quad (7)$$

and the exact entropy S and exact free energy F (defined up to an additive constant) are

$$S_m = -k_B\int_m \rho^B([\alpha_k])\ln\rho^B([\alpha_k])d\alpha_1....\alpha_K \quad (8)$$

and

$$F_m = \int_m \rho^B([\alpha_k])\{E([\alpha_k]) + k_BT\ln\rho^B([\alpha_k])\}\, d\alpha_1....\alpha_K \quad (9)$$

It should be pointed out that the fluctuation of the *exact* F is zero,[71] because by substituting $\rho^B([\alpha_k])$ (eq 7) inside the curly brackets of eq 9 one obtains $E([\alpha_k]) + k_BT \ln \rho^B([\alpha_k]) = -kT \ln Z_m = F_m$, i.e. the expression in the curly brackets is constant and equal to F_m for any set $[\alpha_k]$ within m. This means that the free energy can be obtained from *any single* conformation if its Boltzmann probability density

is known. However, the fluctuation of an approximate free energy (i.e., which is based on an approximate probability density) is finite, and it is expected to decrease as the approximation improves.[49,71−74] Using the HSMC(D) method, it is possible to estimate the free energy of the system from any single structure.

With MD the bond stretching energy is taken into account in eq 9 (and in free energy functionals defined later), while the corresponding entropy is ignored. The contribution of this energy to the free energy becomes an additive constant if one accepts the assumptions about the stretching energy and the corresponding Jacobian made prior to eq 6. This is a very good approximation; however, if the bond stretching entropy should be considered, we argue in section II.6 that it can be estimated *approximately* within the framework of HSMD by assuming that bond stretching is independent of the other interactions.

II.3. On the Practical Definition of a Microstate Thus far we have defined microstates in general terms and discussed various techniques for calculating their populations, p_m (or the ratios p_m/p_n); however, such calculations cannot be carried out without first establishing a *practical* definition of a microstate, which is not straightforward. Therefore, before discussing the theory further, we elaborate about this important issue that has been ignored to a large extent in the literature but has been given considerable thought by us over the course of years.[9,45,46,74−77] For simplicity we consider (for this discussion) an *N*-residue peptide in a helical microstate Ω_h with constant bond lengths and bond angles (ω_i=180°) meaning that its backbone conformation is solely defined by the angles, φ_i and ψ_i (i=1,N); in Ω_h these angles are expected to vary within relatively small ranges $\Delta\varphi_i$ and $\Delta\psi_i$ around $\varphi_i = -60°$ and $\psi_i = -50°$ (we ignore for a moment the side chains). However, if N is not too small, the correct limits of Ω_h in terms of $[\varphi_i,\psi_i]$ are unknown even for this simplified model because the strongly correlated angles define a complicated narrow "pipe" within the region, $\Delta\varphi_1 x \Delta\psi_1 x \Delta\varphi_2 x \Delta\psi_2 \cdots\cdots \Delta\varphi_N x \Delta\psi_N$. Obviously, these correlations are taken into account by an exact simulation method, and, thus, in practice, Ω_h can be defined (or more correctly, represented) by a *local* MC (MD) sample of conformations initiated from an α-helical structure (as mentioned in section I.1).

However, this definition should be used with caution. Thus, a short simulation will span only a small part of Ω_h, and this part will grow constantly as the simulation continues; correspondingly, the calculated average potential energy, E_h, and the entropy, S_h (obtained by any method), will both increase, and the free energy, F_h, is expected to change as well. As the simulation time is increased further, side-chain dihedrals will "jump" to different rotamers, which according to our definition should also be included within Ω_h; for a long enough simulation the peptide is expected to "leave" the α-helical region moving to a different microstate. Thus, *in practice*, the microstate size and the corresponding thermodynamic quantities depend on the simulation time. In some cases, one can better define Ω_h by discarding structures with dihedral angles beyond predefined $\Delta\varphi_i$ and $\Delta\psi_i$ values or structures that do not satisfy a certain number of hydrogen bonds; one can also apply energetic restraints where their bias should later be removed. However, these restrictions are somewhat arbitrary and are difficult to apply for calculating the differences ΔF_{mn} and ΔS_{mn} between microstates m and n, *which is our main interest*. Therefore, in practice there is always some arbitrariness in the definition of a microstate, which affects the calculated averages. This arbitrariness is severe with some methods and can be controlled (minimized) by others, as is discussed in sections II.9 and II.10.

II.4. Exact Scanning Procedure. The HS, LS, and HSMC(D) methods are based on the ideas of the *exact* scanning method, which is a step-by-step construction procedure for a peptide.[78,79] For simplicity this construction is described for an *N*-residue polyglycine molecule (with dihedral and bond angles denoted α_k, $1 \le \alpha_k \le 6N=K$) in a microstate m. Thus, starting from nothing, a conformation of this molecule is built by defining the angles α_k step-by-step with transition probabilities (TPs) and adding the related atoms;[79] for example, the angle φ determines the coordinates of the two hydrogens connected to C^α, while the bond angle $N-C^\alpha-C'$ determines the position of C'. Thus, at step k, $k-1$ angles $\alpha_1, \cdots, \alpha_{k-1}$ have already been determined; these angles and the related structure (the past) are kept constant, and α_k is defined with the *exact* TP density $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$

$$\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) = Z_{\text{future}}(\alpha_k \cdots \alpha_1)/[Z_{\text{future}}(\alpha_{k-1} \cdots \alpha_1)] \quad (10)$$

where $Z_{\text{future}}(\alpha_k \cdots \alpha_1)$ is a future partition function defined over m by integrating over the future conformations defined by $\alpha_{k+1} \cdots d\alpha_K$ (within m) where the past angles, $\alpha_1 \cdots \alpha_k$ (and their corresponding atoms), are held fixed

$$Z_{\text{future}}(\alpha_k \cdots \alpha_1) = \int_m \exp - [(E(\alpha_K, \cdots, \alpha_1))/k_B T] d\alpha_{k+1} \cdots d\alpha_K \quad (11)$$

For simplicity, from now on we shall omit in most cases the subscript m from the thermodynamic functions. The product of the TPs (eq 10) leads to the probability density of the entire conformation (eq 7)

$$\rho^B(\alpha_K, \cdots, \alpha_1) = \prod_{k=1}^{K} \rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \quad (12)$$

This construction procedure is not feasible for a large molecule because the scanning range grows exponentially and the limits of the microstate m are practically unknown, as discussed in section II.3 (for a practical use of this method see ref 79). However, the exact scanning method constitutes the basis for HS as well as for the much less restricted HSMC(D) and LS methods. The exact scanning method is applicable to a peptide (loop) with side chains,[55] where for a loop all the backbone future conformations should also satisfy the loop closure condition.

The *exact* scanning method is equivalent to any other exact simulation technique (in particular MC and MD) in the sense that large samples generated by such methods lead to the same averages and fluctuations. Therefore, one can assume that a given MC or MD sample has rather been generated

by the exact scanning method, which enables one to reconstruct each conformation i by calculating the TP densities that *hypothetically* were used to create it step-by-step. With HSMC(D) (unlike HS) the *entire* future is considered in the reconstruction process, and in this respect HSMC(D) can be considered to be exact.

II.5. The HSMC(D) Method. The theory is described for HSMD (which is more efficient and practical than HSMC[55]) as applied (for simplicity) to an N-residue polyglycine molecule. One starts by generating an MD sample of microstate m; the conformations are then represented in terms of the dihedral and bond angles α_k, $1 \le \alpha_k \le 6N-K$, and the variability range $\Delta\alpha_k$ is calculated

$$\Delta\alpha_k = \alpha_k(\max) - \alpha_k(\min) \tag{13}$$

where $\alpha_k(\max)$ and $\alpha_k(\min)$ are the maximum and minimum values of α_k found in the sample, respectively. $\Delta\alpha_k$, $\alpha_k(\max)$, and $\alpha_k(\min)$ enable one to verify that the sample spans correctly the microstate m.

As pointed out in section II.4, with our approach a sample conformation i is reconstructed step-by-step by calculating the TP density of each α_k (eq 10) from the future partition functions Z_{future} (eq 11). However, a deterministic integration of Z_{future} based on the *entire* future (within the limits of m) is difficult and becomes impractical for a large peptide where m is unknown (see section II.3). The idea of HSMD is to obtain the TPs (eq 10) by carrying out MD simulations of the future part of the chain rather than by evaluating the integrals defining Z_{future} (eq 11) in a deterministic way. Thus, at reconstruction step k of conformation i the TP density, $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$, is calculated from an MD sample of n_f conformations (generated in Cartesian coordinates), where the *entire* future of the peptide is moved (i.e., the atoms defined by $\alpha_k, \cdots, \alpha_K$) while the past (the atoms defined by $\alpha_1, \cdots, \alpha_{k-1}$) are kept fixed at their values in conformation i. A small segment (bin) $\delta\alpha_k$ is centered at $\alpha_k(i)$, and the number of visits of the future chain to this bin during the simulation, n_{visit}, is calculated; one obtains

$$\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \approx \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1) = n_{\text{visit}}/[n_f \delta\alpha_k] \tag{14}$$

where $\rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ becomes exact for very large n_f ($n_f \rightarrow \infty$) and a very small bin ($\delta\alpha_k \rightarrow 0$). This means that in practice $\rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ will be somewhat approximate due to insufficient future sampling (finite n_f), a relatively large bin size $\delta\alpha_k$, an imperfect random number generator, etc. Because this TP is also applicable to HSMC, we denote it (and functions derived from it) with 'HS' (rather than 'HSMD'). Notice that with HSMD the future conformations generated by MD at each step k remain in general within the limits of m, which is represented by the analyzed MD sample. The corresponding probability density is

$$\rho^{\text{HS}}(\alpha_K, \cdots, \alpha_1) = \prod_{k=1}^{K} \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1) \tag{15}$$

$\rho^{\text{HS}}([\alpha_k])$ defines approximate entropy and free energy functionals, S^{A} and F^{A}, which can be shown using Jensen's inequality to constitute *rigorous* upper and lower bounds, respectively[50]

$$S^{\text{A}} = -k_{\text{B}} \int_m \rho^{\text{B}}([\alpha_k]) \ln \rho^{\text{HS}}([\alpha_k]) d\alpha_1 \cdots \alpha_K \tag{16}$$

$$F^{\text{A}} = \langle E \rangle - TS^{\text{A}} = \langle E \rangle + k_{\text{B}}T \int_m \rho^{\text{B}}([\alpha_k]) \ln \rho^{\text{HS}}([\alpha_k]) d\alpha_1 \cdots \alpha_K \tag{17}$$

where $<E>$ is the Boltzmann average of the potential (force field) energy, estimated from the MD sample, and ρ^{B} (eq 7) is the Boltzmann probability density with which the sample has been generated. S^{A} is estimated from a Boltzmann sample of size n by the arithmetic average, $\overline{S^{\text{A}}}$

$$\overline{S^{\text{A}}} = \frac{1}{n} \sum_{t=1}^{n} \ln \rho_t^{\text{HS}} \tag{18}$$

As discussed in section II.2, the fluctuation (standard deviation) of the correct free energy (eq 9) is zero, while the approximate F^{A} has finite fluctuation, σ_{A} (estimated by its arithmetic average, $\overline{\sigma_{\text{A}}}$), which is expected to decrease as the approximation improves (i.e., as n_f increases and/or $\delta\alpha_k$ decreases)[49,71-74]

$$\overline{\sigma_{\text{A}}} = \left[\frac{1}{n} \sum_{t=1}^{n} [\bar{F}^{\text{A}} - E_t - k_{\text{B}}T \ln \rho_t^{\text{HS}}]^2\right]^{1/2} \tag{19}$$

While (for simplicity) in the theory above only a single angle is reconstructed at each step k, in practice a pair of angles is treated simultaneously, where each pair consists of a dihedral angle and its successive bond angle (e.g., φ and the bond angle $N-C^{\alpha}-C'$). Thus, at each step both α_k and α_{k+1} are considered, and n_{visit} is increased by 1 only if α_k and α_{k+1} are located within the limits of $\delta\alpha_k$ and $\delta\alpha_{k+1}$, respectively. The HSMD process described above for polyglycine is also applicable to a side chain and a loop, where the reconstruction process of the latter starts from the first residue (which is connected to one end of the template), and the future chains are always connected (by the force field) to the second end of the template. Clearly, the conformational freedom of the future chains decreases as step k increases.

It should be pointed out again that in the case of HSMD the dependence of F^{A} (eq 17) on the bond stretching energy is only through $<E>$, while this interaction is ignored in S^{A}. However, under the assumptions leading to eq 6 this is not expected to affect differences in free energy which are our main interest. Still, if one seeks to include the bond stretching entropy, one can use a transition probability density, $\rho(a_k)$, similar to eq 14 for the bond length a_k which corresponds to the pair of atoms k and $k+1$; considering the Jacobian, one obtains $\rho(a_k) \approx n_{\text{visit}}(a_k)/[n_f 3^{-1} \delta(a_k^3)]$, where δa_k is small compared to a_k and $n_{\text{visit}}(a_k)$ is the number of visits to a_k. In this *approximation* the bond stretching is independent of the other interactions and thus $\rho_{\text{TP}}^{\text{HS}} = \rho^{\text{HS}}(\alpha_k|\alpha_{k-1} \cdots \alpha_1)\rho(a_k)$. Both probability densities can be calculated simultaneously, which in practice would not increase computer time.

II.6. The Reconstruction Procedure with HSMD. The HSMD reconstruction procedure needs further discussions.

Thus, the MD simulation of the future chain at step k starts from the reconstructed conformation i, and every g fs the current conformation is considered, where the n_{init} initial considered conformations are discarded for equilibration. The next n_f (considered) future conformations are represented in internal coordinates, and their contribution to n_{visit} (eq 14) is calculated. An essential issue is how to guarantee an adequate coverage of microstate m, i.e., that the future chains will span its entire region (in particular the side-chain rotamers) while avoiding their "overflow" to neighbor microstates, conditions that will occur for a too small and a too large n_f, respectively. To be able to control the extent of coverage of m the following procedure has been applied: n_f has been divided into several (j) shorter repetitive procedures ("units"), each based on $n'_f < n_f$ conformations where $n_f = jn'_f$, and each unit starts from the reconstructed structure i with a different set of velocities followed by equilibration of size, n_{init}; obviously, one would seek to determine the minimal values for n'_f, j, and n_{init}, which would keep the future chains within m while allowing its adequate sampling. A similar procedure was first suggested by Brady and Karplus[80−82] within the framework of the QH method and was also used in implementations of the LS method to peptides.[9,75]

To estimate the extent of coverage of the reconstructed samples of the future chains one can generate a sample of the *entire* peptide (or loop) in the same way it is generated in the reconstruction process. Thus, starting from conformation i and discarding n_{init} conformations for equilibration, the sample can be of size n'_f (using g=10 fs) or of j consecutive samples of size n'_f, each starting from i with a different set of velocities. The $\Delta\alpha_k$ values (eq 13) of the dihedral angles (of both backbone and side chains) of this sample are then calculated and compared to the corresponding values obtained for the studied sample. The $\Delta\alpha_k$ values of the studied sample can also be compared with $\Delta\alpha_k$ values calculated during the reconstruction process itself for randomly chosen one or two conformations. These measures enable one to optimize the values of n_{init}, n'_f, and j. It should be pointed out, however, that in general we are interested in an entropy difference, $\Delta S^{\text{A}}_{m,n}$, between two microstates, where (as discussed later) the set n_{init}, n'_f, and j should be optimized simultaneously for both microstates; the result for $\Delta S^{\text{A}}_{m,n}$ is considered reliable if it is found to be stable for a large range of the parameters n_{init}, n'_f, and j. From now on we shall replace in most cases n'_f by the word unit.

II.7. Upper Bound and Exact Expressions for the Free Energy. In addition to $F^{\text{A}}(\rho^{\text{HS}}([\alpha_k]))$ (eq 17), which in practice is a lower bound, one can define an upper bound functional denoted F^{B} [47]

$$F^{\text{B}} = \frac{\int_m \rho^{\text{B}}[\rho^{\text{HS}} \exp[E/k_{\text{B}}T](E + k_{\text{B}}T \ln \rho^{\text{HS}})]d\alpha_1 \cdots d\alpha_K}{\int_m \rho^{\text{B}}[\rho^{\text{HS}} \exp[E/k_{\text{B}}T]]d\alpha_1 \cdots d\alpha_K} \quad (20)$$

Notice that (unlike F^{A}) the statistical reliability of estimating F^{B} decreases sharply with increasing system size. The inequalities $F^{\text{A}} \le F \le F^{\text{B}}$ will hold provided that the assumptions leading to eq 6 are valid. In this case F^{B} (like F^{A}) is increased by an additive constant (contributed by the bond stretching energy) which is cancelled in free energy differences of microstates. However, if deviations from these assumptions occur, F^{B} will be affected more significantly than F^{A} because $E/k_{\text{B}}T + \ln \rho^{\text{HS}}$ is exponiated in the numerator and denominator of eq 20; thus, to observe $F \le F^{\text{B}}$ one might need to consider the bond-stretching entropy as well (see discussions in refs 46, 50, and 55).

As shown for fluids in ref 50, an *exact* expression for F, denoted F^{D}, is[55]

$$F^{\text{D}} = k_{\text{B}}T\ln\left(\frac{1}{Z_m}\right) = k_{\text{B}}T\ln[\int_m \rho^{\text{B}} \exp[F^{\text{HS}}/k_{\text{B}}T][d\alpha_k]] \quad (21)$$

where $[d\alpha_k] = d\alpha_1 \cdots d\alpha_K$ and $F^{\text{HS}}/k_{\text{B}}T = E([\alpha_k])/k_{\text{B}}T + \ln \rho^{\text{HS}}([\alpha_k])$. The above discussion for F^{B} also applies to F^{D}, where its estimation is statistically more reliable than that of F^{B} which is defined as a ratio of two summations similar to that defining F^{D}.

II.8. The Local States (LS) Method. With the LS method[44−46] (applied to an N-residue polyglycine with $6N = K$ backbone angles, α_k) the ranges $\Delta\alpha_k$ (eq 13) are divided into l equal segments, where l is the discretization parameter. These segments are denoted by ν_k (ν_k=1,l), where an angle α_k is represented by the segment ν_k to which it belongs, and a conformation i is expressed by the corresponding vector of segments $[\nu_1(i), \nu_2(i), ...,\nu_K(i)]$. $\rho(\alpha_k|\alpha_{k-1} \cdots \alpha_1)$ can be estimated by $n(\nu_k, \cdots ,\nu_1)/\{n(\nu_{k-1}, \cdots ,\nu_1)[\Delta\alpha_k/l]\}$, where $n(\nu_k, \cdots ,\nu_1)$ is the number of times the *local state* [i.e., the vector $(\nu_k, \cdots ,\nu_1)$] appears in the sample. However, in practice, one uses smaller local states $(\nu_k,\nu_{k-1},...,\nu_{k-b})$ consisting of ν_k and its b preceding angles, where b is the correlation parameter. $n(\nu_k,\nu_{k-1},...,\nu_{k-b})$ lead to a set of transition probabilities $p(\nu_k|\nu_{k-1},..., \nu_{k-b})$ and *approximate* probability density, $\rho_i(b,l) = \prod_{k=1}^{K} p(\nu_k|\nu_{k-1},...,\nu_{k-b})/(\Delta\alpha_k/l)$, the larger are b and l the better the approximation (for enough statistics). The $\rho_i(b,l)$ lead to *rigorous* upper and lower bounds, S^{A} (eqs 16 and 18) and F^{A} (eq 17), respectively, where $\rho_i(b,l)$ replaces ρ^{HS}.

II.9. Calculation of Differences $S_m - S_n$. With QH, LS, and HSMC(D) calculation of $\Delta S_{mn} = S_m - S_n$ is based on the absolute values for each microstate. However, in section II.3 we have argued that the definition of a microstate m depends to a large extent on the simulation time t where *typically* m and its energy and entropy all grow with t. To be able to carry out a reliable estimation of ΔS_{mn} (ΔF_{mn}, etc.) we simulate both m and n for the same t looking for a range of t values where $\Delta F_{mn}(t)$, $\Delta S_{mn}(t)$, and $\Delta E_{mn}(t)$ are stable within the statistical errors [due to the simultaneous increase of $E_m(t)$, $E_n(t)$, etc.]. For the QH method such stable results constitute the best final answer. For the LS method, on the other hand, one can calculate $\Delta S^{\text{A}}_{mn}(b,l)$ [and $\Delta F^{\text{A}}_{mn}(b,l)$] for a set of improved approximations (by increasing b and l); if these differences converge within the statistical errors, the converged values are considered to be the correct differences due to cancellation of equal systematic errors in $S^{\text{A}}_m(b,l)$ and $S^{\text{A}}_n(b,l)$ (see discussion in section II.10). Notice that LS, unlike QH,[43] is applicable to a sample which covers several microstates and, in principle, even to a random coil.[49]

Obviously, if m is less stable than n, then the t values should be adjusted (i.e., decreased) to fit the stability of m. If m is significantly larger than n, then t_m should be large enough to allow adequate coverage of m $t_m \sim t_n[\Pi\Delta\alpha_k(m)/[\Pi\Delta\alpha_k(n)]$, where t_n is the time required to obtain an adequate sample for n. However, if $\Delta S_{mn}(t)$ increases monotonically, then it constitutes a lower bound. If the microstate is restrictive, e.g., side chains should populate a single rotamer, then the MD sample can be composed of several smaller samples each starts from the same structure with a different set of velocities. It should be pointed out that with LS and QH relatively large samples are required for obtaining converged TPs[46] and converged terms of the correlation matrix, σ (eq 5),[43] respectively. Therefore, one should verify that the samples remain in the original microstates and have not "escaped" to neighboring ones. We have developed methods for analyzing the stability of a microstate by calculating distribution profiles of dihedral angles.[9,75,77]

Unlike QH and LS, HSMC(D) is not based on gathering statistics from the studied sample; therefore, the required sample size is relatively small; also, F[HSMC(D)] (but not necessarily E and S[HSMC(D)]) can be obtained from a very small sample (even from a single conformation).[50] Therefore, in our studies the sample size for HSMC(D) is small, and it has been determined by the range of t values for which the average of E_m (E_n) is approximately constant. Again, one can envisage extreme cases where m is significantly larger than n, which would require increasing the sample size for m as described above for LS. With HSMC(D) the problem is to control the samples generated in the reconstruction process, as discussed in section II.6 and the next section (II.10). All these considerations are applicable to a peptide in different microstates (e.g., a helix, hairpin, or extended microstates[54,55]) as well as to a flexible surface loop, which populates significantly several microstates. In particular, the effect of sample size on $\Delta S_{mn} = S_m - S_n$ can be reduced, while controlling this effect with TI and the counting approaches is difficult (see discussion in ref 19).

II.10. Cancellation of Systematic Errors with HSMD. It should be pointed out that for any practical set of n_{init}, n'_f, j and bin sizes, $\delta\alpha_k$, the calculated S_m^{A} (S_n^{A}) will be approximate, and thus the corresponding difference, $S_m^{\text{A}} - S_n^{\text{A}}$, might be approximate as well. However, if $S_m^{\text{A}} - S_n^{\text{A}}$ is found to be stable for significantly improving approximations, the constant value can be considered to be the correct difference. Indeed, in the previous application of HSMD to peptides[55] and in the present study of a loop (see section III), relatively small values of n'_f and j have already led to stable differences, meaning that systematic errors in both S_m^{A} and S_n^{A} are comparable and thus are cancelled in $S_m^{\text{A}} - S_n^{\text{A}}$. This cancellation of relatively large systematic errors (discussed further below) makes HSMD an efficient procedure for peptides/loops.

To understand the basis for this cancellation, we examine first two one-dimensional harmonic microstates, i.e., two oscillators with equal mass and different spring constants f_1 and f_2. The *exact* entropy difference, $S_2 - S_1$, can be expressed in terms of the variances $\langle x_1^2 \rangle$ and $\langle x_2^2 \rangle$ of the corresponding coordinates

$$\Delta S_{2,1} = S_2 - S_1 = (1/2)k_{\text{B}} \ln\left(\frac{f_1}{f_2}\right) = k_{\text{B}}[\ln(\langle x_2^2 \rangle^{1/2}) - \ln(\langle x_1^2 \rangle^{1/2})] \quad (22)$$

One can estimate $\Delta S_{2,1}$ from two separate MD simulations, where the corresponding variances are calculated. If f_1 is significantly smaller than f_2 (i.e., f_1 defines a flatter parabola) and the same step size is used in both simulations a longer simulation will be required for f_1 than for f_2 to gain the same statistical precision. Therefore, if the same sample size is used for both microstates, then the statistical precision of $\Delta S_{2,1}$ will be determined mostly by that of S_1.

We now examine the entropy contributed by a backbone dihedral angle, α_k (denoted α for simplicity), in the course of the reconstruction process. α varies in microstates 1 and 2 within the ranges $\Delta\alpha_1$ and $\Delta\alpha_2$ (eq 13), which we denote Δ_1 and Δ_2, respectively. The crudest (but sometimes quite reliable) HSMD approximation for the corresponding difference in entropy, $\Delta S_0(\alpha)$, is

$$\Delta S_0(\alpha) = k_{\text{B}}[\ln\Delta_2 - \ln\Delta_1] \quad (23)$$

which is similar to that of eq 22 above (for brevity we shall omit α from the equations below). For better HSMD approximations, $\Delta S_0^{n_f}(l)$, we define the bins $\delta_1 = \Delta_1/l$ and $\delta_2 = \Delta_2/l$, where l is an increasing integer; the corresponding probabilities are $p_1^{n_f}(l)$ and $p_2^{n_f}(l)$ which are defined by n_{visit}/n_f (eq 14). One obtains

$$\Delta S_0^{n_f}(l) = k_{\text{B}}[\ln(p_1^{n_f}(l)/\delta_1) - \ln(p_2^{n_f}(l)/\delta_2)] = k_{\text{B}}\{\ln[p_1^{n_f}(l)/p_2^{n_f}(l)] + \ln(\Delta_2/\Delta_1)\}$$

or

$$\Delta S_0^{n_f}(l) = \Delta S^{n_f}(l) + \Delta S_0 \quad (24)$$

where $\Delta S^{n_f}(l)$ can be viewed as an anharmonic term. One can write $p_i^{\text{exact}}(l) = p_i^{n_f}(l)x_i^{n_f}(l)$ for $i = 1,2$, where $p_i^{\text{exact}}(l) = p_i^{n_f=\infty}(l)$ and $x_i^{n_f}(l)$ are thus factors (systematic errors) satisfying $x_i^{n_f}(l) \to 1$ for very large n_f; for a given l (bin) one obtains

$$\Delta S^{n_f}(l) = k_{\text{B}}\{\ln p_1^{\text{exact}}(l) - \ln p_2^{\text{exact}}(l) + \ln[x_2^{n_f}(l)/x_1^{n_f}(l)]\} \quad (25)$$

However, for large bins, δ (small l), one would expect to obtain probabilities that are close to the exact ones, $p_1^{\text{exact}}(l)$ and $p_2^{\text{exact}}(l)$ [i.e., $x_1^{n_f}(l)$ and $x_2^{n_f}(l)$ are ~1] for a relatively small n_f due to adequate statistics, i.e., relatively large n_{visit} values. To obtain the exact probabilities (within the statistical errors) for decreased bin sizes, n_f should be increased adequately, which might increase computer time significantly. Thus, for practical values of n_f, $x_1^{n_f}(l)$ and $x_2^{n_f}(l)$ might differ significantly from 1 (i.e., large systematic errors). However, we argue that already for relatively small n_f, $x_2^{n_f}(l) \approx x_1^{n_f}(l)$, and the last logarithmic term (eq 25) becomes smaller than the *statistical* error leading to the correct value, $\Delta S(l)$, within the statistical error. To obtain the correct

contribution (ΔS) of dihedral angle α to the *entropy difference* one has to define small enough bins, i.e., large enough l_{min}, where for $l > l_{min}$ $\Delta S(l)$ remains unchanged within the statistical error. As expected, l_{min} has to be smaller for a linear peptide than for a protein loop due to the restriction of loop closure, which requires relatively small bins (see sections III.1, 5, and 7).

The relation $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ stems from two reasons, where the first one is the fact that HSMD takes all interactions into account, and thus for a given n_f the future part of the chain is treated with the same level of approximation in both microstates. Second, because with MD the atoms are moved along their potential gradients, the simulations are equally efficient in both microstates. For peptides[55] the condition $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ occurs for much smaller n_f with HSMD than with HSMC[53] because the efficiency of the MC procedure used by us depends on the compactness of a structure (e.g., hairpin versus extended). Again, as for the parabolas above, if one microstate is significantly "flatter" than the other (i.e., larger $\Delta\alpha_k$ values), the required n_f value for obtaining convergence of ΔS will be determined mainly by the flatter microstate.

It should be noted that a $\Delta\alpha_k$ value of the studied sample might be significantly larger than the actual $\Delta\alpha_k$ available for α_k at step k of the reconstruction process of conformation i, due to geometrical constraints imposed by the constant "past", i.e., the partial structure reconstructed in the previous steps, 1....$k-1$. This limiting effect is expected to be more significant for dihedral angles than for bond angles; moreover, because at step k n_{visit} depends on mutual visits to the dihedral angle bin, $\delta\alpha_k$, and to its successive bond angle bin, $\delta\alpha_{k+1}$ (i.e., the modified eq 14 is $\rho^{HS}(\alpha_k,\alpha_{k+1}|\alpha_{k-1} \cdots \alpha_1) = n_{visit}/[n_f\, \delta\alpha_k \delta\alpha_{k+1}]$), $\delta\alpha_k$ and $\delta\alpha_{k+1}$ can be optimized to reduce S^A for a given n_f, which would lead to higher efficiency, i.e., to converged $S_m^A - S_n^A$ for smaller n_f (see sections III.4 and III.5). One can envisage a situation where for some side chains all rotamers are populated in one microstate, but only one rotamer is populated in the other microstate and vice versa (see section II.7). These differences might compensate each other in $S_m^A - S_n^A$; therefore, evaluating the reconstruction calculations should be carried out with extra caution. Again, the ultimate test for accuracy is the occurrence of stable $S_m^A - S_n^A$ values for increasing n_f and decreasing bin sizes, as previously discussed.

As mentioned above, with the MC method used by us[53] an open peptide structure (e.g., the extended microstate of a peptide) is simulated more efficiently than a compact hairpin structure and therefore relatively large n_f was needed to achieve $x_2^{n_f}(l) \approx x_1^{n_f}(l)$ within the statistical errors. Thus far we have studied by HSMD microstates of three systems, degaglycine, $NH_2(Val)_2(Gly)_6(Val)_2CONH_2$,[55] and in this paper the 7-residue loop of α-amylase in vacuum and implicit solvent. In all these studies the cancellation of systematic errors has been found to occur for relatively small n_f, which has been verified by comparing entropy differences obtained for a wide range of n_f values. For decaglycine, for example, n_f ranges between 500 and 24 000, where $n_f = 500$ (5 ps) leads to the correct results, and HSMD is thus ~100 times more efficient (in terms of computer time) than HSMC.[53] We expect this cancellation of errors to occur also for models of peptides and loops in explicit water.

III. Results and Discussion

III.1. Simulation Details for the Loop in Vacuum. An MD simulation (at 300 K) starting from the free PDB structure has led to a stable sample of size 600 (where a structure is added to the sample every 0.5 ps) around the PDB structure with an average energy of ~ -138 kcal/mol. However, in simulations of this size, starting from the bound PDB structure, the initial energy (~ -98 kcal/mol) was decreased constantly and significantly; we have found this energy to stabilize around -110 kcal/mol only after a very long MD run. Because we are interested in studying the stability of the bound microstate in the free protein (see sections I.7 and II.1), its sample was generated by combining partial samples obtained from short MD runs each started from the PDB bound structure with a different set of velocities. The average energy of these short samples remained close to -98 kcal/mol. This procedure can be used with HSMC(D), which operates on small samples (and in an extreme case even on a single structure), while it is less effective with the LS[9] and the QH methods, which require much larger samples, as discussed earlier (section II.9). Generating such a combined sample will be useful also for studying the entropy (and energy) of a transition state.

The free and bound samples and the reconstruction simulations (future samples) were carried out with the velocity-Verlet algorithm[28] based on a time step of 1 fs, where the Berendsen[28] heat bath controlled the temperature. Cut-offs on long-range interactions were not imposed, and in the reconstruction process a structure was added to the sample every $g = 10$ fs, where the $n_{init} = 250$ initial structures (2.5 ps) were discarded for equilibration. The future samples were generated for four bin sizes, $\delta = \Delta\alpha_k/30$, $\Delta\alpha_k/15$, $\Delta\alpha_k/10$, and $\Delta\alpha_k/5$, centered at α_k (i.e., $\alpha_k \pm \delta/2$) (eqs 13 and 14). If the counts of the smallest bin are smaller than 50, then the bin size is increased to the next size and if necessary to the next one, etc. In the case of zero counts, n_{visit} is taken to be 1; however, zero counts is a very rare event. In this context it should be pointed out that if one had tried to build loop structure i by selecting angles at random within the ranges $\alpha_k \pm \delta/2$, the constructed structure would differ from i and in the case of a loop would not satisfy the loop closure condition leading to a very high bond stretching energy. Therefore, the smallest bin chosen for a loop ($\Delta\alpha_k/30$) is smaller than that used for the linear peptides[55] ($\Delta\alpha_k/15$). Notice, however, that this structural deviation from i would affect both microstates, and the bins used are the largest that still lead to converging results of ΔS^A.

For each microstate, two sets of results were calculated, one is based on unit $n'_f = 250$ (2.5 ps) and n_f values of 250 ($j=1$), 500 ($j=2$), 750 ($j=3$), and 1250 ($j=5$). The second set is based on unit $n'_f = 1000$ (10 ps) and n_f values of 1000 ($j=1$), 2000 ($j=2$), 4000 ($j=4$), and 8000 ($j=8$) (see section II.6). These sets that lead to an increasing coverage of the studied microstates, enable one to examine the convergence of $S^A(n'_f, j)$ as well as $\Delta S_{mn}^A(n'_f,j)$, which is our main interest.

Table 1. Differences $\Delta\alpha_k$ (in deg) between the Minimum and Maximum Values of Dihedral Angles in the Free and Bound Samples in Vacuum[a]

	free loop				bound loop			
	studied sample		1 × 2.5 ps (5 × 2.5 ps)		studied sample		1 × 2.5 ps (5 × 2.5 ps)	
residue	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$
Gly 1	46	74	43 (52)	61 (108)	43	90	51 (48)	96 (74)
His 2	75	88	62 (98)	71 (189)	89	61	78 (87)	51 (68)
Gly 3	58	112	64 (298)	80 (109)	68	72	54 (73)	63 (88)
Ala 4	99	89	73 (103)	70 (87)	91	94	71 (73)	53 (76)
Gly 5	99	105	77 (83)	80 (139)	84	101	58 (70)	43 (60)
Gly 6	112	85	88 (118)	85 (67)	59	64	43 (57)	42 (46)
Ser 7	69	54	52 (88)	65 (65)	81	44	42 (58)	35 (39)
χ^1(His)	58		36 (66)		45		44 (50)	
χ^2(His)	145		117 (116)		103		62 (96)	
χ^1 (Ser)	155		55 (163)		39		31 (51)	

[a] $\Delta\alpha_k$ are defined in eq 13. The studied samples of $n = 600$ conformations were generated with the AMBER force field by retaining a conformation every 500 fs. The 1 × 2.5 ps samples (of 250 conformations each) were started from two chosen conformations of the free and bound (studied) samples, by retaining a conformation every 10 fs and ignoring the first 250 conformations for equilibration. The sample denoted (5 × 2.5 ps) consists of five 2.5 ps samples (altogether 1250 conformations) each started from the chosen structure with a different set of velocities where the initial 250 conformations are ignored for equilibration.

III.2. Entropy Results in Vacuum. In Table 1 we present the values of $\Delta\alpha_k$ (eq 13) for the free and bound microstates obtained from the corresponding MD samples. These values suggest that the two samples indeed are concentrated in conformational space. Table 2 contains two sets of results for the entropy, TS^A (eq16) based on units of 2.5 and 10 ps for the free and bound microstates. As mentioned in section III.1, these results were calculated for four different future sample sizes, n_f and four bin sizes; however, the extent of convergence is demonstrated by the results obtained for the three smallest bin sizes, $\Delta\alpha_k/10$, $\Delta\alpha_k/15$, and $\Delta\alpha_k/30$, which are presented in the table. The statistical errors were obtained from the fluctuations and results obtained for partial samples. These results were obtained without considering the Jacobian ($\Pi_j \sin(\theta_j)$ (see discussion following eq 6), which enables one to compare them to those obtained previously for peptides[55] without considering the Jacobian as well. As shown later in section III.4, the contribution of the Jacobian to the entropy cancels out to a very good approximation in entropy and free energy differences—our main interest. Therefore, ignoring the Jacobian, which increases the statistical errors (hence requires larger samples), is justified.

One would expect S^A to decrease with decreasing the bin and increasing n_f—an expectation, which is fully satisfied by the results of Table 2. In particular, for a given n_f, S^A always decreases as the bin is decreased; however, a complete convergence occurs only for the free loop (unit=2.5 ps), where $TS^A(\Delta\alpha_k/15, n_f=1250)$ is equal to $TS^A(\Delta\alpha_k/30, n_f=1250)$ within the relatively large statistical errors; for the other cases the deviation from convergence are small. Convergence of the two best results (i.e., for the largest n_f) for each bin occurs only for unit = 10 (for both microstates), while for unit = 2.5 ps the deviations from full convergence are again small $T[S^A(\delta, n_f=750) - S^A(\delta, n_f=1250) \leq 1.2$ kcal/mol for all δ]. It should be pointed out that the errors for the bound loop are smaller than those for the free loop probably due to the fact that the sample of the bound loop consists of several subsamples that were generated from the same initial structure.

The HSMD results for the entropy are compared in the table to those obtained with the LS and QH methods from larger MD samples of 5000, 8000, and 10 000 conformations. These samples consist of several subsamples each started from the same structure with a different set of velocities, where a conformation was retained every 50 fs. The QH results for *TS* exceed the HSMD values by 12.8 and 11.3 kcal/mol for the free and bound microstate, respectively, which is in accord with S^{QH} being an upper bound; these differences are probably also affected (i.e., increased) by the significantly larger samples used for QH than for HSMD (see discussion in section II.9). The LS results (calculated for $b=1$, $l=10$), which also constitute upper bounds, are larger than the corresponding QH values, as was also found in previous studies.[53−55]

III.3. Free Energy Results in Vacuum. Results for the free energy functional, F^A (eq 17), its fluctuation, σ_A (eq 19), and the energies are presented in Table 3. These results are given only for the smallest bin, $\Delta\alpha_k/30$ and unit = 10, because F^A values for the other bins can be obtained from the entropies of Table 2 and the energies provided in the bottom of Table 3. F^A (like S^A) does not change (within the errors bars) as n_f is increased from 5000 to 8000, and the slight (expected) decrease of the central values of σ_A with increasing n_f is, however, insignificant within the error bars. As expected, the QH and LS results for *F* underestimate the correct values, and the central values of the energy fluctuations are always larger than those for $\sigma_A(n_f=8000)$. Finally, the table shows the differences in free energy and energy between the free and bound microstates. It is evident that the free energy differences, ΔF^A, are all equal within the statistical errors, and they are also equal to the energy difference, ΔE and ΔF^A_{LS}, obtained with LS. This suggests that the ΔS^A results are ∼0, as indeed shown in the next section, III.4, i.e., the free microstate is more stable than the bound one by ∼38.8 kcal/mol which is contributed mainly by ΔE.

The results for F^B (eq 20) are not provided in the table because they do not behave as expected, i.e., they do not decrease as n_f is increased and the bin is decreased. This "misbehavior" can be attributed to a too small sample size *n* and to the fact that the bond stretching energy is included in the potential energy, while the corresponding entropy is not taken into account in ρ^{HS} (eq 15) (see discussion in ref 55). Still, the results obtained for F^B are always larger than those of F^A and thus probably provide upper bounds; the deviations, however, are relatively large ($F^B = -191.7$ and -150.2 kcal/mol, deviating from F^A by ∼12 and 14 kcal/mol for the free and bound microstates, respectively). Due to the almost convergence of the F^A values it is plausible to assume that the F^B results do not lead to improved approximations for the free energy, i.e., the average values, $F^M = (F^A+F^B)/2$ are probably less reliable than those of F^A. Notice, however, that $\Delta F^B = -39.7$ is only 1 kcal/mol

Table 2. HSMD Results (in kcal/mol) for the Entropy, TS^{A} (Eqs 16 and 18), at $T = 300$ K Calculated from Samples of 600 Conformations of the Free and Bound Microstates in Vacuum[a]

	free loop				bound loop			
	unit = 2. 5 ps (250)		unit = 10 ps (1000)		unit = 2.5 ps (250)		unit = 10 ps (1000)	
Δ	$n_f(j)$	TS^{A}	$n_f(j)$	TS^{A}	$n_f(j)$	TS^{A}	$n_f(j)$	TS^{A}
$\Delta\alpha_k/10$	250 (1)	72.1 (3)	1000 (1)	68.6 (2)	250 (1)	71.01 (3)	1000 (1)	67.78 (3)
$\Delta\alpha_k/10$	500 (2)	69.5 (3)	2000 (2)	68.2 (2)	500 (2)	68.84 (3)	2000 (2)	67.35 (3)
$\Delta\alpha_k/10$	750 (3)	68.6 (3)	4000 (5)	68.1 (1)	750 (3)	67.81 (4)	4000 (5)	67.28 (3)
$\Delta\alpha_k/10$	1250 (5)	**67.9 (2)**	8000 (8)	**68.0 (1)**	1250 (5)	**67.15 (3)**	8000 (8)	**67.26 (2)**
$\Delta\alpha_k/15$	250 (1)	71.9 (3)	1000 (1)	67.4 (2)	250 (1)	70.90 (3)	1000 (1)	66.89 (3)
$\Delta\alpha_k/15$	500 (2)	69.0 (3)	2000 (2)	66.8 (2)	500 (2)	68.46 (5)	2000 (2)	66.28 (3)
$\Delta\alpha_k/15$	750 (3)	67.6 (3)	4000 (5)	66.7 (1)	750 (3)	67.14 (4)	4000 (5)	66.24 (3)
$\Delta\alpha_k/15$	1250 (5)	**66.6 (2)**	8000 (8)	**66.7 (1)**	1250 (5)	**66.10 (3)**	8000 (8)	**66.22 (3)**
$\Delta\alpha_k/30$	250 (1)	71.9 (2)	1000 (1)	67.2 (2)	250 (1)	70.89 (3)	1000 (1)	66.77 (4)
$\Delta\alpha_k/30$	500 (2)	68.9 (2)	2000 (2)	66.3 (2)	500 (2)	68.42 (4)	2000 (2)	65.97 (4)
$\Delta\alpha_k/30$	750 (3)	67.5 (2)	4000 (5)	65.9 (1)	750 (3)	67.06 (4)	4000 (5)	65.60 (4)
$\Delta\alpha_k/30$	1250 (5)	**66.3 (2)**	8000 (8)	**65.8 (1)**	1250 (5)	**65.91 (3)**	8000 (8)	**65.51 (4)**
TS^{QH}		78.61 (7)		78.61 (7)		76.8 (2)		76.8 (2)
TS^{LS}		91.6 (4)		91.6 (4)		90.9 (7)		90.9 (7)

[a] The bin sizes are $\delta = \Delta\alpha_k/l$ (eq 13). The two units of 2.5 and 10 ps used are also defined (in parentheses) by their number of conformations, 250 and 1000, respectively. n_f denotes the sample size of the future chains used in the reconstruction process, $n_f =$ unit $\times$ j, where j is the number of simulations of unit size applied at each reconstruction step. The statistical errors are given in parentheses, e.g., 66.3 (2) = 66.3 $\pm$ 0.2. S^{QH} is the quasi-harmonic entropy (eq 5), and S^{LS} (eqs 16 and 18 and section II.8) is S^{A} obtained by the local states method using $b = 1$ and discretization parameter $l = 10$; these results were obtained from larger samples (for details see text). All calculations were carried out with the AMBER force field. The entropy is defined up to an additive constant that is the same for both microstates.

Table 3. HSMD Results at $T = 300$ K for the Free Energy, F^{A}, the Interaction Energy, E_{int}, Their Fluctuations, and ΔF^{A} and ΔE for the Free and Bound Microstates in Vacuum[a]

	free loop		bound loop		free−bound	
n_f	$-F^{A}$	σ_A, σ_E	$-F^{A}$	σ_A, σ_E	ΔF^{A}	ΔE_{int}
1000	204.7 (3)	4.4 (3)	165.79 (8)	4.5 (3)		−38.9 (4)
2000	203.7 (3)	4.3 (3)	165.0 (1)	4.4 (2)		−38.7 (4)
4000	203.3 (3)	4.3 (3)	164.62 (6)	4.4 (2)		−38.7 (3)
8000	203.3 (2)	4.2 (3)	164.54 (6)	4.4 (2)		−38.7 (3)
$-F^{QH}$	216.1 (1)		175.8 (3)			−40.2 (4)
$-F^{LS}$	229.1 (4)		190.0 (7)			−39.1 (5)
$-E_{int}$	137.5 (3)	4.4 (3)	99.02 (5)	4.49 (4)		−38.4 (3)

[a] F^{A} (eq 17) is a lower bound of the free energy, and σ_A (eq 19) is its fluctuation. The results were obtained from samples of $n = 600$ conformations for the smallest bin size, $\delta = \Delta\alpha_k/30$, unit = 10 ps, and all future sample sizes n_f. F^{QH} (see eq 5) and F^{LS} (eq 17 and section II.8) are free energies obtained by the quasi-harmonic approximation and the local states method (b=1, l=10), respectively, and are based on larger samples (see text). The average potential energy, E_{int}, of the studied samples appears in the bottom row; σ_E is the energy fluctuation (these results are in kcal/mol). All free energies are in kcal/mol and are defined up to the same additive constant for both microstates. All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

Table 4. Entropy Differences, $T\Delta S^{A} = T[S^{A}_{free} - S^{A}_{bound}]$ (in kcal/mol) at $T = 300$ K in Vacuum[a]

	unit = 2.5 ps (250)			unit = 10 ps (1000)		
	n_f	$T\Delta S^{A}$	$T\Delta S^{A}$ (Jacobian)	n_f	$T\Delta S^{A}$	$T\Delta S^{A}$ (Jacobian)
$\Delta\alpha_k/15$	250	1.0 (2)	1.2 (1)	1000	0.5 (2)	0.6(2)
$\Delta\alpha_k/15$	500	0.5 (2)	0.6 (2)	2000	0.5 (2)	0.7 (1)
$\Delta\alpha_k/15$	750	0.5 (2)	0.6 (2)	4000	0.5 (1)	0.6 (1)
$\Delta\alpha_k/15$	1250	**0.5 (1)**	**0.6 (1)**	8000	**0.5 (1)**	**0.6 (1)**
$\Delta\alpha_k/30$	250	1.0 (2)	1.1 (1)	1000	0.5 (2)	0.6 (1)
$\Delta\alpha_k/30$	500	0.5 (2)	0.6 (2)	2000	0.3 (2)	0.4 (1)
$\Delta\alpha_k/30$	750	0.4 (2)	0.5 (2)	4000	0.3 (1)	0.4 (1)
$\Delta\alpha_k/30$	1250	**0.4 (1)**	**0.5 (1)**	8000	**0.3 (1)**	**0.4 (1)**
$T\Delta S^{QH}$		1.8 (2)			1.8 (2)	
$T\Delta S^{LS}$		0.6 (3)			0.6 (3)	

[a] S^{A} is an upper bound of the entropy (eqs 16 and 18). The results for $T\Delta S^{A}$ were obtained from samples of $n = 600$ conformations for the two smallest bins, $\delta = \Delta\alpha_k/15$ and $\delta = \Delta\alpha_k/30$, unit = 2.5 and 10 ps, and all the future sample sizes, n_f. We also present $T\Delta S^{A}$ results obtained by HSMD, where S^{A} is calculated with the Jacobian (see discussion following eq 6). $T\Delta S^{QH}$ (eq 5) and $T\Delta S^{LS}$(eqs 16 and 18 and section II.8) are entropy differences calculated by the quasi-harmonic approximation and the local states method (b=1, l=10); they are based on larger samples (see text). All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

smaller than ΔF^{A}. We also calculated the corresponding F^{D} values, −196.9 and −154.0 kcal/mol (eq 21), which are smaller than the related F^{B} results but are larger than those for F^{M}, leading to $\Delta F^{D} = -42.9$ kcal/mol. While it would be desirable to have converging results for F^{B} and F^{D}, we demonstrate below that reliable *differences* in S (and F) can be obtained from differences in S^{A} (and F^{A}).

III.4. Entropy Differences in Vacuum. Because computer time increases linearly with n_f, it is of interest to check the effect of decreased n_f on entropy differences. In Table 4 results are presented for $T\Delta S^{A} = T[S^{A}_{free} - S^{A}_{bound}]$ calculated for the two smallest bins, the four n_f values, and unit = 2.5 and 10 ps. We also present results for $T\Delta S^{A}$ calculated with the Jacobian. The table reveals that the corresponding results obtained with and without the Jacobian are equal within the error bars. This is important because the calculations without the Jacobian have converged statistically already for samples of 400 conformations, while including the Jacobian required increasing the sample size to 600. The results for unit = 2.5 and $n_f \geq 500$ are converged (within the error bars) with respect to both n_f and the two bin sizes. The fact that the

$T\Delta S^A$ value obtained for unit = 2.5 ps and $n_f = 500$ is equal to that obtained for a four times larger unit (of 10 ps) and for a 16 times larger n_f (8000) suggests that $T\Delta S^A = 0.3 \pm 0.1$ kcal/mol is the correct result.

This stems from the cancellation (in $T\Delta S^A$) of approximately equal systematic errors for both microstates, as discussed in section II.10. Thus, Table 2 shows that the worst approximations that still lead to the correct $T\Delta S^A$ differ from the best ones by $TS^A(\Delta\alpha_k/15, n_f=500) - TS^A(\Delta\alpha_k/30, n_f=8000) = 3.2$ and 3.0 kcal/mol for the free and bound microstates, respectively; these differences constitute lower bounds because the correct TS values might be significantly smaller than $TS^A(\Delta\alpha_k/30, n_f=8000)$. The table shows that the difference obtained by the LS method [0.6 (3) kcal/mol] is equal to that obtained by HSMD, while QH leads to a significantly higher difference, 1.8 (2) kcal/mol. However, this good LS result might be accidental as unreliable differences were obtained by LS for the extended, helix, and hairpin microstates of decaglycine.[53,55]

The similar results obtained for unit = 2.5 and 10 ps suggest that already for unit =2.5 (n_f=500) the coverage of both microstates by the future chains is adequate and that for unit = 10 and n_f = 8000 the future chains still remain within these microstates. To get an idea about the extent of this coverage, we selected a structure from each of the two studied samples from which MD simulations of the *entire* loop were started (see the last paragraph of section II.6). Two samples of 250 conformations and two samples of $5 \times 250 = 1250$ conformations were generated in the same way the future chains are simulated during the reconstruction process, i.e., 1250 consists of five subsamples (i.e., units) of 250 conformations, each starting from the initial structure with a different set of velocities where the first 250 structures are ignored for equilibration. In these simulations a structure is retained (as in the reconstruction process) every g = 10 fs (unlike the two studied samples that were generated with g=500 fs). The $\Delta\alpha_k$ results (eq 13) for the dihedral angles for these samples are presented in Table 1 which shows that in most cases the results for 2.5 ps are slightly smaller than the corresponding results obtained for the studied samples, while the (expected) larger results for 5×250 are still close to those of the studied samples; this applies also to the side chains. Deviations from this picture occur for 5×250, where $\Delta\psi$ and $\Delta\varphi$ of His^2 and Gly^3, respectively, are significantly larger than the corresponding values of the studied samples. This picture suggests that the reconstruction simulations cover adequately the two studied microstates.

In view of the discussion in section II.10 we have also tried to optimize the bins' sizes. As pointed out there, the values of $\Delta\alpha_k$(dihedral) in Table 1 (for the entire samples) are expected to overestimate the actual $\Delta\alpha_k$ available for α_k at step k of the reconstruction process. Therefore, a relatively large number of visits of α_k(dihedral) to its bin $\delta\alpha_k = \Delta\alpha_k/l$ are not followed by visits of α_{k+1}(bond angle) to its bin, α_{k+1}/l; thus, the number of counts (at both $\delta\alpha_k$ and $\delta\alpha_{k+1}$) is relatively small, while $\Delta\alpha_k$(dihedral)/l is large, leading to small $\rho^{HS}(\alpha_k,\alpha_{k+1}|\alpha_{k-1}\cdots\alpha_1) = n_{visit}/[n_f\delta\alpha_k\delta\alpha_{k+1}]$, i.e., to a large contribution, $-k_B\ln\rho^{HS}$, to the entropy. This undesirable effect can be reduced by decreasing the values of $\Delta\alpha_k$-(dihedral) used for defining the dihedral bins or increasing the values of $\Delta\alpha_k$(bond angle) used for defining the bond angles' bins. We have adopted the latter option by increasing *all* of the bond angles bins to 50°/l (typically $\Delta\alpha_k$(bond angle) ranges from 20 to 25°) and applied HSMD with a relatively small unit = 1 ps and small n_f = 100, 200, 300, and 400. The results for $T\Delta S^A$ appear in Table 5 and are shown to be very close to those of Table 4, which suggests that HSMD can be optimized further leading to a further reduction in computer time.

Table 5. Entropy Differences, $T\Delta S^A = T[S^A_{free} - S^A_{bound}]$ (in kcal/mol) at T = 300 K in Vacuum Using Equal Bins for the Bond Angles[a]

unit = 1 ps (100)		
	n_f	$T\Delta$S
$\Delta\alpha_k$/10	100	0.7 (1)
$\Delta\alpha_k$/10	200	0.7 (1)
$\Delta\alpha_k$/10	300	0.7 (2)
$\Delta\alpha_k$/10	400	**0.5 (1)**
$\Delta\alpha_k$/15	100	0.6 (1)
$\Delta\alpha_k$/15	200	0.6 (1)
$\Delta\alpha_k$/15	300	0.6 (2)
$\Delta\alpha_k$/15	400	**0.5 (2)**
$\Delta\alpha_k$/30	100	0.6 (1)
$\Delta\alpha_k$/30	200	0.6 (1)
$\Delta\alpha_k$/30	300	0.6 (2)
$\Delta\alpha_k$/30	400	**0.4 (2)**

[a] S^A is an upper bound of the entropy (eqs 16 and 18). The results for $T\Delta S^A$ were obtained from samples of n = 600 conformations for the three smallest bins, using unit = 1 ps. The bond angles bins are $\delta = 50°/l$, while for the dihedral angles they are $\delta = \Delta\alpha_k/l$ (eq 13), where l = 10, 15, and 30. n_f is the sample size of the future chains in the reconstruction procedure. All calculations were carried out with the AMBER force field. The statistical error is defined in footnote *a* of Table 2.

These results support the conclusions obtained for peptides[55] that correct differences ΔS^A_{mn} can be obtained for relatively short reconstruction simulations, which leads to considerable savings in computer time. In fact, reconstruction of a structure based on n_f = 500 and 100 requires, respectively, ~30 and 14 min CPU on a 2.1 GHz Athlon processor. This time can be reduced by a factor of 2 if the MD integration is carried out with a time step of 2 fs (rather than 1 fs). Due to strong correlations among the dihedrals and bond angles within a microstate, it might be possible to treat four successive angles (two dihedrals and two bond angles) rather than two angles considered presently at each reconstruction step. One can increase efficiency further by applying a cutoff on long-range interactions and running the simulations on the best machines available to date. One would seek to decrease computer time further by considering the conformational restraints imposed by the loop closure condition on the pair of dihedral angles, α_k, and its successive bond angle, α_{k+1}, at each reconstruction step. However, in spite of this restraint the fluctuations in these angles (partially due to bond stretching) are significant in all reconstruction steps besides the last two (K- 4, K-3 and K-2, K-1, where K is the last angle in the loop). While one could probably ignore the reconstruction of these two last steps in both microstates (as long as differences, ΔS^A are of interest), the gain in

Table 6. Differences $\Delta\alpha_k$, (in deg) between the Minimum and Maximum Values of Dihedral Angles in the Free and Bound Samples in Solvent[a]

	free loop (solvent)				bound loop (solvent)			
	entire sample		1×5 ps (10×5 ps)		entire sample		1×5 ps (10×5 ps)	
residue	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$	$\Delta\varphi$	$\Delta\psi$
Gly 1	76	153	54 (101)	122 (175)	92	148	85 (109)	125 (200)
His 2	139	130	114 (140)	95 (360)	125	105	105 (122)	101 (162)
Gly 3	175	124	88 (285)	99 (176)	80	95	100 (202)	139 (151)
Ala 4	131	94	68 (170)	75 (89)	143	288	134 (168)	125 (360)
Gly 5	107	100	89 (119)	111 (131)	199	360	130 (226)	122 (360)
Gly 6	126	109	154 (351)	97 (134)	285	267	205 (293)	357 (357)
Ser 7	83	64	75 (87)	56 (71)	243	109	127 (188)	90 (114)
X^1 (His)	55		43 (66)		53		33 (77)	
X^2 (His)	130		102 (161)		108		95 (130)	
X^1 (Ser)	317		60 (188)		321		51 (167)	

[a] $\Delta\alpha_k$ are defined in eq 13. The studied samples of $n = 500$ conformations were generated by retaining a conformation every 500 fs. The 1×5 ps samples (of 500 conformations each) were started from two chosen conformations of the free and bound (studied) samples, by retaining a conformation every 10 fs and ignoring the first 250 conformations for equilibration. The sample denoted (10×5 ps) consists of ten 5 ps samples (altogether 5000 conformations) each started from the chosen structure with a different set of velocities where the initial 250 conformations are ignored for equilibration. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA.

computer time would be small. Finally, while the structure of TINKER makes it a very convenient tool for developing new programs, the code is not very efficient, decreasing the performance of HSMD as well.

III.5. Results for the Loop in Implicit Water. These MD simulations are based on the AMBER force field[68] and the GB/SA solvation model of Still and co-workers[69] which (like AMBER) is implemented within TINKER.[70] The samples for the free and bound microstates were generated at $T = 300$ K in a similar way to those in vacuum with some changes: the sample size is $n = 500$ (rather than 600), and the MD step size was increased to 2 fs, where bonds involving hydrogens were frozen to their ideal values by using the RATTLE algorithm;[28] also, the smallest bin size in the reconstruction process was decreased to $\delta = \Delta\alpha_k/45$, and thus the other three bins are $\Delta\alpha_k/30$, $\Delta\alpha_k/15$, and $\Delta\alpha_k/10$.

The $\Delta\alpha_k$ results for the free and bound samples appear in Table 6, and they are shown to be larger than the corresponding values obtained for the loop in vacuum in Table 1. This increase in $\Delta\alpha_k$ is expected due to the protein−solvent interactions that lead to an increase in the loop flexibility, hence to its larger entropy. The table also reveals that in most cases the $\Delta\alpha_k$ values of the bound microstate are larger than their counterparts in the free microstate and in some cases, for $\Delta\varphi$ of Gly[5], Gly[6], and Ser[7] and $\Delta\psi$ of Ala[4], Gly[5], and Gly[6], the difference is significant where $\Delta\alpha_k$-(bound) ranges from 200 to 360°. This might lead to the conclusion that the entropy of the bound microstate is significantly larger than that of the free microstate. However, one should bear in mind that the α_k are highly correlated, and in the case of small residues, such as Gly and Ala significant simultaneous changes in neighbor dihedral angles

Table 7. HSMD Results (in kcal/mol) for the Entropy, TS^A (Eqs 16 and 18) at $T = 300$ K Calculated from a Sample of 200 Conformations of the Free and Bound Microstates in Solvent[a]

bin size	n_f	TS^A_{free}	TS^A_{bound}	$T[S^A_{free} - S^A_{bound}]$
$\Delta\alpha_k/15$	625	70.7 (1)	71.0 (2)	- 0.3 (1)
$\Delta\alpha_k/15$	1250	69.9(1)	70.3 (2)	−0.4 (1)
$\Delta\alpha_k/15$	2500	69.7 (2)	70.1 (2)	−0.4 (1)
$\Delta\alpha_k/15$	5000	**69.6 (2)**	**70.0 (2)**	**−0.4 (1)**
$\Delta\alpha_k/30$	625	70.6 (1)	70.9 (2)	−0.3 (1)
$\Delta\alpha_k/30$	1250	69.4 (1)	69.8 (2)	−0.4 (1)
$\Delta\alpha_k/30$	2500	68.8 (2)	69.2 (2)	−0.4 (1)
$\Delta\alpha_k/30$	5000	**68.4 (2)**	**68.8 (2)**	**−0.4 (1)**
$\Delta\alpha_k/45$	625	70.6 (1)	70.9 (2)	−0.3 (1)
$\Delta\alpha_k/45$	1250	69.4 (2)	69.7 (2)	−0.4 (1)
$\Delta\alpha_k/45$	2500	68.7 (2)	69.1 (2)	−0.4 (1)
$\Delta\alpha_k/45$	5000	**68.2 (2)**	**68.5 (1)**	**−0.4 (1)**
TS^{QH}		89 (1)	92 (1)	−3 (1)
TS^{LS}		100 (1)	108 (1)	−8 (1)

[a] The bin sizes for the bond angles are $\delta = 100°/l$ (i.e., $\Delta\alpha_k$=100°) and are $\delta = \Delta\alpha_k/l$ for the other angles, where $\Delta\alpha_k$ is defined in eq 13. n_f, the sample size of the future chains generated in the reconstruction process, is based on unit = 500 conformations (5 ps). S^{QH} is the quasi-harmonic entropy (eq 5), and S^{LS} (eqs 16 and 18 and section II.8) is S^A obtained by the local states (LS) method using $b = 1$ and the discretization parameter, $l = 10$; these results were obtained from larger samples (for details see text). The entropy is defined up to an additive constant which is the same for both microstates. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA. The statistical error is defined in footnote *a* of Table 2.

can lead mainly to small *localized* conformational changes and thus to a relatively narrow "pipe" of low entropy (see section II.3). One has also to verify that the large $\Delta\alpha_k$ values of the bound microstate do not lead to an overlap of the two microstates. Comparing structures of the two samples generated at the same time along the trajectories shows that the energies differ by ∼25 kcal/mol (see also Table 8), the rmsd of all heavy atoms is ∼2.2 Å, and the corresponding dihedral angles are different and in some cases significantly different (e.g., 109 vs −45° for ψ of Ala[4]).

Using the (relatively large) $\Delta\alpha_k$ values of Table 6 as a basis for defining the bin sizes for the reconstruction process will lead to a set of results S^A_{bound} with a lower level of approximation than the corresponding results of S^A_{free}; consequently, relatively small bins and large n_f values will be needed to obtain a set of converging results for the difference $\Delta S^A = S^A_{free}(\Delta\alpha_k/l,n_f) - S^A_{bound}(\Delta\alpha_k/l,n_f)$. Indeed, preliminary calculations have led to decreasing nonconverging results where $T\Delta S^A \sim -1$ kcal/mol for the best approximation, $n_f = 5000$ and $\delta = \Delta\alpha_k/45$. To obtain sets of results of S^A_{bound} which are on the same level of approximation as those of S^A_{free} we have defined (similar to the vacuum case in section III.4) a uniform set of bins for the bond angles (for both microstates) as $100°/l$, where $l = 45$, 30, 15, and 10, while the dihedral angles' bins, $\Delta\alpha_k/l$, are based on the $\Delta\alpha_k$ values of Table 6.

III.6. Entropy in Implicit Water. The computations with GB/SA are much more time-consuming than those carried out in vacuum; therefore, we performed only one set of calculations based on unit = 500 (5 ps) with $n_f = 625$, 1250,

Table 8. HSMD Results at $T = 300$ K for the Free Energy, F^A, the Potential Energy, E_{int}, Their Fluctuations, and the Differences, ΔF^A and ΔE_{int}, between the Free and Bound Microstates in Solvent[a]

	free loop		bound loop		free−bound	
n_f	$-F^A$	σ_A, σ_E	$-F^A$	σ_A, σ_E	ΔF^A	ΔE
625	939.0 (1)	3.7 (1)	913.4 (2)	3.7 (2)		−25.6 (1)
1250	937.8 (2)	3.7 (1)	912.3 (2)	3.7 (1)		−25.5 (1)
2500	937.1 (2)	3.8 (1)	911.6 (1)	3.7 (2)		−25.5 (1)
5000	936.6 (2)	3.8 (1)	911.1 (2)	3.7 (2)		−25.5 (1)
$-F^{QH}$	955 (1)		935 (1)			−21 (1)
$-F^{LS}$	966 (2)		950 (2)			−16 (3)
$-E_{int}$	868.3 (5)	4.1 (1)	842.6 (3)	4.0 (1)		−25.7 (1)

[a] F^A (eq 17) is a lower bound of the free energy, and σ_A (eq 19) is its fluctuation. The results were obtained from samples of $n = 200$ conformations for the smallest bin size, $\delta = \Delta\alpha_k/45$, unit = 5 ps, and all future sample sizes n_f. F^{QH} (see eq 5) and F^{LS} (eq 17 and section II.8) are free energies obtained by the quasi-harmonic approximation and the local states method, respectively, and are based on larger samples (see text). The average potential energy, E_{int}, of the studied samples appears in the bottom row; σ_E is the energy fluctuation (these results are in kcal/mol). All free energies are in kcal/mol and are defined up to the same additive constant for both microstates. All calculations were carried out with the AMBER force field and the implicit solvation GB/SA. The statistical error is defined in footnote *a* of Table 2.

2500, and 5000 conformations. The results for TS^A and $T\Delta S^A$ for the three smallest bin sizes appear in Table 7. For both microstates, the table shows the expected behavior, i.e., that for each bin, S^A decreases as n_f is increased and for a given n_f, S^A decreases as the bin is decreased. The results are not completely converged where the extent of convergence for the free microstate is slightly better than for the bound one. Thus, $T[S^A(\delta,n_f=2500) - S^A(\delta,n_f=5000)] \sim 0.1$, 0.4, and 0.4 kcal/mol for $\delta = \Delta\alpha_k/15$, $\Delta\alpha_k/30$, and $\Delta\alpha_k/45$, respectively, where the corresponding values for the bound microstate are 0.2, 0.3, and 0.6 kcal/mol. On the other hand, for $n_f = 5000$ $T[S^A(\delta=\Delta\alpha_k/30) - S^A(\delta=\Delta\alpha_k/45)] \sim 0.2$ for both microstates. As expected, the entropies in solvent are larger than in vacuum, where $TS^A(\Delta\alpha_k/45,n_f=5000) = 68.2$ and 68.5 kcal/mol for the free and bound microstates, respectively, in solvent, while the corresponding results in vacuum are $TS^A(\Delta\alpha_k/30,n_f=5000) = 65.8$ and 65.5 (the additive constant is assumed to be the same for both environments).

The HSMD results for the entropy are also compared in the table with those obtained using the LS and QH methods, for which larger MD samples (composed of subsamples, see section III.2) of 5000, 8000, and 10 000 conformations were generated (for each microstate) by retaining a conformation every 200 fs (100 MD steps). While both methods are expected to provide overestimations, the QH results for TS are significantly larger than the HSMD values by ∼21 and ∼24 kcal/mol for the free and bound microstate, respectively, where the LS results (based on $b=2$, $l=10$) exceed those of QH. These large QH and LS values are also affected by the significantly larger samples used for the QH and LS calculations than for HSMD (see section II.9).

III.7. Entropy Differences in Implicit Water. Table 7 also shows that the results for $T\Delta S^A = T[S^A_{free} - S^A_{bound}]$ are converged nicely to -0.4 ± 0.1 kcal/mol for *all* n_f values and bins (even for the not shown $\delta=\Delta\alpha_k/10$) (this convergence suggests that decreasing the smallest bin to $\delta=\Delta\alpha_k/45$ was not necessary). Thus, in solvent the entropy of the bound microstate is slightly larger than that of the free microstate, unlike in vacuum where this relation is reversed. Again, the QH and LS results for $T\Delta S^A$, −3(1) and −8(1) kcal/mol, respectively differ significantly from the HSMD value.

These perfectly converged results for $T\Delta S^A$ stem from an exact cancellation (see section II.10) of the systematic errors in TS for both microstates, where equal bins $\delta = 100°/l$ are used for the bond angles. This cancellation occurs for a relatively large range of approximations; thus, the (not provided) worst TS^A results for $\delta = \Delta\alpha_k/10$ differ from the best results in Table 7 by $TS^A(\Delta\alpha_k/10,n_f=500) - TS^A(\Delta\alpha_k/45,n_f=5000) = 3.4$ kcal/mol for both the free and bound microstates; as discussed in section III.4, these differences still constitute lower bounds.

The equal $T\Delta S^A$ results obtained for different n_f values suggest that the level of coverage of both microstates by the future chains during the reconstruction process is comparable and adequate, i.e., the future chains remain within these microstates. To estimate the extent of this coverage, we carried out MD simulations of the *entire* loop generating two samples (for the free and bound microstates) of 1×500 (5 ps) conformations and two samples of 10×500 [$10 \times$ (5 ps)] based on $g = 10$ fs in the same way the future chains are simulated during the reconstruction process (see sections II.6 and III.4 for the loop in vacuum). The $\Delta\alpha_k$ results (eq 13) for the dihedral angles for these samples are presented in Table 6, which shows that in most cases the results for 500 are somewhat smaller and the results for 10×500 are larger (but still close) to those of the studied samples; however, several strong deviations from this picture are also observed. Notice that the 500×10 results for χ^1(Ser), 188 and 167°, are still significantly smaller than 317 and 321° obtained for the free and bound microstates of the studied samples, respectively. However, the effect of these too small (almost equal) values is expected to get cancelled in entropy differences.

III.8. Free Energy in Implicit Water. Results for the free energy functional, F^A (eq 17), its fluctuation, σ_A (eq 19), and the energies are presented in Table 8. As in vacuum (Table 3), these results are given only for the smallest bin, $\Delta\alpha_k/45$. F^A increases slightly as n_f is increased from 2500 to 5000, while σ_A is unchanged (within the error bars). As expected, the QH and LS results for F underestimate the correct values, and the energy fluctuations are always larger than those for σ_A ($n_f=5000$). Finally, the table shows the differences in free energy, ΔF^A, and energy, ΔE, between the free and bound microstates. It is evident that the ΔF^A results are all equal within the statistical errors, and they are also equal to ΔE meaning again that the higher stability (by ∼25.5 kcal/mol) of the free microstate over the bound microstate is mostly due to ΔE.

The computer time required in solvent is significantly larger than in vacuum, where reconstructing a structure based on $n_f = 500$ requires ∼3.6 h CPU on a 2.1 GHz Athlon processor. This stems from the fact that at each MD step

GB/SA is applied to all of the ~700 atoms, while only the contributions of atoms close to the loop are expected to be affected by the conformational changes of the loop. We have not attempted to reduce the calculation time by eliminating the computation of the (constant) contribution of the atoms remote from the loop.

IV. Summary and Conclusions

As pointed out in section I.7, this study is focused mainly on theoretical and implementation aspects of HSMD as applied (for the first time) to a flexible loop of a protein; for that it has been convenient to treat initially the relatively short loop of pancreatic α-amylase which consists of small residues. The role of this loop in the enzymatic function of pancreatic α-amylase is presently of secondary interest and will be discussed in future studies where the ligand, which interacts with the loop in the bound state (and is missing in the free protein), will be considered, and explicit water—the preferred solvation model—will be introduced.

Still, the relatively large energy (hence free energy) differences between the free and bound microstates (~38 and ~25 kcal/mol in vacuum and solvent, respectively) suggest that the bound microstate would not be visited by the loop in the free protein, i.e., the response of the loop to ligand binding is probably an induced fit rather than a selected fit (see sections I.1, I.7, and II.1). The higher free energy of the bound microstate stems mainly from electrostatic interactions that are contributed by many of the loop atoms (rather a specific one). Thus, while the crystal structures of 1pif and 1pig are similar, they still differ in the structural arrangement of specific side chains, which leads to (relatively) unfavorable electrostatic interactions between the (1pif) template and the bound loop structure (which is superimposed on 1pif). Indeed, the crystal structure of 1pig is significantly better resolved than that of 1pif, where atoms with elevated B-factors (larger than 40) appear in 61 and 153 residues (predominantly charged and polar) of these structures, respectively; also, on average, the B-factors of 1pif are significantly larger than those of 1pig. The unstable MD trajectory obtained initially for the bound microstate is a result of the unfavorable electrostatic interactions, which has led us to generate a bound sample consisting of short trajectories (see section II.1).

In this context we would like to discuss further our result, $S_{\text{free}} \sim S_{\text{bound}}$. Thus, in the presence of the ligand in the active site one would expect $S_{\text{free}} > S_{\text{bound}}$ in accord with the measured B-factors. However, (as discussed above) for both models studied (where the ligand is missing) the energy of the free microstate is significantly lower than that of the bound microstate which suggests that S_{free} should be significantly lower than S_{bound}. The unexpected result, $S_{\text{free}} \sim S_{\text{bound}}$ (also obtained approximately with the quasi-harmonic method) might be attributed to the fact that our bound sample consists of several partial (relatively short) MD samples all starting from the X-ray structure of the bound protein with different sets of velocities, which leads to a relatively concentrated sample of low entropy (see section III.1).

This discussion demonstrates the problems involved in the computational definition of a microstate—a topic that has been ignored to a large extent in the literature but has been given a great deal of thought in this paper. In particular, we have provided strong theoretical arguments that systematic errors in S^{A}(HSMD) for different microstates are comparable and thus get cancelled in differences, ΔS^{A} — our main interest. This means that one can apply highly crude approximations (i.e., small reconstruction samples) decreasing computer time dramatically. Indeed, such cancellation has been observed for peptides[55] and for the loop of α-amylase modeled by the AMBER force field[68] and AMBER with the implicit solvation GB/SA,[69] leading to efficient computations and providing support to our theory. Notice that calculating transition probabilities for different steps k is completely independent, and the reconstruction process is thus completely parallelized. As for peptides, the small statistical errors in $T\Delta S^{\text{A}}$ of 0.1−0.2 kcal/mol is very satisfactory.

An important development has been the realization that the bins can be optimized, leading to improved (i.e., smaller) results for S^{A} hence to reliable results for $T\Delta S^{\text{A}}$ for smaller reconstruction samples. We have not carried out a full bins' optimization but have demonstrated its potential effectiveness by applying to both microstates a uniform set of bins, $\delta =$ constant/l for the bond angles. It is plausible to assume that further bins' optimization would lead to an improved probability density, $\rho^{\text{HS}}(\alpha_K, \cdots, \alpha_1)$ (eq 15), hence to more accurate free energy functionals, F^{D} (eq 22) and F^{B} (eq 21), where the latter exhibits an upper bound behavior. It has also been shown that the contributions of the Jacobian are cancelled out in entropy and free energy differences. The quasi-harmonic approximation and the local states method (as expected) overestimate the entropy but more significantly than for peptides,[53,55] which might reflect strong long-range correlations and anharmonic effects within the loop due to the loop-template interactions.

The theoretical developments introduced in this paper and the conclusions gathered from the application of HSMD to the 7-residue loop of pancreatic α-amylase constitute a mandatory basis for the next step in the development of HSMD—its extension to the same loop modeled by the AMBER force field and explicit water. In this treatment the loop is capped with explicit water,[83] and the entropy is calculated from the sample in a two-stage process, where the loop structure i is reconstructed first leading to S_i (the surrounding waters, which constitute part of the future, are moved as well during the reconstruction of i); next, the water configuration is reconstructed step-by-step in the presence of the frozen loop structure i leading to $S_{\text{w}/i}$. One is interested to estimate $\langle S_i \rangle$ and $\langle S_i + S_{\text{w}/i} \rangle$, which constitute measures of flexibility, and their values for the free and bound microstates can be compared (unlike the energy and free energy that due to the ligand depend on different sets of interactions in the free and bound proteins). This study is being carried out presently. After completing these developmental stages HSMD will become a mature tool for studying other flexible loops of interest, such as the 11-residue lid loop of TIM proteins, and problems which require calculating the relative and absolute free energy of binding.

Acknowledgment. We thank Dr. Ron P. White for his contributions in the initial stage of this work. This work was supported by NIH grant 2-R01 GM066090-4 A2.

References

(1) Alder, B. J.; Wainwright, T. E. *J. Chem. Phys*. **1959**, *31*, 459.

(2) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585.

(3) Elber, R.; Karplus, M. *Science* **1987**, *235*, 318.

(4) Stillinger, F. H.; Weber, T. A. *Science* **1984**, *225*, 983.

(5) Getzoff, E. D.; Geysen, H. M.; Rodda, S. J.; Alexander, H.; Tainer, J. A.; Lerner, R. A. *Science* **1987**, *235*, 1191.

(6) Rini, J. M.; Schulze-Gahmen, U.; Wilson, I. A. *Science* **1992**, *255*, 959.

(7) Constantine, K. L.; Friedrichs, M. S.; Wittekind, M.; Jamil, H.; Chu, C. H.; Parker, R. A.; Goldfarb, V.; Mueller, L.; Farmer, B. T. *Biochemistry* **1998**, *37*, 7965.

(8) Kessler, H.; Matter, H.; Gemmecker, G.; Kottenhahn, M.; Bates, J. W. *J. Am. Chem. Soc.* **1992**, *114*, 4805.

(9) Baysal, C.; Meirovitch, H. *Biopolymers* **1999**, *50*, 329.

(10) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586.

(11) Eisenmesser, E. Z.; Millet, O.; Labeikovski, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky J. J.; Kay, L. E.; Kern D. *Nature* **2005**, *438*, 117.

(12) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.

(13) Beveridge, D. L.; DiCapua, F. M. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431.

(14) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395.

(15) Jorgensen, W. L. *Acc Chem. Res*. **1989**, *22*, 184.

(16) Meirovitch, H. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1998; Vol. 12, p 1.

(17) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047.

(18) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.

(19) Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, *17*, 181.

(20) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345.

(21) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249.

(22) Ikeda, K.; Galzitskaya, O. V.; Nakamura, H.; Higo, J. *J. Comput. Chem.* **2003**, *24*, 310.

(23) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C. K.; Li, M. S. *Proteins* **2005**, *61*, 795.

(24) Lange, O. F.; Grubmüller, H. *J. Chem. Phys.* **2006**, *124*, 214903.

(25) MacDonald, I. R.; Singer, K. *J. Chem. Phys*. **1967**, *47*, 4766.

(26) Hansen, J.-P.; Verlet, L. *Phys. Rev.* **1969**, *184*, 151.

(27) Hoover, W. G.; Ree, F. H. *J. Chem. Phys*. **1967**, *47*, 4873.

(28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarenden Press: Oxford, 1987.

(29) Kirkwood, J. G. *J. Chem. Phys*. **1935**, *3*, 300.

(30) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

(31) Squire, D. R.; Hoover, W. G. *J. Chem. Phys*. **1969**, *50*, 701.

(32) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett*. **1974**, *28*, 578.

(33) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.

(34) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.

(35) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett*. **1989**, *63*, 1195.

(36) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kolmann, P. A. *J. Comput. Chem*. **1995**, *16*, 1339.

(37) Kumar, S.; Payne, P. W.; Vásquez, M. *J. Comput. Chem*. **1996**, *17*, 1269.

(38) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.

(39) Gō, N.; Scheraga, H. A. *J. Chem. Phys.* **1969**, *51*, 4751.

(40) Gō, N.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 535.

(41) Hagler, A. T.; Stern, P. S.; Sharon, R.; Becker, J. M.; Naider, F. *J. Am. Chem. Soc.* **1979**, *101*, 6842.

(42) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325.

(43) Chang, C. E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017.

(44) Meirovitch, H. *Chem. Phys. Lett.* **1977**, *45*, 389.

(45) Meirovitch, H.; Vásquez, M.; Scheraga, H. A. *Biopolymers* **1987**, *26*, 651.

(46) Meirovitch, H.; Koerber, S. C.; Rivier, J.; Hagler, A. T. *Biopolymers* **1994**, *34*, 815.

(47) Meirovitch, H. *Phys. Rev. A* **1985**, *32*, 3709.

(48) Meirovitch, H.; Scheraga, H. A. *J. Chem. Phys*. **1986**, *84*, 6369.

(49) Meirovitch, H. *J. Chem. Phys.* **2001**, *114*, 3859.

(50) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2004**, *121*, 10889.

(51) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2006**, *124*, 204108.

(52) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2005**, *123*, 214908.

(53) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2005**, *122*, 054903.

(54) Cheluvaraja, S.; Meirovitch, H. *J. Phys. Chem. B* **2005**, *109*, 21963.

(55) Cheluvaraja, S.; Meirovitch, H. *J. Chem. Phys.* **2006**, *125*, 024905.

(56) Qian, M.; Haser, R.; Payan, F. *J. Mol. Biol*. **1993**, *231*, 785.

(57) Qian, M.; Haser, R.; Buisson, G.; Duee, E.; Payan, F. *Biochemistry* **1994**, *33*, 6284.

(58) Qian, M.; Haser, R.; Payan, F. *Protein Sci.* **1995**, *4*, 747.

(59) Machius, M.; Vertesy, L.; Huber, R.; Wiegand, G. *J. Mol. Biol*. **1996**, *260*, 409.

(60) Brayer, G. D.; Sidhu, G.; Maurus, R.; Rydberg, E. H.; Braun, C.; Wang, Y. et al. *Biochemistry* **2000**, *39*, 4778.

(61) Rydberg, E. H.; Li, C.; Maurus, R.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 4492.

(62) Numao, S.; Maurus, R.; Sidhu, G.; Wang, Y.; Overall, C. M.; Brayer, G. D.; Withers, S. G. *Biochemistry* **2002**, *41*, 215.

(63) Steer, M. L.; Levitzki, A. *FEBS Lett.* **1973**, *31*, 89.

(64) Levitzki, A.; Steer, M. L. *Eur. J. Biochem.* **1974**, *41*, 171.

(65) Aghajari, N.; Feller, G.; Gerday, C.; Haser, R. *Protein Sci.* **2002**, *11*, 1435.

(66) Brayer, G. D.; Luo, Y.; Withers, S. G. *Protein Sci.* **1995**, *4*, 1730.

(67) Ramasubbu, N.; Paloth, V.; Luo, Y.; Brayer, G. D.; Levine, M. J. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1996**, *52*, 435.

(68) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(69) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem.* **1997**, *101*, 3005.

(70) Ponder, J. W. *TINKER - software tools for molecular design, version 3.9*; Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 2001.

(71) Meirovitch, H.; Alexandrowicz, Z. *J. Stat. Phys.* **1976**, *15*, 123.

(72) Meirovitch, H. *J. Chem. Phys.* **1999**, *111*, 7215.

(73) Szarecka, A.; White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2003**, *119*, 12084.

(74) White, R. P.; Meirovitch, H. *J. Chem. Phys.* **2003**, *119*, 12096.

(75) Meirovitch, H.; Meirovitch, E. *J. Phys. Chem.* **1996**, *100*, 5123.

(76) Meirovitch, H.; Hendrickson, T. F. *Proteins* **1997**, *29*, 127.

(77) Baysal, C.; Meirovitch, H. *Biopolymers* **2000**, *53*, 423.

(78) Meirovitch, H. *J. Chem. Phys.* **1988**, *89*, 2514.

(79) Meirovitch, H.; Vásquez, M.; Scheraga, H. A. *Biopolymers* **1988**, *27*, 1189.

(80) Brady, J.; Karplus, M. *J. Am. Chem. Soc.* **1985**, *107*, 6103.

(81) Gibbs, W. *Elementary Principles in Statistical Mechanics*; Yale University Press: 1902; Chapter XI.

(82) White, R. P.; Meirovitch, H. *J. Chem. Theory Comput.* **2006**, *2*, 1135.

CT700116N