Relative stability of the open and closed conformations of the active site loop of streptavidin

Ignacio J. General and Hagai Meirovitch[a)]
Department of Computational and Systems Biology, University of Pittsburgh School of Medicine, 3059 BST3, Pittsburgh, Pennsylvania 15260, USA

(Received 23 August 2010; accepted 5 November 2010; published online 14 January 2011)

The eight-residue surface loop, 45–52 (Ser, Ala, Val, Gly, Asn, Ala, Glu, Ser), of the homotetrameric protein streptavidin has a "closed" conformation in the streptavidin-biotin complex, where the corresponding binding affinity is one of the strongest found in nature ($\Delta G \sim -18$ kcal/mol). However, in most of the crystal structures of apo (unbound) streptavidin, the loop conformation is "open" and typically exhibits partial disorder and high B-factors. Thus, it is plausible to assume that the loop structure is changed from open to closed upon binding of biotin, and the corresponding difference in free energy, $\Delta F = F_{\text{open}} - F_{\text{closed}}$ *in the unbound protein,* should therefore be considered in the total *absolute* free energy of binding. ΔF (which has generally been neglected) is calculated here using our "hypothetical scanning molecular-dynamics" (HSMD) method. We use a protein model in which only the atoms closest to the loop are considered (the "template") and they are fixed in the x-ray coordinates of the free protein; the x-ray conformation of the closed loop is attached to the same (unbound) template and both systems are capped with the same sphere of TIP3P water. Using the force field of the assisted model building with energy refinement (AMBER), we carry out two *separate* MD simulations (at temperature $T = 300$ K), starting from the open and closed conformations, where only the atoms of the loop and water are allowed to move (the template-water and template-loop interactions are considered). The *absolute* F_{open} and F_{closed} (of loop + water) are calculated from these trajectories, where the loop and water contributions are obtained by HSMD and a thermodynamic integration (TI) process, respectively. The combined HSMD-TI procedure leads to total (loop + water) $\Delta F = -27.1 \pm 2.0$ kcal/mol, where the entropy $T\Delta S$ constitutes 34% of ΔF, meaning that the effect of S is significant and should not be ignored. Also, ΔS is positive, in accord with the high flexibility of the open loop observed in crystal structures, while the energy ΔE is unexpectedly negative, thus also adding to the stability of the open loop. The loop and the 250 capped water molecules are the largest system studied thus far, which constitutes a test for the efficiency of HSMD-TI; this efficiency and technical issues related to the implementation of the method are also discussed. Finally, the result for ΔF is a prediction that will be considered in the calculation of the absolute free energy of binding of biotin to streptavidin, which constitutes our next project.
 [doi:10.1063/1.3521267]

I. INTRODUCTION

An important objective of this paper is to further develop our method, the hypothetical scanning molecular dynamics (HSMD) for calculating the absolute entropy, S, and the absolute Helmholtz free energy, F ($F = E\text{-}TS$, where E is the energy and T is the absolute temperature). Calculation of these fundamental thermodynamic quantities is extremely difficult, in spite of the significant progress that has been made in the last 50 years.[1–10] Calculation of F is in particular challenging in structural biology due to the flexibility and strong long-range interactions characterizing bio-macromolecules such as proteins. Thus, the potential energy surface of a protein, $E(\mathbf{x})$, is rugged ($\mathbf{x}$ is the $3N$-dimensional vector of the Cartesian coordinates of the molecule's N atoms), i.e., it is "decorated" by a tremendous number of localized wells and "wider" wells (called microstates) defined over regions, Ω_m, with each wider well consisting of many localized ones. A microstate Ω_m (e.g., the α-helical region of a peptide), 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)[11,12] simulation (see further discussions in Refs. 13 and 14). A molecule will visit a localized well for a very short time [several femtoseconds (fs)] while staying much longer within a microstate,[15,16] which is therefore of a greater physical significance. A central aim in protein folding is the daunting task of finding the most stable microstate, i.e., that with the lowest free energy, $F_m = -k_B T \ln Z_m$, where k_B is the Boltzmann constant, and the partition function Z_m is integrated over Ω_m (rather than over the entire space).

In addition to the difficult problem of protein folding (where interest is typically in a single microstate), more manageable problems are commonly studied, where smaller systems are involved (e.g., cyclic peptides), or the focus is on

a)Author to whom correspondence should be addressed. Electronic mail: hagaim@pitt.edu, Tel.: 412-648-3338.

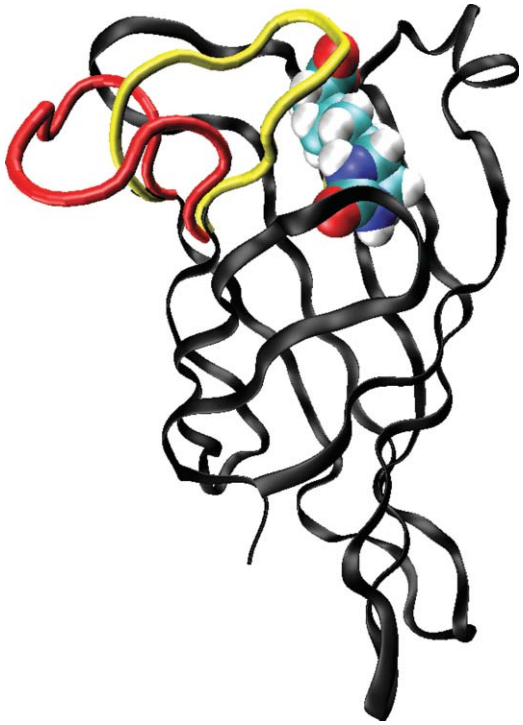

FIG. 1. The conformations of the closed loop (yellow tube) and the open loop (red tube) attached to a single monomer of streptavidin (black ribbon) complexed with biotin. The significant difference between these structures is evident.

specific protein regions, such as *flexible* surface loops, side chains, or bound ligands. These limited systems can populate significantly *several* microstates, Ω_m in thermodynamic equilibrium, which should be identified, and their populations, $p_m = \exp(-F_m/k_BT)$, calculated. It is of interest to know whether the conformational change adopted by a loop (a side chain, ligand, etc.) upon ligand binding has been induced by the ligand (induced fit[17,18]) or alternatively whether the free loop interconverts among different microstates, one of which is selected upon binding (selected fit[19]). (Notice again that not only is the calculation of p_m difficult, but defining a microstate in the high-dimensional conformational space is also not straightforward.) The topic of the present paper is related to this category of problems, as described later in the paper.

Thus, our central aim is to study the relative stability of the "open" and "closed" conformations (microstates) of a mobile loop existing in each chain of the homotetrameric protein streptavidin (see Fig. 1); this loop is found in the "closed" conformation in the streptavidin-biotin complex. To study the loop flexibility and stability, we carry out MD simulations of the protein in water and calculate the differences in free energy and entropy (ΔF_{mn} and ΔS_{mn}, respectively) between these microstates (denoted m and n). Notice, however, that carrying out such calculations with thermodynamic integration (TI) would be practically unfeasible if the structural variance between m and n were significant. Therefore, alternatively (and as in previous work) we use our HSMD method mentioned earlier,[20–25] which enables one to calculate the *absolute* S and F. Thus, only two (separate) *local* MD simulations of the loop in m and n are carried out, from which F_m, F_n, and $\Delta F_{mn} = F_m - F_n$ (and $\Delta S_{mn} = S_m - S_n$) are obtained *directly* and the integration process is avoided. (Other methods for calculating the absolute F and S are reviewed, for example, in Ref. 7.)

With HSMD (or HSMC when Monte Carlo replaces MD), each conformation of a sample is reconstructed step-by-step (from nothing) using transition probabilities (TPs); the product of these TPs leads to an approximation P_i for the correct Boltzmann probability P_i^B, where from P_i various free-energy functionals can be defined. While the TPs of HSMC(D) are stochastic in nature (calculated by MD or MC simulations), all the system interactions are taken into account (from now on, for simplicity, we shall omit in most cases the letters MC); in this respect, HSMD can be viewed as exact,[20] where the only approximation involved is due to insufficient MD sampling for calculating the TPs. HSMD has unique features: it provides rigorous lower and upper bounds for F, which enable one to determine the accuracy from HSMD results alone without the need to know the correct answer. Furthermore, F can be obtained from a very small sample and in principle even from any single conformation (e.g., see the results for argon in Ref. 20). The HSMC(D) methodology has been developed systematically, as applied to systems of increasing complexity. The initial (HSMC) calculations of liquid argon, TIP3P water,[20] self-avoiding walks,[22] and polyglycine molecules[23] have verified the validity of the theoretical predictions stated above by comparisons with accurate results obtained by other well-established techniques; a subsequent application of HSMD to peptides has led to an ~100 times reduction in computer time.[25]

HSMD has also been applied to mobile loops which change their structure due to ligand binding. We studied initially the seven-residue surface loop, 304–310 (Gly-His-Gly-Ala-Gly-Gly-Ser), of the enzyme porcine pancreatic α-amylase in vacuum and in the generalized Born surface area (GB/SA) implicit solvent.[13] In a subsequent paper, the loop was capped with 70 TIP3P water molecules and an HSMD-TI procedure was developed in which the contribution of water to F is calculated by a TI process, which is more efficient than HSMD.[14] Subsequently, HSMD-TI was applied to the loop 287–290 with the bulky residues, Ile, Phe, Arg, and Phe of acetylcholinesterase (AChE) from *Torpedo California*;[26] reaction of AChE with the inhibitor di*iso*propylphosphorofluor*i*date (DFP) leads to a displacement of the loop's backbone by roughly 4 Å, and experimental evidence suggests that the free-energy penalty for the loop displacement is on the order of 4 kcal/mol (i.e., $F_{\text{free}} - F_{\text{bound}} \sim -4$ kcal/mol). Therefore, AChE has been an ideal system for checking the performance of HSMD-TI, and in particular for examining the minimal number of water molecules needed to cap the loop. We have found that to recover the experimental free-energy difference, the water density should be close to that of bulk water, and our results, $F_{\text{free}} - F_{\text{bound}} = -3.1 \pm 2.5$ and -3.6 ± 4 kcal/mol for a sphere containing 160 and 180 waters, are equal within error bars to the experimental value.[26]

The present study constitutes an additional step in the application of HSMD-TI to mobile loops. Streptavidin (isolated from the bacterium *Streptomyces avidinni*) is a tetrameric protein consisting of 159 residues per chain, where each subunit

can bind noncovalently one biotin molecule ($C_{10}H_{16}N_2O_3S$). This binding (characterized first by Chaiet and Wolf[27]) is of extraordinarily high affinity ($K_a \sim 10^{13}$ M^{-1}, $\Delta G \sim -18$ kcal/mol)—one of the strongest found in nature.[28] It is a property that has been exploited to devise powerful tools for affinity chromatography, biochemical assays, and many other application.[29–32] Biotin binds to the open end of the β barrel of each subunit of streptavidin, and an eight-residue surface loop (45–52) (Ser, Ala, Val, Gly, Asn, Ala, Glu, Ser) folds so as to cap the barrel (Fig. 1). This loop, which usually has an "open" structure in apo streptavidin, has been identified before as a major factor in the affinity of streptavidin to biotin.[31] Deletion of the loop via circular mutation decreases the binding affinity by approximately 10 kcal/mol.[33]

The x-ray structures of streptavidin complexed with biotin were first determined in 1989 by Hendrickson *et al.*[34] and Weber *et al.*,[31] where the latter study also provided the structure of apo streptavidin and its comparison with the complexed one. Today the protein data bank (PDB) contains 134 crystal structures of wild-type and mutated streptavidin, complexed with biotin and other ligands. To calculate the free-energy difference between the closed and open microstates of the loop, one needs to know the corresponding crystal structures. However, the open structure in apo streptavidin is (in most cases) partially disordered and the closed structure appears with a bound biotin. We were able to overcome these hurdles by suitably matching structures of the loop and protein taken from a set of crystal structures of the free and complexed streptavidin obtained by Freitag *et al.*,[30] as described in detail in the next section.

Finally, it should be noted that the eight-residue loop and the 250 water molecules capping it constitute the largest system treated by HSMD-TI thus far. Testing the performance of HSMC-TI for systems of increasing size (N) is important due to the $N^{1/2}$ increase in the fluctuation of the absolute entropy and energy. In this paper, we suggest and test new reconstruction procedures for a loop. Our results shed more light on the somewhat unusual structural behavior of the loop of streptavidin exhibited in crystal structures. Also, the free-energy difference between the closed- and open-loop microstates (in the apo protein) should be taken into account in the calculation of the absolute free energy of binding of biotin to streptavidin. Thus, the present study is the first step in our long-range goal of developing HSMD-TI as a tool for calculating the absolute free energy of binding, where we intend to apply HSMD-TI initially to the streptavidin-biotin and avidin-biotin complexes.

II. THEORY AND METHODOLOGY

A. Definition of the system

As pointed out earlier, we study the surface loop of $N_{res} = 8$ residues (45–52) (Ser, Ala, Val, Gly, Asn, Ala, Glu, Ser) of streptavidin. While the loop exists in each of the four subunits, HSMD-TI is applied only to a single loop of one subunit of the *unbound* protein. However, even such a limited treatment might be problematic because in the unbound structures of streptavidin which appear in the PDB, the (open) loop conformations are typically disordered, as can also be learned from the five x-ray structures (I–V) obtained by Freitag *et al.*[30] under various crystallization conditions. Indeed, in structures I–IV there are 14 unbound subunits with 12 loops in the open conformation, where for 10 of these loops at least three residues could not be traced in the electron density maps and the other residues appear with elevated B factors. Only in structure III (PDB 1swc) were Freitag *et al.* able to resolve the free loop conformations of subunits 2 and 4 with average B factors of 33.8 and 48.6 $Å^2$, respectively. It should also be noted that the loop of subunit 1 of the unbound structures I and II appears in the *closed* conformation, a fact that is attributed by Freitag *et al.* to crystal-packing interactions; however, the existence of closed conformations might suggest that the structural preference of the loop in unbound streptavidin is somewhat uncertain or that the loop responds to binding by a selected fit process.

We have a special interest in structure IV (1swd), in which two biotin molecules are bound to subunits 1 and 4; as expected, the corresponding loops are closed, and the loops of the unbound subunits 2 and 3 have partially disordered open conformations. We decided to use subunit 2 of structure IV as a basis to which initial open- and closed-loop structures will be attached for further optimization and MD simulations. First, we deleted the original loop coordinates of subunit 2 of IV and attached instead the closed-loop conformation of subunit 1 of IV by superimposing the structure of subunit 1 on subunit 2 (both of IV); similarly, the open-loop conformation of subunit 2 of III (1swc) (which has lower B-factors than the loop of subunit 4 of III) was attached to the same (subunit) basis to create an initial open-loop conformation; we denote these two tetramer structures as A and B for the closed and open loops, respectively. In both cases, the incomplete open loop in subunit 3 was also replaced by the open loop in subunit 2 of structure III. Notice that in the crystal structure of IV, only the coordinates of the (core) residues, 16–133, are provided.

Before discussing the various stages of system optimization and simulations, it should be pointed out that all calculations were performed with the AMBER99 force field[35] where the amino acids Lys, Arg, Glu, Asp, and His are charged. The MD simulations were carried out in the *NVT* ensemble where some of the preliminary results of the tetramer were obtained with the AMBER 10 package;[36] however, implementation of HSMD-TI is more convenient with the TINKER 5.0 program,[37] therefore it was used in all of the free-energy calculations. The temperature was kept around 300 K with the Berendsen thermostat based on a time constant of 1.0 ps.[38] The time step employed in the MD simulations was 2 fs, applying the RATTLE algorithm to constrain bonds involving hydrogen atoms.[38] No periodic boundary conditions or cutoffs were used unless otherwise stated. The TIP3P model for water was used.[39]

B. Structural optimization for the free-energy calculations

We carried out two sets of MD simulations. In one set, which is aimed at testing the stability of the loop, the

entire tetramer (soaked in water) was treated (see Sec. III A). However, for calculating the free energy of the loop, only part of the protein was considered; below we define this partial system and describe its optimization, which constitutes the starting point for a set of simulations described in Sec. III B.

The structural optimization started from structure *B* (1swd with the open-loop conformation attached to subunits 2 and 3). This structure was solvated with TIP3P water molecules in a sphere of radius $R_{\text{water}} = 15$ Å around the center of the loop of subunit 2; to hold these water molecules around the loop, they were restrained with a flat-welled half-harmonic potential [a force constant of 10 kcal/(mol Å^2)] based on the distance from the "center" of the loop region. That is, the distance of each water molecule (in practice, the oxygen atom) is measured from the restraining center. If this distance is greater than 15 Å, a harmonic restoring force is applied; otherwise, the restraining force is zero. Also, harmonic restraints with a force constant of 5 kcal/(mol Å^2) were applied to the heavy atoms of the protein to eliminate bad atomic overlaps and strains in the original structure, while still keeping the atoms reasonably close to the PDB coordinates. The system was energy-minimized using 10^4 steps of steepest descent followed by 10^4 steps of conjugate gradient. The root-mean-square deviation (RMSD) values between the PDB coordinates of the structure without the loops before and after minimization are 0.15 Å for the backbone and 0.29 Å for the backbone and side chains, meaning that most of the change is due to the side-chain motion.

As in our previous HSMD-TI applications to loops, to reduce computer time we do not consider the entire protein but only a spherical part of it (the template) which is the closest to the loop and whose coordinates are held fixed during future simulations of the loop. To define this system, we first removed the water molecules from the minimized tetramer configuration obtained previously and defined a spherical template with a radius of 18 Å, centered at the middle point on the line connecting the α-carbon atoms of the first and last residues of the loop attached to subunit 2 of 1swd. Thus, if the distance of any atom of a residue from the middle point is less than 18 Å, the entire residue is included in the template; otherwise, the residue is eliminated. This "cutting" procedure reduced the number of protein atoms considered from 6805 to 957, representing a significant gain in simulation speed. We should note that most of the residues in the spherical cut belong to subunit 2, but some correspond to other subunits. Nevertheless, it is important to keep these extra residues because through visualization of the PDB files it became clear that they form a "wall" over the β barrel of subunit 2 which prevents the loop from moving to regions that are not available in the case of the whole tetramer.

Next, the loop was solvated with 250 TIP3P water molecules, distributed in a spherical cap with a radius of $R_{\text{water}} = 15$ Å. Initially we used the same spherical center defined above for the template, but have found our template to be too "thin," i.e., during MD simulations water molecules could "seep" through cavities in the template or around it to its "back side." To avoid this undesired situation, we shifted the water center by 3.3 Å toward the "loop side" of the loop-template system, which was found to be the smallest distance necessary to prevent this effect. [Such a shift would not be needed for larger templates like that defined for the larger (single chain) protein AChE of 535 residues;[26] see also Refs. 14 and 40]. The number of water molecules in the cap (250) was determined by the condition that their density is close to the bulk density of water in normal conditions of pressure and temperature (0.0350 Å^{-3}); the actual density, 0.0355 Å^{-3}, is indeed close to the target value. (Notice that the volume in which the waters move is not the full spherical volume, but a smaller one due to the presence of the loop and part of the template; it is about half the total volume.) The water molecules were restrained to stay in the spherical cap by a harmonic force with a force constant of 10 kcal/(mol Å^2) as described previously.

In this system, the template coordinates are always held fixed, i.e., only the loop and water atoms are allowed to move (no harmonic restraints were applied to the loop). Thus, the total potential energy E_{total} is a sum of partial energies (the *constant* template-template energy is ignored),

$$\begin{aligned} E_{\text{total}} &= (E_{\text{loop-loop}} + E_{\text{loop-tmpl}}) \\ &\quad + (E_{\text{water-water}} + E_{\text{water-tmpl}} + E_{\text{water-loop}}) \\ &= E_{\text{loop}} + E_{\text{water}}, \end{aligned} \tag{1}$$

where $E_{\text{loop-loop}}$ is the intra loop energy and $E_{\text{loop-tmpl}}$ is the energy due to the loop-template interactions; these energies define the total loop energy, E_{loop}. The interactions related to water are defined in a similar way, where their total is denoted by E_{water}. This system was energy-minimized using the same procedure mentioned before, and equilibrated in a 500 ps MD simulation. This optimization procedure was also applied to structure *A* (1swd with its open subunit 2 loop replaced by its closed subunit 1 loop). Hence, both loops share exactly the same "frozen" template and the same 15 Å water sphere with 250 water molecules. The last loop/water configurations of the 500 ps simulations for the open and closed loops become starting configurations for 2 ns "production" runs from which the free energy is calculated (see Secs. III B and III C).

One might consider our model to be limited as it is based on a partial frozen template (which reduces the system size, keeping the fluctuations of E and S manageable). However, this model is expected to be adequate, as a recent 250 ns MD simulation of the bound streptavidin tetramer in solution by Cerutti *et al.*[41] has found the backbone of the 67 core residues of each subunit to be very stable, i.e., with RMSD values smaller than 0.7 Å; stronger fluctuations were observed there only in the loop regions. Also, as pointed out earlier, our free-energy results for the loop of AChE based on a frozen template are in a very good agreement with experiment.[26] Since HSMD-TI can also treat a fluctuating template, we intend to test it in future studies, initially for a template whose heavy atoms are restrained by harmonic forces.

C. Statistical mechanics of a loop in internal coordinates

The theory of HSMD-TI has been developed in previous publications; therefore, we describe it briefly providing mainly equations that are related directly to the calculations.

The reconstruction of the loop structure (of $N_{\text{res}} = 8$ residues) is carried out in internal coordinates; therefore, the loop conformations simulated by MD are transferred from Cartesians to the dihedral angles φ_i, ψ_i, and ω_i ($i = 1, N_{\text{res}}$), the bond angles $\theta_{i,l}$ ($i = 1, N_{\text{res}}$, $l = 1, 3$), the side-chain angles χ, and the corresponding bond angles. For convenience, all these angles (ordered along the backbone) are denoted by α_k, $k = 1, K$; as discussed in Sec. II G, in this work we define two reconstruction procedures: one is based on $K = 64$ and the other on $K = 104$. We have argued in Refs. 13 and 25 that to a good approximation, bond stretching can be ignored, thus the bond lengths are considered to be constant.

The partition function of the loop/water/template system is

$$Z_m = \int_m \exp[-E(\mathbf{x}_{\text{loop}}, \mathbf{x}^N)/k_{\text{B}}T]d\mathbf{x}_{\text{loop}}d\mathbf{x}^N, \tag{2}$$

where $E(\mathbf{x}_{\text{loop}}, \mathbf{x}^N) = E_{\text{total}}$ is defined in Eq. (1), $\mathbf{x}_{\text{loop}}$ are the Cartesian coordinates of the loop in microstate m, and $\mathbf{x}^N$ are the $9N_{\text{water}}$ Cartesian coordinates of the water molecules; E_{total} also depends on the "frozen" template coordinates, which are omitted for simplicity. For the same reason, the letter m will be omitted in most of the equations and N_{water} will be replaced in the theoretical section by N ($N = N_{\text{water}}$). After changing the variables of integration from $\mathbf{x}_{\text{loop}}$ to internal coordinates, the integral becomes a function of the K dihedral and bond angles, α_k, $k = 1, \ldots, K$, and a Jacobian $\cos(\theta_{i,l})$ that depends only on each of the bond angles $\theta_{i,l}$,[42–44]

$$Z = \int_m \exp\{[-E_{\text{loop}}([\alpha_k]) - E_{\text{water}}([\alpha_k], \mathbf{x}^N)]/k_{\text{B}}T\}d[\alpha_k]d\mathbf{x}^N, \tag{3}$$

where $[\alpha_k] = [\alpha_1, \ldots, \alpha_K]$ and $d[\alpha_k] = d\alpha_1 \ldots d\alpha_K$. In Eq. (3), we have omitted a factor that depends on the bond lengths and is *assumed* to be the same (i.e., constant) for different microstates of the same loop and therefore does not affect entropy *differences* (e.g., see the discussion in Ref. 26). The Jacobian is also omitted for simplicity because we have shown[13] that it cancels out (within the error bars) in entropy and free-energy differences (this conclusion was checked again and verified in the present study). The Boltzmann probability density corresponding to Z [Eq. (3)] is

$$\rho^{\text{B}}([\alpha_k], \mathbf{x}^N) = \exp\{-E([\alpha_k], \mathbf{x}^N)/k_{\text{B}}T\}/Z, \tag{4}$$

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

$$S = -k_{\text{B}} \int_m \rho^{\text{B}}([\alpha_k], \mathbf{x}^N) \ln \rho^{\text{B}}([\alpha_k], \mathbf{x}^N)d[\alpha_k]d\mathbf{x}^N \tag{5}$$

and

$$F = \int_m \rho^{\text{B}}([\alpha_k], \mathbf{x}^N)\{E([\alpha_k], \mathbf{x}^N) + k_{\text{B}}T \ln \rho^{\text{B}}([\alpha_k], \mathbf{x}^N)\}d[\alpha_k]d\mathbf{x}^N. \tag{6}$$

It should be noted that the fluctuation of the *exact* F is zero,[45,46] because by substituting $\rho^{\text{B}}([\alpha_k])$ [Eq. (4)] inside the curly brackets of Eq. (6), one obtains $E([\alpha_k]) + k_{\text{B}}T \ln \rho^{\text{B}}([\alpha_k]) = -kT \ln Z = F$, i.e., the expression in the curly brackets is constant and equal to F 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. [Notice, however, that the calculation of $\rho^{\text{B}}([\alpha_k], \mathbf{x}^N)$ for a single conformation depends on the entire microstate, as is also evident from the HSMC(D) procedure discussed later.] Still, the fluctuation of an approximate free energy (i.e., one based on an approximate probability density) is finite and is expected to decrease as the approximation improves.[45,46,20–22] Because HSMC(D) provides an approximation for $\rho^{\text{B}}([\alpha_k], \mathbf{x}^N)$, one can, *in principle*, estimate the free energy of the system from *any single* structure.[20–22] This is the reason why in practice reliable HSMC(D) results for F (but not necessarily for S and E) can be obtained from a relatively small sample.

D. Exact stochastic future scanning procedure

It should first be pointed out that the Metropolis Monte Carlo (MC) and MD are exact *dynamical* methods that enable one to sample system configuration i correctly with its Boltzmann probability, P_i^{B}, while the *value* of P_i^{B} is not provided (due to the dynamical character of these methods). [To simplify the discussion, we use the probability P_i^{B} rather than the probability density $\rho^{\text{B}}([\alpha_k], \mathbf{x}^N)$ defined in Eq. (4).] Thus, properties such as the energy that are "written" on i can easily be calculated, while a *direct* calculation of the absolute S is difficult because $\ln P_i^{\text{B}}$ is unknown (it depends not only on i but on the entire ensemble through the partition function Z, which cannot be obtained from a finite sample).

Unlike the dynamical MC and MD, the exact future scanning method, which constitutes the basis of HSMC(D), is a growth procedure that enables one (at least in principle) to generate any system configuration (including fluids) from nothing by determining the atoms' positions step-by-step with the help of TPs; the product of these TPs leads to the value of P_i^{B}, and hence to S and F. Practically, a loop/water/template configuration would be generated by initially building a loop structure (in the presence of moving water) followed by the construction of a configuration of the surrounding water molecules (in the presence of a *fixed* loop conformation). In this way, a sample of statistically independent system configurations can be obtained.[47]

For simplicity, this construction is described for a loop without side chains, consisting of M Gly residues, i.e., ordered along the chain are the $3M$ heavy atoms denoted $k' = 1, \ldots, 3M$ and the $6M$ dihedral and bond angles denoted α_k, $1 \le \alpha_k \le 6M = K$, with values within microstate m; the loop is surrounded by N_{water} water molecules moving within the volume defined by a sphere of radius, R_{water}, the template, and the loop. We seek to generate a configuration of the entire system by first generating a loop conformation and then a configuration of the water molecules.

With the scanning procedure, the position of the heavy atoms ($\mathbf{x}_{\text{loop}}$) is determined step by step. Thus the position of the first atom $k' = 1$ is defined by the simultaneous determination of the first pair of dihedral and bond angles α_1,

and α_2. The maximum range $\Delta\alpha_1\Delta\alpha_2$ which will keep the loop within m is defined, and each of $\Delta\alpha_1$ and $\Delta\alpha_2$ is divided into n_b small bins (of sizes $\Delta\alpha_1/n_b$ and $\Delta\alpha_2/n_b$) denoted j_1 and j_2, $j_1 = 1, \ldots, n_b$, $j_2 = 1, \ldots, n_b$, respectively. A *long* MD simulation of the whole system (loop + water) is carried out within microstate m, where a conformation is retained every l fs leading to a huge sample of size n; then, the number of conformations $n_{j_1 j_2}$ that visit simultaneously the (double) bin $j_1 j_2$ is calculated from which the corresponding TP is obtained, $p_{j_1 j_2} = n_{j_1 j_2}/n$ (or $\rho_{j1j2} = [n_{j_1 j_2}/n]/[\Delta\alpha_1\Delta\alpha_2/n_b^2]$). A double bin is then selected by a random number according to the $p_{j_1 j_2}$ which defines the position of atom $k' = 1$ (and its hydrogen or oxygen). The position of this atom is not changed in the next steps of the build-up process, i.e., it becomes part of the "past". The position of the second atom ($k' = 2$) is determined in the same manner from a long MD simulation of the future part of the system (i.e., atoms $k' = 2, \ldots, 3M$ and water) where α_3 and α_4 are considered, bins $\Delta\alpha_3/n_b$ and $\Delta\alpha_4/n_b$ are defined, probabilities are calculated, and a "lottery" (like above) determines the values of α_3 and α_4 which define the position of atom $k' = 2$; the process continues until the positions of all the loop's atoms (and their hydrogens or oxygens) have been determined. A configuration of the N_{water} molecules is then determined in a similar way step by step in the presence of the fixed loop structure previously constructed (for details, see Ref. 20). Obviously, the smaller the bins are, the higher is the accuracy of the construction process, provided that the statistics is adequate, i.e., that the (future) MD simulations are long enough; this *stochastic* scanning method becomes exact as the bin size $\to 0$ ($n_b \to \infty$) and $n \to \infty$. Notice that in applications of the (deterministic) scanning method to *lattice* models, only part of the future has been considered (i.e., only f steps ahead), where this part has been scanned completely; therefore, the corresponding TPs are approximate but deterministic (rather than stochastic), and accurate results were obtained by using an additional importance sampling procedure.[47]

This procedure can be described more formally as follows: at step k' ($k = 2k'$), the positions of $k' - 1$ atoms have already been determined (from the values of the corresponding $k - 2$ angles $\alpha_1, \ldots, \alpha_{k-2}$) and they are kept fixed (defining the "past"); α_{k-1} and α_k (which will determine the position of atom k') are defined with the *exact* TP density $\rho(\alpha_{k-1}\alpha_k|\alpha_{k-2}, \ldots, \alpha_1)$

$$\rho(\alpha_{k-1}\alpha_k|\alpha_{k-2}, \ldots, \alpha_1) = Z_{\text{future}}(\alpha_k\alpha_{k-1}, \ldots, \alpha_1)/[Z_{\text{future}}(\alpha_{k-2}, \ldots, \alpha_1)], \quad (7)$$

where $Z_{\text{future}}(\alpha_k, \ldots, \alpha_1)$ is a future partition function. The term "future" indicates that the integration defining Z_{future} is carried out over the positions of atoms $k' = k/2 + 1, \ldots, K/2$ (which affect angles, $\alpha_{k+1}, \ldots, \alpha_K$) and the $9N$ coordinates $\mathbf{x}^N$ of the water molecules (which will be determined in future steps of the build-up process). Notice that this integration is carried out in a restrictive way where the corresponding conformations (of the loop) remain within microstate m. Also, in this integration the atoms treated in the past $(1, \ldots, k' - 1)$ (which were determined by $\alpha_1 \cdots \alpha_{k-2}$) are held fixed in their coordinates. For simplicity, the integrations below are written over the angles rather than the Cartesian coordinates ($\mathbf{x}_{\text{loop}}$) of the loop atoms, $k' = k/2 + 1, \ldots, K/2$. Thus

$$Z_{\text{future}}(\alpha_k, \ldots, \alpha_1) = \int_m \exp[-(E(\alpha_K, \ldots, \alpha_1, \mathbf{x}^N)/k_B T]d\alpha_{k+1} \cdots d\alpha_K d\mathbf{x}^N, \quad (8)$$

where E [Eq. (1)] is the total potential energy of the loop/template/water system, which also imposes the loop closure condition. The product of the TPs [Eq. (7)] leads to the (Boltzmann) probability density of the entire loop conformation, $\rho^{\text{B}}_{\text{loop}}(\alpha_K, \ldots, \alpha_1)$. After the loop structure has been constructed, a configuration of water molecules is generated step by step [in the presence of the constant loop structure where the product of the corresponding TPs leads to the probability density of the water configuration, $\rho^{\text{B}}_{\text{water}}(\alpha_K, \ldots, \alpha_1, \mathbf{x}^N)$. The probability density $\rho^{\text{B}}([\alpha_k], \mathbf{x}^N)$ of the loop/water/template configuration is the product of $\rho^{\text{B}}_{\text{loop}}([\alpha_k])$ and $\rho^{\text{B}}_{\text{water}}([\alpha_k], \mathbf{x}^N)$]. One can define for m "the loop entropy of mean force," S_{loop},

$$S_{\text{loop}} = -k_B \int_m \rho^{\text{B}}_{\text{loop}}([\alpha_k]) \ln \rho^{\text{B}}_{\text{loop}}([\alpha_k]) d[\alpha_k], \quad (9)$$

where S_{loop} is defined up to an additive constant. Extending the exact scanning procedure to side chains is straightforward, where again the position of a side-chain atom is defined by two angles as described previously.

However, implementation of the exact scanning procedure [as described prior to Eq. (7)] for generating Boltzmann-weighted configurations of a large loop/water/template system would be inefficient, due to the need to calculate (at each step) a large set of accurate TPs (i.e., for an extremely large number of small bins); this would require extremely long (future) MD simulations. Also, with long simulations it is difficult to guarantee that the loop will remain in microstate m (see the discussion in Ref. 13). However, the exact scanning procedure provides the theoretical basis for HSMC(D). Thus, the exact scanning method is equivalent to any other exact simulation technique (in particular, to MC or MD) in the sense that large samples generated by such methods lead to the same averages and fluctuations (the sample does not carry a memory of the simulation method with which it has been generated). 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), the efficiency problems discussed earlier for the *exact* scanning procedure are alleviated to a large extent since only a *single* TP is calculated at each step [rather than many TPs (p_j) required with the scanning method], as described below; also, because we are mainly interested in entropy differences, approximations (e.g., ignoring the Jacobians and bond stretching) can be applied without compromising the accuracy of the results.

E. The HSMC(D) method

The theory of HSMC(D) is described again for a loop consisting of M Gly residues. Notice that while HSMD and

the exact scanning method (described in the previous section) have some resemblance, they are different. With the scanning method a loop conformation is *generated* with P_i^{B} whereas with HSMD an already generated structure (by MD) is reconstructed in order to obtain its probability. Thus, one starts by generating an MD sample of the loop in water in microstate m; the configurations of this sample will be reconstructed by HSMD (application of HSMC is similar). First, the conformations are represented in terms of the dihedral and bond angles α_k,$1 \le \alpha_k \le 6M = K$, and the variability range $\Delta\alpha_k$ is calculated,

$$\Delta\alpha_k = \alpha_k(\max) - \alpha_k(\min), \tag{10}$$

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 has not "escaped" from microstate m. Notice that in the discussion below, we define the loop conformation by the set of angles $[\alpha_k]$ rather than by the positions of the loop atoms, $\mathbf{x}_{\mathrm{loop}}$, which are determined by $[\alpha_k]$.

Each of the configurations (frames) $([\alpha_k], \mathbf{x}^N)$ (denoted i for brevity) of the sample is reconstructed in two stages, where the biotin structure is reconstructed first followed by the reconstruction of the water configuration $\mathbf{x}^N$. Because the position of atom k' is defined by a dihedral and a bond angle, one has to calculate their TP simultaneously. Thus, at step k' ($k = 2k'$) of stage 1, the $k-2$ angles $\alpha_{k-2} \cdots \alpha_1$ have already been reconstructed and the TP density of $\alpha_{k-1}\alpha_k$, $\rho(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1)$, is calculated from an MD run, where the *entire* future of the loop and water is moved (i.e., the loop's atoms k', $k'+1, \ldots, K/2$ and their connected hydrogens, and the water coordinates $\mathbf{x}^N$) while the past (loop's atoms $1, 2, \ldots, k'-1$ and their connected hydrogens) are held fixed at their values in conformation i. By considering a future conformation every 6 fs, a sample of size n_f is generated. Two small segments (bins) $\delta\alpha_{k-1}$ and $\delta\alpha_k$ are centered at $\alpha_{k-1}(i)$ and $\alpha_k(i)$, respectively, and the number of *simultaneous* visits, n_{visit}, of the future chain to these two bins during the simulation is calculated; one obtains (see Fig. 2)

$$\begin{aligned}\rho_{\mathrm{loop}}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1) &\approx \rho^{\mathrm{HS}}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1)\\ &= n_{\mathrm{visit}}/[n_f\,\delta\alpha_{k-1}\delta\alpha_k]\end{aligned} \tag{11}$$

where $\rho^{\mathrm{HS}}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1)$ becomes exact for very large n_f ($n_f \to \infty$) and very small bins ($\delta\alpha_{k-1}$, $\delta\alpha_k \to 0$). This means that in practice $\rho^{\mathrm{HS}}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1)$ will be somewhat approximate due to insufficient future sampling (finite n_f) and relatively large bin sizes. In Ref. 13, we have shown that $\delta\alpha_k$ and $\delta\alpha_{k+1}$ can be optimized. Notice that with HSMD the future loop 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 related to the loop is

$$\begin{aligned}\rho^{\mathrm{HS}}(\alpha_K, \ldots, \alpha_1) &= \rho^{\mathrm{HS}}([\alpha_k])\\ &= \prod_{k=2,2}^{K} \rho^{\mathrm{HS}}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2}, \ldots, \alpha_1).\end{aligned} \tag{12}$$

$\rho^{\mathrm{HS}}([\alpha_k])$ defines an approximate entropy functional, denoted $S^{\mathrm{A}}_{\mathrm{loop}}$, which can be shown (using Jensen's inequality) to constitute a *rigorous* upper bound for S_{loop} [Eq. (9)],[20,48]

$$S^{\mathrm{A}}_{\mathrm{loop}} = -k_{\mathrm{B}} \int_m \rho^{\mathrm{B}}_{\mathrm{loop}}([\alpha_k]) \ln \rho^{\mathrm{HS}}([\alpha_k]) d[\alpha_K]. \tag{13}$$

$\rho^{\mathrm{B}}_{\mathrm{loop}}$ [Eq. (9)] is the Boltzmann probability density of $[\alpha_K]$ in m. ($S^{\mathrm{A}}_{\mathrm{loop}} \ge S_{\mathrm{loop}}$ is also known as Gibbs' inequality.) Being an upper bound suggests that $S^{\mathrm{A}}_{\mathrm{loop}}$ will decrease as the approximation improves. In practice, however, this inequality is satisfied only if the probabilities are well defined (e.g., they are correctly normalized). This is pointed out because $\rho^{\mathrm{HS}}(\alpha_K, \ldots, \alpha_1)$ is calculated stochastically and its error [and the error in $\bar{S}^{\mathrm{A}}_{\mathrm{loop}}$, the estimation of $S^{\mathrm{A}}_{\mathrm{loop}}$, see Eq. (14)] increases with increasing system size (i.e., increasing the number of TPs; see later discussions in Secs. II F and III C).

$S^{\mathrm{A}}_{\mathrm{loop}}$ can be estimated (for microstate m) from a Boltzmann sample (of size n_s) generated by MD using the arithmetic average,

$$\bar{S}^{\mathrm{A}}_{\mathrm{loop}}(m) = -\frac{k_{\mathrm{B}}}{n_{\mathrm{s}}} \sum_{t=1}^{n_{\mathrm{s}}} \ln \rho^{\mathrm{HS}}(t, m), \tag{14}$$

where $\rho^{\mathrm{HS}}(t, m)$ is the value of $\rho^{\mathrm{HS}}([\alpha_k])$ obtained for configuration t of the sample of m. $\bar{S}^{\mathrm{A}}_{\mathrm{loop}}$ (with the bar) is an estimation of the ensemble average $S^{\mathrm{A}}_{\mathrm{loop}}$ [Eq. (13)]; correspondingly, the ensemble averages of the energy are estimated from

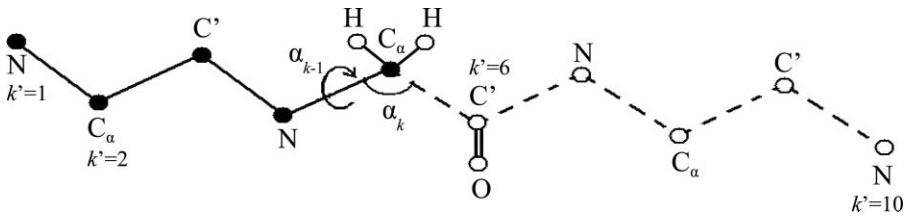

FIG. 2. An illustration of the HSMD reconstruction process of conformation i of a peptide consisting of three glycine residues, where for simplicity the oxygens and most of the hydrogens are discarded. At step $k' = 6$, the TPs related to the "past" atoms $k' = 1, \ldots, 5$ (depicted by full spheres) have already been determined and these atoms are kept fixed in their positions at conformation i. At this step (6) one calculates the TPs of bond angle α_k (defined by C′-C_α-N) and dihedral angle α_{k-1} (defined by C′-C_α-N-C′) which are related to C′, where $k = 2k'$ and thus $k = 12$. The TPs are obtained from an MD simulation where the as yet unreconstructed atoms $k' = 6, \ldots, 10$ (the "future" atoms) are moved (depicted by empty spheres connected by dashed lines) while the past atoms $k' = 1, \ldots, 5$ are kept fixed; note that the future part should remain within the limits of the microstate and future-past interactions are taken into account. Small bins $\delta\alpha_{k-1}$ and $\delta\alpha_k$ are centered at the values α_{k-1} and α_k in i. The TP is calculated from the number of visits [n_{visit}, Eq. (11)] of the future part to $\delta\alpha_{k-1}$ and $\delta\alpha_k$ *simultaneously* during the simulation. After $\rho^{\mathrm{HS}}(\alpha_{11}\alpha_{12}\,|\alpha_{10}, \ldots, \alpha_1)$ [Eq. (12)] has been determined, the coordinates of C′ (and O) are fixed at their positions in i, i.e., they become "past" atoms and the process continues.

a sample of size n_s and should appear with a bar as well. However, from now on only estimations will be considered, and for simplicity, all of them will appear without the bar, like the energies defined in Eq. (1). $S^{\mathrm{A}}_{\mathrm{loop}}$ [Eqs. (13) and (14)] constitutes a measure of a pure geometrical character for the loop flexibility, i.e., with no *direct* dependence on the interaction energy. When the *converged* or the *best* value of $S^{\mathrm{A}}_{\mathrm{loop}}$ is considered, it will be denoted by S_{loop}; thus, $F_{\mathrm{loop}} = E_{\mathrm{loop}} - TS_{\mathrm{loop}}$ is defined as the loop's contribution to the total free energy, where E_{loop} is defined in Eq. (1). In the same way, the difference in the loop entropies between the open (o) and closed (c) microstates obtained for a specific set of parameters is denoted by $\Delta S^{\mathrm{A}}_{\mathrm{loop}}$ while the converged values are denoted without "A" (i.e., $\bar{S}_{\mathrm{loop}}$) and their difference is denoted by ΔS_{loop},

$$\Delta S^{\mathrm{A}}_{\mathrm{loop}} = \bar{S}^{\mathrm{A}}_{\mathrm{loop}}(\mathrm{o}) - \bar{S}^{\mathrm{A}}_{\mathrm{loop}}(\mathrm{c}) \tag{15a}$$

where

$$\Delta S_{\mathrm{loop}} = \bar{S}_{\mathrm{loop}}(\mathrm{o}) - \bar{S}_{\mathrm{loop}}(\mathrm{c}). \tag{15b}$$

One can define a free-energy difference for the loop, ΔF_{loop},

$$\Delta F_{\mathrm{loop}} = \Delta E_{\mathrm{loop}} - T\Delta S_{\mathrm{loop}}, \tag{16}$$

where ΔE_{loop} is obtained from Eq. (1).

To reconstruct the water configuration, one can use in principle the HSMC(D) procedure for fluids mentioned previously,[20] which would lead to $\rho^{\mathrm{HS}}_{\mathrm{water}}([\alpha_k], \mathbf{x}^N)$ and then to the contribution of the water configuration to the free energy, $F_{\mathrm{water}}([\alpha_k], \mathbf{x}^N) = E_{\mathrm{water}}([\alpha_k], \mathbf{x}^N) + k_{\mathrm{B}}T \ln \rho^{\mathrm{HS}}_{\mathrm{water}}([\alpha_k], \mathbf{x}^N)$. However, this procedure for fluids has not been optimized yet and it is relatively time consuming.

Alternatively, as in Refs. 14 and 26, one can obtain $F_{\mathrm{water}}([\alpha_k], \mathbf{x}^N)$ by a TI procedure based on the same reference state for the open and closed structures. In this state, the water-water and water-template interactions are preserved but the (fixed) loop structure $[\alpha_k]$ does not "see" the surrounding waters, i.e., the loop-water interactions [electrostatic and Lennard-Jones (LJ)] are switched off. These interactions are gradually increased (from zero) during an MD simulation of water (while the loop structure remains fixed at $[\alpha_k]$). For $[\alpha_k]$ of microstate m the integration leads to $F^{\mathrm{TI}}_{\mathrm{water}}([\alpha_k], m)$, which is then averaged over the n_s sample configurations [as in Eq. (14)]. The integration is performed in two stages but in an *opposite* direction to that described above, i.e., first the charges are gradually *decreased* to zero (by decreasing the parameter λ from 1 to zero; see Sec. III D), followed by a similar decrease in the LJ potential, leading to $F^{\mathrm{TI}}_{\mathrm{water}}([\alpha_k], m, \mathrm{ch})$ and $F^{\mathrm{TI}}_{\mathrm{water}}([\alpha_k], m, \mathrm{LJ})$, respectively. Denoting the set of $[\alpha_k]$ in the sample by t and omitting m, one obtains

$$\begin{aligned} F^{\mathrm{TI}}_{\mathrm{water}}(m) &= F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{ch}) + F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{LJ}) \\ &= \frac{1}{n_{\mathrm{s}}} \sum_{t=1}^{n_{\mathrm{s}}} \left[F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{ch}, t) + F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{LJ}, t) \right]. \end{aligned} \tag{17}$$

The difference in the free energy of water between the open and closed microstates is

$$\Delta F_{\mathrm{water}} = F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{o}) - F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{c}) \tag{18}$$

and the difference in the total free energy between the open and closed microstates is

$$\Delta F_{\mathrm{total}} = \Delta E_{\mathrm{loop}} - T\Delta S_{\mathrm{loop}} + \Delta F_{\mathrm{water}}. \tag{19}$$

The corresponding thermodynamic cycle is described in Fig. 3, which shows that the free-energy calculation for both loops starts from the same reference state (0) of a (fixed) template surrounded by water, i.e., the equilibrium state is defined by water-water and water-template interactions. In the first step, n_s loop structures are reconstructed independently in water and the corresponding free energies, $F_{\mathrm{loop}}(\mathrm{o})$ and $F_{\mathrm{loop}}(\mathrm{c})$, are obtained for the open and closed loops. The hatched background in the reference state and the first stage demonstrates that the free energy of water is not considered. In the second step, the loop-water interaction is gradually decreased to zero by the TI procedure described above for the n_s configurations, where the (negative) average results are $F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{o})$ and $F^{\mathrm{TI}}_{\mathrm{water}}(\mathrm{c})$; the crossed-hatched background here demonstrates that the free energy of water is considered. The sums of the loop and water contributions are $F_{\mathrm{total}}(\mathrm{o})$ and $F_{\mathrm{total}}(\mathrm{c})$, and their difference is $\Delta F_{\mathrm{total}} = F_{\mathrm{total}}(\mathrm{o}) - F_{\mathrm{total}}(\mathrm{c})$ [Eq. (19)].

F. Analysis of the reconstruction results

First, notice that the heavy atoms are ordered along the loop and denoted $k' = 1, \ldots, K/2$. Because in the reconstruction of step k' the whole future is simulated (i.e., atoms k', $k'+1, \ldots, K/2$), one can order the atoms in different ways, e.g., residue by residue, or the entire backbone first and the side-chain atoms later. In the present work, the atoms are considered residue by residue where for each residue the backbone is treated before the side chain. Also notice that the position of atom k' is defined by the bond and dihedral angle, α_{k-1} and α_k, where $k = 2k'$.

The MD simulation (in water) of the future chain at step k' starts from conformation i, which we want to reconstruct, and every $g = 6$ fs the current conformation is retained; the n_{init} initial retained conformations are discarded for equilibration. For each of the next n_f (retained) future conformations (i.e., based on the positions of the atoms $k', k'+1, \ldots, K/2$) the dihedral and bond angles, α_{k-1} and α_k, are calculated and if both are found to be within their corresponding bins, n_{visit} [Eq. (11)] is increased by 1, which leads to the TP, $\rho_{\mathrm{loop}}(\alpha_{k-1}\alpha_k | \alpha_{k-2}, \ldots, \alpha_1)$ [Eq. (11)]; when the entire loop structure has been reconstructed (i.e., the TPs have been calculated for the $K/2$ pairs of angles), one obtains an estimation for the probability density of this structure, $\rho^{\mathrm{HS}}(\alpha_K, \ldots, \alpha_1)$ [Eq. (12)]. Notice that in practice, one might encounter two opposite scenarios of over- or undercoverage of the future part of m.

Thus, if n_f is too large, the loop might "overflow" to a neighbor microstate, leading to a number of counts (n_{visit}) which is too small, hence to too small TPs and probabilities, $\rho^{\mathrm{HS}}(\alpha_K, \ldots, \alpha_1)$; the larger n_f is, the smaller are these probabilities. Therefore, the corresponding values of $\bar{S}^{\mathrm{A}}_{\mathrm{loop}}(m)$ [Eq. (14)] will increase with increasing n_f rather than decrease as expected by theory for improving approximations

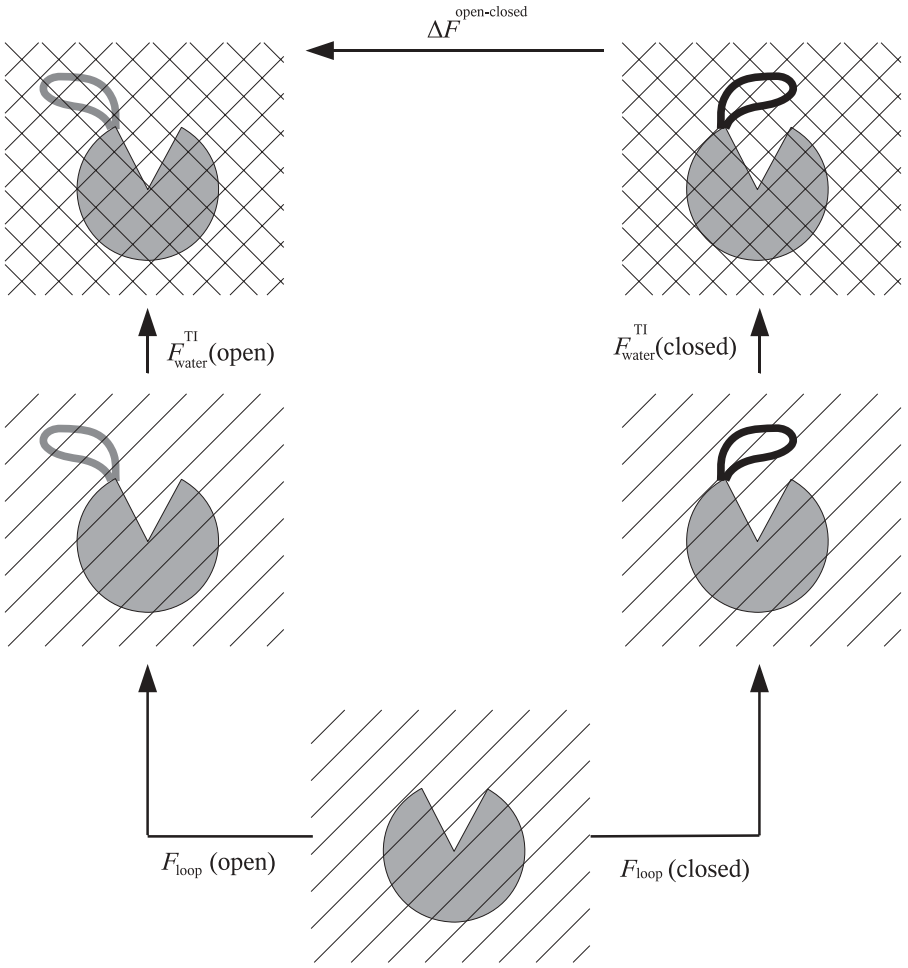

FIG. 3. A schematic illustration of the thermodynamic cycle used to calculate the difference in free energy between the open and closed microstates. For both microstates, the calculations start from the same reference state: a fixed template (gray partial circle) soaked in water (parallel lines) which appears in the bottom. In step 1, a set of 40 loop conformations selected (for each microstate) from a 2 ns MD trajectory are reconstructed in water by HSMD leading to the loop contribution to the free energy, $F_{\text{loop}} = E_{\text{loop}} - TS_{\text{loop}}$, where E_{loop} [Eq. (1)] consists of the loop-loop and loop-template energy and S_{loop} is the converged results of the entropy defined in Eqs. (13) and (14); the parallel lines indicate that the loop is soaked in water, which affects its behavior, while the free energy of water is not considered directly. In step 2, the contribution of water to the free energy, $F^{\text{TI}}_{\text{water}}$ [Eq. (17)], is calculated for each of the 40 configurations by a thermodynamic integration procedure where the loop-water interactions are increased gradually from zero (the reference state) to their full value (in practice these interactions were decreased to zero). The squared background means that the free energy of water is calculated. The total free-energy difference is $\Delta F_{\text{total}} = \Delta E_{\text{loop}} - T\Delta S_{\text{loop}} + \Delta F_{\text{water}}$ [Eq. (19)].

[see the discussion following Eq. (13)]. [Note that even at step k', where the "past" of the loop (atoms $1, \ldots, k'-1$) is kept fixed, the (future) unfixed part ($k', \ldots, K/2$) can leave the microstate during long MD simulations; such an "overflow" is more likely to happen for small residues such as Gly and for small k.] To control an overflow, we have suggested carrying out the reconstruction in j shorter repetitive "units,"[13,14,25] each based on $n'_f < n_f$ conformations where $n_f = jn'_f$. Using units of increasing length (n'_f) and larger values of n_f (i.e., larger j) enables one to gain control on the extent of coverage of a microstate by the future simulations (again, very small n'_f values will lead to undercoverage of the microstate while large n'_f might lead to an overflow). Obviously, each unit should start from the reconstructed structure i with a different set of velocities followed by an adequate equilibration of size n_{init}; an important test for an adequate coverage is verifying that $\bar{S}^{\text{A}}_{\text{loop}}(m)$ decreases with increasing j [i.e., improving the estimation of $\rho^{\text{HS}}(\alpha_K, \ldots, \alpha_1)$] which indeed has been found in several previous studies.[13,14,23,25]

In the second scenario, the future conformation is located within m but n_f is too small for adequately calculating a TP [Eq. (11)], i.e., the ratio n_{visit}/n_f as yet has not been stabilized. As an example, consider a χ angle which visits (in m) more than one rotamer or all of them (i.e., $\Delta\alpha_k = 360^\circ$); the corresponding ratio n_{visit}/n_f will decrease systematically as n_f is increased, since most of the sampled angles will fall out of the considered bin, meaning that stabilization will occur only for very large n_f. Thus, for practical n_f values, TP(n_f) will also decrease with increasing n_f and if the number of such angles is significant they might lead to a systematic increase of $\bar{S}^{\text{A}}_{\text{loop}}(m)$.

In a normal situation, however, increasing n_f will lead to an improved estimation of the TPs, and hence of ρ^{HS}, and the corresponding results for $\bar{S}^{\text{A}}_{\text{loop}}(n_f)$ are expected to decrease as required by theory, provided that the probabilities are well defined [see the discussion following Eq. (13)]. Indeed, in Sec. III C we discuss a case where insufficient equilibration (i.e., too small n_{init}) leads to unnormalized ρ^{HS}

which is followed by a systematic increase in $\bar{S}^{A}_{loop}(n_f)$ as n_f is increased.

Therefore, the analysis of the results for $\bar{S}^{A}_{loop}(m)$ requires caution. It should first be noted that our main interest is in the differences ΔS^{A}_{loop} between microstates m and n rather than in the absolute values themselves. For any practical set of n_f and bin sizes, $\delta\alpha_k$, $\bar{S}^{A}_{loop}(m)$ and $\bar{S}^{A}_{loop}(n)$ will be approximate and thus their difference, ΔS^{A}_{loop}, might be approximate as well. However, if ΔS^{A}_{loop} is found to be stable for significantly improving sets of parameters, the stable value can be considered as the correct difference (within the statistical errors). Indeed, in applications of HSMD to peptides[25] and loops,[13,14,26] relatively small values of n_f have already led to stable differences, meaning that the *systematic* errors in both $\bar{S}^{A}_{loop}(m)$ and $\bar{S}^{A}_{loop}(n)$ are comparable and thus are canceled in ΔS^{A}_{loop} [we define the deviation, $\bar{S}^{A}_{loop}(m) - S_{loop}$ as the systematic error]. In Ref. 13, we have provided theoretical arguments supporting this error cancellation, which, however, should be verified for each system studied.

Thus, using HSMC(D)-TI, the objective is not to obtain the most accurate free energies, F_m and F_n, but to minimize computer time by finding the *worst* HSMC(D)-TI approximations for F_m and F_n for which their difference is still correct within a required statistical error. It should be pointed out that the difficulties involved in the definition of a microstate discussed above are not characteristic only for HSMD-TI but are common to all methods for calculating entropy.

G. Various implementations of the reconstruction procedure

A somewhat technical issue is how to program the reconstruction procedure in the most general way that will be applicable to loops of different sequence and length. We describe two such procedures, where the reconstruction is performed atom by atom, and hence an order of the atoms along the chain should initially be determined, as discussed earlier.

Thus, at step k' atom k' is treated and the related dihedral and bond angles, α_{k-1}, α_k ($k = 2k'$), are obtained in the usual way from the positions of atoms k', $k'-1$, $k'-2$, and $k'-3$, and k', $k'-1$, and $k'-2$, respectively (see Fig. 2). In general, all heavy atoms and polar hydrogens are taken into account, but one can consider approximations based on a smaller number of atoms or include methyl hydrogens if, for example, the rotational entropy of the methyl groups of valine is of interest. In this way, if C′ ($k' = 4$) is treated and the previous atoms are C^{α} ($k' = 3$), N ($k' = 2$), and H(N) ($k' = 1$), these last three atoms are held fixed, while C′ and the atoms numbered $k' > 4$ move in the MD reconstruction simulation; n_{visit} is calculated for the dihedral ψ and the bond angle defined by the positions of C′, C^{α}, and N. After the reconstruction simulation for $k' = 4$ has been completed, C′ becomes fixed in its position at conformation i and O(C′) ($k' = 5$) should be treated next (where C′, C^{α}, and N are its previous atoms to be considered); thus, the dihedral angle is ψ and the bond angle is defined by the positions of O, C′ and C^{α}. Notice, however, that after O has been fixed, the position of the next N ($k' = 6$) is determined to a large extent as well. Here we define two procedures, the "conventional" procedure I and procedure II.

With procedure I, the atom O(C′) is not considered (i.e., O is not numbered) and ψ is defined by N($k' = 5$), C′ (4), C^{α}(3), and N(2) and the corresponding bond angle by N(5)-C′ (4)-C^{α}(3), where O [like N(5)] is allowed to move in the simulation because these angles also define the position of O to a large extent. With procedure II, on the other hand, O(C′) is considered as described above followed by the treatment of N ($k' = 6$) for which the pair of angles are defined now by the positions of N ($k' = 6$), O ($k' = 5$), C′ ($k' = 4$), and C^{α} ($k' = 3$); however, these angles are not the conventional ones since the dihedral angle is defined around C′O and the bond angle N-O-C′ is based on the nonstandard N-O bond.

While procedure II is somewhat more accurate than procedure I, its implementation is more time consuming (more atoms and angles) and the contribution of the additional reconstructed angles to the entropy is expected to be comparable in different microstates. Therefore, entropy *differences* (our main interest) obtained with procedure I are not expected (in general) to change significantly by procedure II.

III. RESULTS AND DISCUSSION

A. Simulation of the entire tetramer

As discussed in the Introduction, an interesting question is whether the conformational transition of the loop upon binding demonstrates an induced fit, i.e., whether it moves due to interactions induced by the ligand, or alternatively the open loop interconverts among different microstates, one of which is selected upon binding (selected fit). In the x-ray structures of Freitag *et al.*[30] and others, the loop in the bound protein is always in the closed conformation with well-resolved coordinates. Without biotin, the loop in most cases is partially disordered and exhibiting elevated B-factors (>40 Å^2); however, in structures I and II of Freitag *et al.* where biotin is not bound, the loop in subunit 1 is closed suggesting that a selected fit is a possibility. To check the extent of flexibility of the loop in the closed and open microstates, we carried our two MD simulations as described below.

The first simulation is based on the x-ray structure of tetramer IV (1swd) of Freitag *et al.*, where the disordered loop conformations of subunits 2 and 3 were replaced by the complete open-loop conformation of subunit 2 in crystal structure III (1swc). Thus, the "retouched" 1swd structure consists of two open loops (subunits 2 and 3) and two closed loops (subunits 1 and 4) which cover the corresponding bound biotins. This tetramer of 6820 atoms (including hydrogens, not counting biotins) was solvated with a 10 Å buffer of TIP3P water (13688 water molecules) and 6 Na+ ions to neutralize the total charge. The system was treated with periodic boundary conditions, where the particle mesh Ewald (PME) summation method was used for electrostatics; the real-space terms were evaluated with a cutoff distance of 15 Å and the simulations were carried out in the *NVT* ensemble.

The system was minimized without restraints for 15000 steps of steepest descent and 15000 steps of conjugate gradient, followed by a 1 ns MD run where the system was

heated gradually to 300 K. In the subsequent 2.5 ns production run, no significant structural changes were observed in the protein; in particular, the conformations of the two closed loops (with biotins) and the two open ones only exhibited localized fluctuations.

Next, our tetramer was changed further by moving the biotin of subunit 1 (below the closed loop) to the empty pocket of subunit 3 where the loop has an open conformation. Our objective has been to check whether the open loop of subunit 3 would move to a closed conformation due to its interaction with biotin, and whether the empty pocket in subunit 1 would cause the closed loop there to open.

The system was first energy minimized for 10^4 steps of steepest descent and another 10^4 steps of conjugate gradient, then was heated to 300 K during a 50 ps MD run, followed by additional 50 ps run at pressure of 1 atm and 300 K; however, the production simulation was carried out at fixed temperature and volume (*NVT* ensemble). During the first 2 ns, the loops underwent only small local conformational fluctuations. To check the stability further, the temperature was increased gradually to 500 K followed by a 5 ns MD run. The local fluctuations of the closed loop without biotin (subunit 1) increased significantly and the loop moved to a different microstate, but no transition from the closed to the open conformation has been observed. The fluctuations of the open loop in subunit 3 were also increased significantly, but in this case no transition to a different microstate was observed.

While these simulations are relatively short, they suggest that both the open and closed conformations (even without biotin) have considerable local stability; furthermore, according to our later results the free energy of the open loop is lower by ∼27 kcal/mol than that of closed one. Therefore, it is plausible to assume that the conformations of the open loop (in apo streptavidin) do not cover the closed microstate, and the conformational response of the loop to ligand binding is probably not of a selected fit type (i.e., it is an induced fit). The fact that the open-loop structures in the presence of biotin did not change to the closed microstate and the closed-loop structures without biotin did not move to an open microstate is probably due to our relatively short simulation. This interpretation is in accord with a Gaussian network model (GNM) analysis of 1swd, where the three lowest-frequency modes have only a minor effect on the open loop, which undergoes relatively large fluctuations only in the fourth mode. The fluctuations of the closed-loop structures are not significant in any of the (20) lowest modes. In particular, no open-closed transition is observed.[49]

B. Simulations of the loop/water/template systems

In Sec. II B we defined loop/water/template systems (for calculating the loop free energy) and described their optimization by a process that ended in 500 ps MD simulations applied to the open and closed loop. The last loop/water/template structures obtained in these optimization runs became the starting structures for (two) 2 ns MD production runs from which loop/water/template conformations were retained every 2 ps (i.e., 1000 configurations) for a future free-energy analysis. It should be pointed out that determining the exact limits of a microstate in conformational space is practically impossible and therefore it is commonly defined by the underlying MC or MD trajectories initiated from a microstate's structure. Thus, the microstate's size typically increases with simulation time, *t,* and for large enough *t* the loop might move to a significantly different region in conformational space; obviously, *E, S*, and *F* depend on *t* as well. In previous publications, we have developed procedures for checking that the system remains in its microstate during a simulation (see Ref. 13 and references cited therein). Thus, in the case of two microstates *m* and *n* one should verify that the differences, $\Delta E_{mn}(t)$, $\Delta S_{mn}(t)$, and $\Delta F_{mn}(t)$, are stable during a long enough time, Δt, meaning that the conformational changes in both microstates are small and comparable. Notice that the present trajectories of 2 ns are significantly larger than the ∼0.5 ns trajectories studied in our previous papers.[13,14,23–26,54]

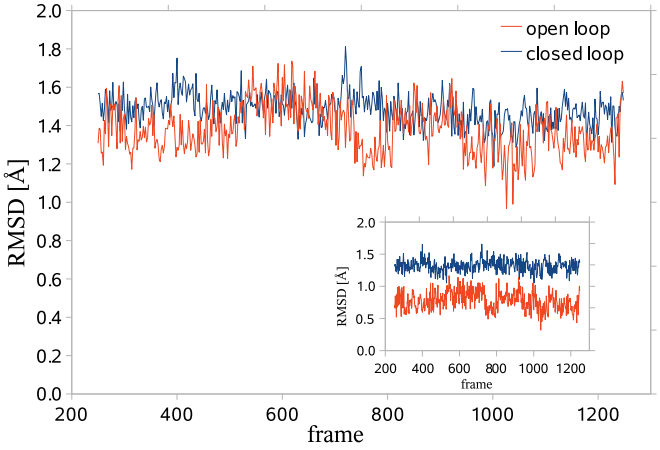

FIG. 4. Root-mean-square deviation (RMSD) of the heavy atoms from the x-ray structure for the open and closed loops obtained from MD trajectories of 2 ns, where a frame is defined every 1 ps. While lower RMSD values are observed for the open loop, their fluctuations are the largest, suggesting that the entropy of the open loop is larger than that of the closed one, $\Delta S_{\text{loop}} > 0$ [Eq. (15b)]. The inset shows the RMSD for backbones.

First, we calculated for each run (trajectory) the RMSD of the loop's conformations from its initial one, and the results are presented in Fig. 4 as a function of simulation time. The figure reveals that the closed loop has moved from its initial conformation slightly more than the open loop, as is also demonstrated by the average RMSD values, 1.49 and 1.38 Å, respectively. This is expected as the closed loop is attached to the template of the open one. On the other hand, larger RMSD oscillations are observed for the open loop than for the closed one, which are expressed by the corresponding standard deviations of the RMSD values, 0.13 and 0.08 Å; this is in accord with the standard deviations observed for the potential energy, 9.8 and 9.5 kcal/mol, respectively (data not shown). These larger fluctuations of the open loop suggest that its entropy is higher than that of the closed loop, which agrees with the experimental observation of elevated B-factors and partial disorder of the open loop in crystal structures of sreptavidin.[30]

The higher flexibility of the open microstate is also demonstrated by results for $\alpha_k(\text{min})$, $\alpha_k(\text{max})$, and $\Delta\alpha_k$

TABLE I. Minimum and maximum values of dihedral angles, α_k(min) and α_k(max), and their differences $\Delta\alpha_k$ (in degrees) for the open and closed samples.[a]

		Open loop			Closed loop		
Residue	Dihedrals	Min	Max	Δ	Min	Max	Δ
SER	φ	–168	–120	48	–159	–110	49
	ψ	–25	45	70	134	176	42
	ω	143	205	62	160	205	45
	χ^1	–178	178	356	–29	–210	181
ALA	φ	25	98	73	–166	–118	48
	ψ	5	68	63	–41	15	56
	ω	144	195	51	147	193	46
VAL	φ	100	214	114	39	96	57
	ψ	–61	82	143	108	176	68
	ω	153	212	59	155	200	45
	χ^1	–103	102	205	–99	–42	57
GLY	φ	–163	–58	105	–175	–92	83
	ψ	–36	87	123	–71	8	79
	ω	151	215	64	154	205	51
ASN	φ	–101	–32	69	–175	–93	82
	ψ	92	152	60	–3	62	65
	ω	158	204	46	170	219	49
	χ^1	145	195	50	–101	–34	67
	χ^2	33	126	93	83	245	162
ALA	φ	–89	–43	46	25	91	66
	ψ	–67	–15	52	124	178	54
	ω	152	195	43	155	202	47
GLU	φ	–137	–87	50	165	236	71
	ψ	–36	10	46	1	77	76
	ω	161	199	38	148	208	60
	χ^1	–78	–28	50	–180	–29	151
	χ^2	55	121	66	137	223	86
	χ^3	–152	164	316	–180	180	360
SER	φ	–137	–87	50	165	235	70
	ψ	–10	29	39	144	190	46
	ω	164	186	22	–175	–150	25
	χ^1	–83	–29	54	172	220	48

[a]α_k(min), α_k(max), and $\Delta\alpha_k$ are defined in Eq. (10); their values were calculated from samples of 1000 loop conformations generated for the open and closed microstates by retaining a conformation every 2 ps from the corresponding 2 ns trajectories.

[Eq. (10)] presented in Table I for the backbone dihedral angles φ, ψ, and ω and the side-chain angles χ. These values, obtained (for the open and closed loops) from the samples of 1000 conformations defined above, show that the $\Delta\alpha_k$ values for the backbone are relatively small (in most cases smaller than 80°), but as expected are sometimes higher for Gly and the χ angles. A detailed comparison of the $\Delta\alpha_k$ values reveals that altogether they are larger for the open loop than for the closed one, which again suggests that $\Delta S_{\rm loop}$ [Eq. (15b)] is positive.

C. Results for the loop entropy

Results for the loop entropy, $S^{\rm A}_{\rm loop}$ [Eq. (14)], appear in Table II for the microstates of the open and closed loops and for their difference $T[S^{\rm A}_{\rm loop}({\rm open}) - S^{\rm A}_{\rm loop}({\rm closed})] = T\Delta S^{\rm A}_{\rm loop}$ [see the discussion preceding Eq. (15a)]. These results were obtained by independently reconstructing $n_s = 40$ loop structures, distributed homogeneously along the entire sample of 1000 system configurations. At each reconstruction step of each configuration, the future part of the loop (and water) was simulated by MD, and a structure was retained every 6 fs for a later analysis. We carried out several rounds of analysis as described below.

As discussed earlier, the reconstruction simulation starts at each step from the structure to be reconstructed and the initial part of the simulation is used for equilibration and is thus discarded. To emphasize the important effect of equilibration, we have carried out two sets of analyses. The first set is based an equilibration of 10 ps ($n_{\rm init} = 1667$) followed by a relatively long MD simulation of 60 ps ($n_f = 10^4$ structures) where the probability $\rho_{\rm loop}(\alpha_{k-1}\alpha_k\,|\alpha_{k-2},\ldots,\alpha_1)$ [Eq. (11)] and $TS^{\rm A}_{\rm loop}$ were calculated for $n_f = 500$, 1000, 2000, 4000, 8000, and 10^4. In the second set, another 1000 loop conformations (6 ps) are ignored, i.e., the equilibration is increased to 16 ps ($n_{\rm init} = 2667$) and results are thus calculated for $n_f = 1000$, 3000, 7000, and 9000; these samples are denoted in the table by (1), (3), (7), and (9), respectively. Results for $TS^{\rm A}_{\rm loop}$appear in Table II for different bin sizes, $\delta = \Delta\alpha_k/l$, $l = 10$, 20, 40, 60, 80, and 100 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, the bin size is increased becoming $\delta_1 = \delta + 0.2\delta$ and if necessary it is increased again to $\delta_2 = \delta_1 + 0.2\delta$, etc. until the number of counts becomes 50; in the case of zero counts, $n_{\rm visit}$ is taken to be 1; however an event of zero counts is very rare.

Relying on the discussion following Eq. (13), one would expect the results for $TS^{\rm A}_{\rm loop}$ to decrease as the approximation improves, e.g., as the bin size is decreased for a given value of n_f. The table always shows this expected behavior for the largest values of n_f, $n_f = 10^4$ 9000, 8000, and 4000, while for the other n_f values the $TS^{\rm A}_{\rm loop}$ results in some cases remain constant (rather than decrease) within the (larger) error bars. Similarly, for a given bin one would expect $TS^{\rm A}_{\rm loop}$ to decrease with increasing n_f, which is always satisfied for the set based on the 16 ps equilibration but is never satisfied for the other set (of 10 ps equilibration), where the results *always* increase rather than decrease; this is an example where insufficient equilibration probably leads to unnormalized probabilities, which cause the unexpected behavior of $TS^{\rm A}_{\rm loop}$ [see the discussion following Eq. (13)]. More specifically, because the reconstruction simulation starts from the bin, the bin will remain overoccupied after a short equilibration, and during the following short production run (small n_f); thus, the number of counts $n_{\rm visit}$ [Eq. (11)] will be too large leading to a too low $TS^{\rm A}_{\rm loop}$; as n_f increases the entropy increases (due to the decrease in $n_{\rm init}$), approaching its correct value (from below) for a large n_f where the effect of the initial conditions is gradually eliminated. Note that the required equilibration time is system-dependent, where the present time (16 ps) is significantly larger than the 2.5 ps used in our previous studies,[13,14,26] probably due to the larger loop studied here and its relatively high stability as discussed in Sec. III A. Also, the reconstruction simulations (~60 ps) are

TABLE II. HSMD results (in kcal/mol) for the loop entropy, TS^{A}_{loop}, and for entropy differences $T\Delta S^{A}_{loop}$ between the open and closed microstates at $T = 300$.[a]

			TS^{A}_{loop}(o)		TS^{A}_{loop}(c)		$T\Delta S^{A}_{loop}$	
Equilbration	bin size, δ	n_f	10 ps	16 ps	10 ps	16 ps	10 ps	16 ps
Procedure I	$\Delta\alpha_k/10$	**10000 (9)**	**150.9**	**157.6**	**147.8**	155.4	3.1	2.2
	$\Delta\alpha_k/20$	**10000 (9)**	**148.7**	**151.8**	**144.8**	148.4	3.9	3.4
	$\Delta\alpha_k/40$	**10000 (9)**	**148.1**	**150.4**	**144.1**	**146.6**	**4.0**	**3.8**
	$\Delta\alpha_k/60$	500	141.9		138.4		3.5	
		1000	143.6		139.3		4.3	
		2000 (1)	145.2	155.4	141.0	150.7	4.2	4.7
		4000 (3)	146.6	151.9	142.4	147.4	4.2	4.5
		8000 (7)	147.7	150.2	143.6	146.2	4.1	4.0
		10000 (9)	**147.9**	**150.1**	**143.8**	**146.3**	**4.1**	**3.8**
	$\Delta\alpha_k/80$	500	141.9		138.4		3.5	
		1000	143.6		139.3		4.3	
		2000 (1)	145.2	155.4	140.8	150.6	4.4	4.8
		4000 (3)	146.5	151.8	142.2	147.3	4.3	4.5
		8000 (7)	147.5	150.1	143.4	145.9	4.1	4.2
		10000 (9)	**147.7**	**149.8**	**143.6**	**146.0**	**4.1**	**3.8**
	$\Delta\alpha_k/100$	500	141.9		138.5		3.4	
		1000	143.6		139.4		4.2	
		2000 (1)	145.1	155.4	140.9	150.8	4.2	4.6
		4000 (3)	146.4	151.8	142.1	147.5	4.3	4.3
		8000 (7)	147.2	150.1	143.2	146.0	4.0	4.1
		10000 (9)	**147.3**	**149.7**	**143.3**	**145.8**	**4.0**	**3.9**
Converged							**4.0 ± 1.0**	**3.9 ± 1.0**
Procedure II	4.0°	3333					**5.3 ± 1.0**	**5.3 ± 1.0**
QH I		8000	198.3	198.3	192.6	192.6		5.7 ± 2.5

[a]Results obtained with procedure I based on two equilibration times of 10 and 16 ps for the loop entropy, S^{A}_{loop} [Eqs. (13) and (14)], and the differences, $T\Delta S^{A}_{loop} = T[S^{A}_{loop}(\text{open}) - S^{A}_{loop}(\text{closed})]$ [Eq. (15a)]; they were obtained by reconstructing 40 loop structures selected homogeneously from larger MD samples (of 1000 water-loop configurations) of the open and closed microstates. The results are calculated as a function of the bin size $\delta = \Delta\alpha_k/l$ [Eq. (10)] and n_f [Eq. (11)], the sample size of the future chains used in the reconstruction process. For the 16 ps results, the n_f values appear in parentheses, e.g., $n_f = 8000$ (7) means that $n_f = 8000$ and 7000 for the 10 and 16 ps results, respectively. S^{A}_{loop} is defined up to an additive constant that is expected to be the same for both microstates. The maximal statistical error of the results for TS^{A}_{loop} is 0.7 kcal/mol. The results for $n_f = 10^4$ are bold-faced. For procedure II, the result is presented only for $T\Delta S^{A}_{loop}$ with a constant bin size of 4°. TS^{QH}_{loop} [Eq. (20)] is the quasiharmonic entropy; these results were obtained from larger samples of 8000 loop conformations (see text).

significantly larger than those used before for loops (e.g., 12.5 ps). Still, the 10 and 16 ps results for the open (closed) microstate are approaching each other from above and below, respectively, where for the largest n_f they differ by ∼2.5 kcal/mol (Extrapolating the results in the table suggests that for $n_f \sim 16\,000$ the entropies obtained by the two sets will become equal within the error bars.) See also Fig. 5.

The results for $T\Delta S^{A}_{loop}(n_f)$ (our main interest) for the 10 ps set converge nicely to ∼4 kcal/mol, where $T\Delta S^{A}_{loop}(n_f = 500)$ is systematically too low (∼3.5 kcal/mol). The convergence of the 16 ps set is less pronounced, probably due to lower statistics, i.e., the removal of the (first) 1000 structures that contribute most significantly to n_{visit}, which leads to a relatively high result, $T\Delta S^{A}_{loop}(1000)$ ∼4.6 kcal/mol; therefore, for the 16 ps set the decrease of the $[T\Delta S^{A}_{loop}(n_f)]$ results to the $n_f = 9000$ value (3.9 kcal/mol for $l = 100$) is stronger than that observed for the 10 ps set. Relying on the $T\Delta S^{A}_{loop}$ results for $n_f = 10^4$ and 9000, we estimate $T\Delta S_{loop} = 3.9 \pm 1.0$ kcal/mol. Notice, however, that estimations for $T\Delta S_{loop}$ exceeding by ∼0.4 the 3.9 kcal/mol value can be obtained from smaller (n_f) samples. We have verified again[13] that the Jacobian has no effect on results for $T\Delta S_{loop}$ within the statistical errors. For comparison, we have also calculated the entropy by the quasiharmonic (QH) approximation using the equation[44]

$$S^{QH}_{loop} = (k_B/2)\{N + \ln[(2\pi)^N \mathrm{Det}(\boldsymbol{\sigma})]\}, \quad (20)$$

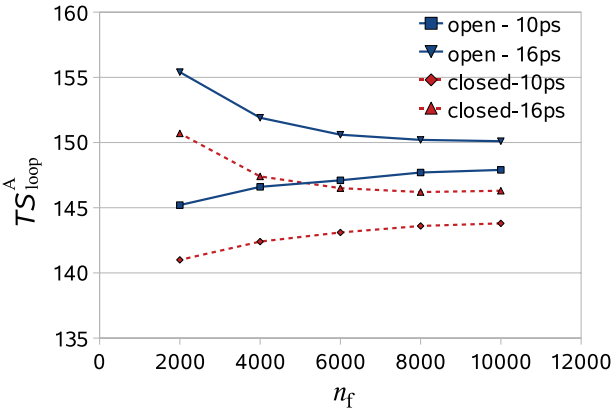

FIG. 5. The approximate loop entropy, $T\bar{S}^{A}_{loop}$ (in kcal/mol + const) [Eq. (14)], is presented as a function of the number of future reconstruction steps, n_f [Eq. (11)] for the open and closed microstates. These results (taken from Table II) are based on a bin size, $\Delta\alpha_k/l$, $l = 60$ [Eq. (10)], and are shown for two equilibrations of 10 and 16 ps. It is demonstrated that for 10 ps, the results for both the open and closed microstates increase while those for 16 ps decrease. Thus, for the open (closed) microstate the 10 and 16 ps results approach each other deviating by ∼2.5 kcal/mol for $n_f = 10^4$ (or 9000).

TABLE III. Free energy of water (in kcal/mol) calculated with thermodynamic integration (TI) for the open and closed loops.[a]

	F^{TI}_{water}(ch)			F^{TI}_{water}(LJ)		F^{TI}_{water}		
Window (ps)	Open	Closed	δ (Å^2)	Open	Closed	Open	Closed	ΔF_{water}
20	–54.4	–117.2	0.5	∞	∞	∞	∞	–
			2.0	45.2	32.6	–9.2	–84.6	75.4
			3.0	41.6	30.1	–12.8	–87.1	74.3
40	**−53.4**	**−116.2**	2.0	**46.5**	**34.9**	**−6.9**	**−81.3**	**74.4**
Errors	±1.0	±0.8		±0.6	±0.4	±1.0	±0.9	±1.3

[a] F^{TI}_{water}(ch) and F^{TI}_{water}(LJ) were obtained by eliminating the loop-water electrostatic (charge) and Lennard Jones (LJ) interactions, respectively; F^{TI}_{water} is their sum [Eq. (17)]. Each integration is based on 22 windows, where the (best) results for windows of 40 ps are bold-faced. δ defines the soft-core LJ potentials [Eq. (21)]. ΔF_{water} [Eq. (18)] is the difference in the water free energy between the open and closed configurations. "Errors" stand for statistical errors [(standard deviation)/$40^{1/2}$].

where $\boldsymbol{\sigma}$ is the covariance matrix and $N = 64$ is the number of internal coordinates. Clearly, S^{QH} (obtained from two samples of 8000 configurations, where a configuration was retained every 1 ps from an 8 ns MD trajectory) constitutes an upper bound for S since correlations higher than quadratic are neglected. Indeed, the QH results are larger (by ∼34%) than the HSMD results (which do take higher-order correlations into account), in accord with our previous studies.[13,14,23,25,26] Still, the QH result $T\Delta S_{loop} = 5.7 \pm 2.5$ kcal/mol is equal to the HSMD value within relatively large error bars; this result required 88 h CPU time on our computers (see below), where most of the time was devoted to generating the two samples of 8000 configurations. It should be pointed out that in a detailed study, Chang *et al.*[50] found QH to be unreliable when used in Cartesian coordinates or applied (in internal coordinates) to several microstates; on the other hand, it was found suitable for treating a single microstate, while the convergence of the results is slow and large samples are typically needed. Still, entropy differences ΔS^{QH}_{loop} are expected to be reliable (see also Ref. 51). We also provide in the table the value of $T\Delta S_{loop}$ obtained with procedure II using only one bin value of 4° for all angles; the result $T\Delta S_{loop} = 5.3 \pm 1.0$ kcal/mol is close to that obtained with procedure I.

The error bars in Table II (and in Tables III and IV) are SD/($n_s^{1/2}$), where SD is standard deviation and n_s is the number of structures. Using this formula is justified because every successive pair of the (40) reconstructed structures is separated by 50 ps along the MD trajectory, and thus the energy and entropy of these structures can be considered as uncorrelated. Notice that 40 structures were reconstructed also in Ref. 26; to verify further that this number is adequate, we have calculated results for $n_s = 20$ and found them to agree within the error bars to those obtained by $n_s = 40$. Reconstruction of a single loop conformation with 250 water molecules based on 10^4 future configurations (60 ps) with 2 ps equilibration requires ∼30 h on a single core of a Quad-Core AMD Opteron(tm) Processor 2380 (∼2500 MHz); however, we have shown that the correct result for $T\Delta S_{loop}$ is already obtained from a reconstruction based on 6 ps (1000 future configurations), which requires only 3 h CPU time. The longer calculations of $n_f = 10^4$ were performed for validating convergence. In general, the required n_f will increase with loop size; in Ref. 26, for example, where a smaller loop of acetylcholineesterase has been studied, the maximal value of n_f is 2000.

D. Contribution of water to the free energy

This contribution is obtained by calculating $-F^{TI}_{water}(m)$ [Eq. (17) and Fig. 3], i.e., by eliminating the loop-water interactions in a TI process keeping the template and loop fixed, as described in Sec. II E. Thus, starting from the complete loop/template/water system, the loop-water interactions were gradually annihilated using a parameter λ, where the electrostatic interactions were removed first followed by the removal of the LJ interactions (in the presence of zero electrostatic interactions). This TI was carried out independently for each of the 40 reconstructed loop structures using soft-core potentials defined by parameters λ and δ; thus, the LJ potential ϕ (based on the usual LJ parameters σ and ε) becomes[52]

$$\phi(r_{ij},\lambda) = \lambda 4\varepsilon\left[\frac{\sigma^{12}}{\left(r_{ij}^2+\delta(1-\lambda)\right)^6} - \frac{\sigma^{6}}{\left(r_{ij}^2+\delta(1-\lambda)\right)^3}\right]. \quad (21)$$

The integration is based on 22 windows (λ = 0.95, 0.90, 0.85, . . . , 0.10, 0.05, 0.03, 0.01, 0.00), i.e., altogether 44 windows for both interactions. The higher density of λ points close to λ = 0 is necessary to accurately integrate the larger LJ fluctuations in this region. As in our previous study,[26] integration step i starts by minimizing the last structure of the previous step, $i-1$, according to the potential energy of the current $\lambda(t)$, followed by a 5 ps equilibration, which uses the set of velocities of the last $i-1$ structure (for λ = 0.95 we used a longer equilibration of 15 ps). After equilibration, a 20 ps production run is performed.

The average TI results (over 40 structures) for F^{TI}_{water}(ch), F^{TI}_{water}(LJ), their sum, F^{TI}_{water} [Eq. (17)], (for the open and closed microstates), and the difference, ΔF_{water} [Eq. (18)], between F^{TI}_{water}(o) and F^{TI}_{water}(c) are presented in Table III. (Note that while the interactions are eliminated by TI, the signs of the above free-energy functionals are reversed describing a TI procedure where the loop-water interactions are *increased* from zero.) The LJ results were calculated for several δ values, and those for $\delta = 2$ also for windows of 20 and 40 ps, where the larger window time was also used in the electrostatic integrations. All these tests enable one to estimate more reliably the errors, in addition to the error estimation provided by the standard deviation.

It should be pointed out that in a typical application of TI (and free-energy perturbation), ligand **a** is transformed into ligand **b** in both the active site of a protein (**P**) and the solvent

TABLE IV. Contribution of the loop and water to the energy, entropy, and free energy (in kcal/mol) of the open and closed microstates.[a]

	E_{water}	TS_{water}	F^{TI}_{water}	E_{loop}	TS_{loop}	F_{loop}
Open	–2609.3 ± 3.6	–2602.4 ± 3.8	–6.9 ± 1.0	–200.8 ± 1.5	149.7 ± 0.6	–350.5 ± 1.6
Closed	–2689.0 ± 3.5	–2607.7 ± 3.7	–81.3 ± 0.9	–103.2 ± 1.3	145.8 ± 0.7	–249.0 ± 1.4
Open-closed	$\boldsymbol{\Delta E_{water}}$	$\boldsymbol{T\Delta S_{water}}$	$\boldsymbol{\Delta F_{water}}$	$\boldsymbol{\Delta E_{loop}}$	$\boldsymbol{T\Delta S_{loop}}$	$\boldsymbol{\Delta F_{loop}}$
	79.7 ± 4.9	5.3 ± 5.1	74.4 ± 1.3	–97.6 ± 1.7	3.9 ± 1.0	–101.5 ± 1.9
		E_{total}	TS_{total}	F_{total}		
	Open	–2810.1 ± 1.6	–2452.7 ± 1.9	–357.4 ± 1.8		
	Closed	–2792.2 ± 1.6	–2461.9 ± 1.7	–330.3 ± 1.6		
	Open-closed	$\boldsymbol{\Delta E_{total}}$	$\boldsymbol{T\Delta S_{total}}$	$\boldsymbol{\Delta F_{total}}$		
		−17.9 ± 1.6	**9.2 ± 2.1**	**−27.1 ± 2.0**		

[a]The water and loop energies, E_{water} and E_{loop}, are defined in Eq. (1). F^{TI}_{water} [Eq. (17)] is the water free energy obtained by a TI procedure. The loop entropy TS_{loop} [Eqs. (13) and (14)] and its difference $T\Delta S_{loop}$ [Eq. (15b)] are taken from Table II; $F_{loop} = E_{loop} - TS_{loop}$. $T\Delta S_{water}$ is obtained from $\Delta E_{water} - \Delta F_{water}$. E_{total} [Eq. (1)] and ΔE_{total} are the total energy and its difference for the open and closed microstates. F_{total} is the sum of the loop and water free energies and its difference is ΔF_{total} [Eq. (19)]; $T\Delta S_{total}$ is obtained from $\Delta E_{total} - \Delta F_{total}$. The errors are defined in Table III. Entropies and free energies are defined up to additive constants, which are expected to be equal for both microstates.

environment, which leads to the free-energy differences, $\Delta F_{\mathbf{Pa,Pb}}$ and $\Delta F_{\mathbf{a,b}}$, respectively. The success of this method lies in the fact that only the interactions of the mutated part of the ligand with the environment are *directly* considered and the fluctuations are therefore small. However, conformational changes in the entire protein (e.g., "jumps" of side chains among rotamers) occur constantly and the results might not converge for long simulation times; in other words, the microstate of **Pb** (and to some extent also of **Pa**) keeps changing as the simulation time increases. This is the main source of errors, which are commonly assessed by calculating ΔF also in the reverse direction, i.e., in going from **b** to **a**.

In our case, however, this problem does not exist because the conformations of both the protein (template) and the loop are kept fixed during TI (where the loop-water interactions are gradually eliminated). Furthermore, we carry out many such TI processes (40 in this paper) and average the results, where the fluctuations define the errors; therefore, performing reversed TI runs is not needed (they might also lead to further complexity; see below). Indeed, Table III shows that the errors of all the free-energy components are relatively small and for the open (closed) conformation the results for each component (based on different parameters) can differ by up to 6 kcal/mol. However, the corresponding "open" and "closed" results are highly correlated as both change in the same direction and thus their differences, ΔF_{water} (which is our main interest), are very stable with much smaller deviations of ∼1 kcal/mol. We obtain $\Delta F_{water} = 74.4 \pm 1.3$ kcal/mol, i.e., the free energy due to water is lower for the closed loop than for the open one.

Returning to the reversed TI, it should be pointed out that in the initial preparation of the systems, some water molecules typically become trapped within the template and they remain there during the generation of the MD trajectories, the reconstruction simulations, and the TI runs based on *eliminating* the loop-water interactions. Therefore, these waters can be considered as part of the fixed template. However, in the equilibration of water, which is the initial step of a *reversed* TI process (under zero water-loop interactions), a new set of trapped water molecules might be created which would lead to changes in the TI results. This undesirable effect can in principle be prevented but would require applying additional computational means.

The LJ results in Table III can intuitively be explained by considering an LJ integration where the loop-water interactions are increased (rather than decreased) from zero to their full value. In this process, the water molecules should create a void for the loop which requires investing energy. The results show that larger energy is required for the open loop (46.5), which is highly exposed to water, than for the closed loop (34.9 kcal/mol), which tends to lie against the template. Explaining the different results for the electrostatic integrations is not straightforward. The integration time per frame is ∼140 h CPU on a single processor.

E. Combined results for the entire system

In the upper part of Table IV, we summarize the contributions of the loop and water (and implicitly also of the template) to the total energy, entropy, and free energy. The water contributions are the energy E_{water} [Eq. (1)] (including water-water, water-loop, and water-template interactions) and the free energy F^{TI}_{water} [Eq. (17) and Table III], from which the water entropy is calculated, $TS_{water} = E_{water} - F^{TI}_{water}$. The loop contributes its energy, E_{loop} [Eq. (1)] (which includes the loop-loop and loop-template interactions), and the reconstructed entropy, TS_{loop} [Eq. (14) and Table II], which both lead to our definition of the loop's free energy, $F_{loop} = E_{loop} - TS_{loop}$. Notice again that S and F are defined up to additive constants and thus only their differences ΔS and ΔF are physically meaningful; thus, the differences of all these parameters between the open- and closed-loop conformations are also provided in the table. Finally, in the lower part of the table, results are presented for the entire loop/water/template systems, i.e., for the total energy, entropy, and free energy, and their differences. It is seen again that the errors of the differences are smaller than the sum of errors of the corresponding open and closed properties, in some cases again due to systematic error cancelation.

It should be noted that the contributions of $T\Delta S_{water}$ to ΔF_{water} and $T\Delta S_{loop}$ to ΔF_{loop} are relatively small, being around 7% and 4%, respectively, i.e., most of the

contribution is due to the corresponding energy differences, ΔE_{water} and ΔE_{loop}. However, these entropy differences are positive, meaning that the *total* entropy of the open system is larger than that of the closed one, in accord with the elevated B factors and conformational disorder observed in crystal structures of unbound sreptavidin.[30] Interestingly, $T\Delta S_{\text{total}} = 9.2$ kcal/mol [the sum of 3.9 (loop) and 5.3 (water)] contributes significantly to ΔF_{total}, about 34%, because ΔE_{water} and ΔE_{loop} are of opposite signs and the absolute value of their sum (ΔE_{total}) is relatively small. This demonstrates again that, in general, ΔE alone is not a reliable criterion of stability since (as in the present case) entropic effects could be significant.[13,14,23,25,26]

Notice that ΔE_{total} is negative whereas $T\Delta S_{\text{total}}$ is positive, while one would expect these quantities to share the same sign, as lower energy is generally correlated with lower entropy. However, as pointed out earlier, the higher entropy of the open-loop system agrees with the crystal structures of the unbound protein. Also, the fact that in most of these unbound structures the loop is open suggests that the open microstate is more stable than the closed one, and according to our calculations, this higher stability is gained from both a higher entropy and a lower energy and hence a lower total free energy, $\Delta F_{\text{total}} = -27.1 \pm 2.0$. The fact that the structure of the closed loop is attached to the template of an unbound protein might also contribute to the decrease of ΔF_{total}.

IV. SUMMARY AND CONCLUSIONS

In a previous publication,[39] we capped loops with an increasing number of TIP3P water molecules and studied the effect of this number on the loop stability by monitoring the RMSD of the loop (from its x-ray structure) along MD trajectories. We have found that for several loops, a minimal number of ~10 water molecules have already stabilized the structure. However, the *general* validity of this somewhat surprising result was tested further in a subsequent study[26] with respect to a stricter free-energy criterion, where HSMD-TI was applied to a loop of the protein AChE. It has been found that the loop must be well soaked in water with the density of bulk water; thus, to cap the relatively large loop of streptavidin, at least 250 water molecules are needed, which makes the entire system the largest treated thus far with HSMD-TI. Testing HSMD-TI as applied to systems of increasing size is important because the fluctuation of the absolute entropy (and energy) increases as $N^{1/2}$ with the number of atoms, N.

Indeed, applying the loop reconstruction part of the method HSMD-TI required longer simulations. Thus, we have found that to obtain well defined (i.e., normalized or close to normalized) probabilities, which lead to the expected decrease of $S^{\text{A}}_{\text{loop}}(n_f)$, one needs significantly longer equilibration (~16 ps) than that applied in previous studies of smaller loops (2.5 ps);[13,14,23,25] the reconstruction simulations (~60 ps) are also significantly larger than those used before (e.g., 12.5 ps). The TI component of the method also appears in methodologies for calculating the absolute free energy of binding.[5–10,53] In this procedure, one should check, in particular, that the gradual elimination in the loop-water LJ interactions is adequately performed, which has been verified for the present larger loop/template/water system.

We initially carried out MD simulations of the entire tetramer in a box of water. While these runs were relatively short, they suggest that both the open and closed conformations have considerable local stability at room temperature even without biotin, and this in turn may suggest that the conformational response of the loop to ligand binding is of an induced- rather than a selected-fit type. However, longer simulations are needed to corroborate this point.

Our finding that the open loop is more stable than the closed one (by $\Delta F_{\text{total}} = F_{\text{open}} - F_{\text{closed}} = -27.1 \pm 2.0$ kcal/mol) is expected since experimentally the open conformation is found in most of the crystal structures of unbound streptavidin (the closed loops observed in two subunits are caused by crystal-packing interactions[30]). While this value seems large, it should be pointed out that $\Delta F_{\text{loop}} = -25.2 \pm 7$ kcal/mol has been obtained recently (using a different method) for the mobile loop of avidin;[53] also, Lazaridis *et al.*[54] have estimated the effect of the conformational changes in avidin upon its binding of biotin (using implicit solvent) to find an energy of ~ –21 kcal/mol, where most of the conformational change is attributed to the loop (avidin and streptavidin are proteins with similar structures). As mentioned earlier, the fact that both loop conformations are attached to the same *open* template also contributes to the higher stability of the open loop. Therefore, one would seek to study each loop attached to its own template. However, a recent application of HSMD-TI to the loop of AChE using two templates has shown that such a comparison might be problematic partially due to the different accuracy of the corresponding x-ray structures.[55]

HSMD-TI provides the entropy of the loop and water as by-products of the simulation. In a recent paper, Singh and Warshel[56] showed that various components of the entropy (e.g., solvation, hydrophobic) can be obtained by applying and releasing harmonic restraints. In the present paper, the contribution of $T\Delta S_{\text{total}}$ (9.2 kcal/mol) to ΔF_{total} is significant at about 34%, which demonstrates that, in general, ΔE alone is not a reliable criterion of stability. Also, ΔE_{total} is negative whereas $T\Delta S_{\text{total}}$ is positive, while one would expect these quantities to share the same sign. Thus, the higher stability of the open loop is a contribution of both higher entropy and lower energy.

On the technical side, we have described in detail the implementation of two reconstruction procedures, and developed a set of programs for analyzing the reconstruction results. Thus, all the reconstructed structures were generated within TINKER and were saved on files for a later analysis based on a set of general programs written in C. This separation between the MD simulation and the analysis allows one to use any MM/MD package and to carry out a flexible analysis (without the need for additional simulations) where parameters (e.g., bin size) can be changed and their effect on the results can be studied.

One advantage of HSMD-TI is the fact that the free energy can be obtained from a small number of structures, in principle even from any single structure. Indeed, the main result of this study, $\Delta F_{\text{total}} = -27.1 \pm 2.0$ kcal/mol, is based

on reconstructing only 40 structures [in agreement with our previous calculations of the loop of AChE (Ref. 26)]. The relatively high accuracy of this result stems also from cancellation of errors in entropy and energy differences, hence in free-energy differences. The fact that the number of water molecules is not large enables one to calculate their contribution to the total entropy, ΔS_{total}, which was found to be slightly larger than the contribution of the loop.

Finally, it should be pointed out again that while mobile loops appear in many enzymes, calculating $\Delta F = F_{\text{open}} - F_{\text{closed}}$ by the conventional methods (e.g., TI) is not straightforward, and very few such calculations are available.[57] However, ΔF (without the ligand) should be considered in the calculation of the total absolute free energy of binding, which has not been done thus far. With HSMD-TI, calculation of ΔF is straightforward and reliable, as has been shown for a loop of AChE in our previous paper.[26] In a subsequent project, we are calculating now the absolute free energy of binding of the biotin-streptavidin complex, where the present result $\Delta F = -27.1$ kcal/mol will be taken into account. In the next step in the development of HSMD-TI, we intend to extend it to a protein model where the template is not fixed.

ACKNOWLEDGMENTS

This work was supported by NIH Grant No. 2-R01 GM066090-4 A.

[1] D. L. Beveridge and F. M. DiCapua, Annu. Rev. Biophys. Biophys. Chem. **18**, 431 (1989).
[2] P. A. Kollman, Chem. Rev. **93**, 2395 (1993).
[3] W. L. Jorgensen, Acc. Chem. Res. **22**, 184 (1989).
[4] H. Meirovitch, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd (Wiley-VCH, New York, 1998), Vol. 12, p. 1.
[5] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon, Biophys. J. **72**, 1047 (1997).
[6] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, J. Phys. Chem. B **107**, 9535 (2003).
[7] H. Meirovitch, Curr. Opin. Struct. Biol. **17**, 181 (2007).
[8] H.-X. Zhou and M. K. Gilson, Chem. Rev. **109**, 4092 (2009).
[9] N. Foloppe and R. Hubbard, Curr. Med. Chem. **13**, 3583 (2007).
[10] W. F. Van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. J. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. Van der Vegt, and H. B. Yu, Angew. Chem. Int. Ed. **45**, 4064 (2006).
[11] B. J. Alder and T. E. Wainwright, J. Chem. Phys. **31**, 459 (1959).
[12] J. A. McCammon, B. R. Gelin, and M. Karplus, Nature (London) **267**, 585 (1977).
[13] S. Cheluvaraja and H. Meirovitch, J. Chem. Theor. Comput. **4**, 192 (2008).
[14] S. Cheluvaraja, M. Mihailescu, and H. Meirovitch, J. Phys. Chem. B **112**, 9512 (2008).
[15] R. Elber and M. Karplus, Science **235**, 318 (1987).
[16] F. H. Stillinger and T. A. Weber, Science **225**, 983 (1984).
[17] E. D. Getzoff, H. M. Geysen, S. J. Rodda, H. Alexander, J. A. Tainer, and R. A. Lerner, Science **235**, 1191 (1987).
[18] J. M. Rini, U. Schulze-Gahmen, and I. A. Wilson, Science **255**, 959 (1992).
[19] K. L. Constantine, M. S. Friedrichs, M. Wittekind, H. Jamil, C. H. Chu, R. A. Parker, V. Goldfarb, L. Mueller, and B. T. Farmer, Biochemistry **37**, 7965 (1998).
[20] R. P. White and H. Meirovitch, J. Chem. Phys. **121**, 10889 (2004).
[21] R. P. White and H. Meirovitch, J. Chem. Phys. **124**, 204108 (2006).
[22] R. P. White and H. Meirovitch, J. Chem. Phys. **123**, 214908 (2005).
[23] S. Cheluvaraja and H. Meirovitch, J. Chem. Phys. **122**, 054903 (2005).
[24] S. Cheluvaraja and H. Meirovitch, J. Phys. Chem. B **109**, 21963 (2005).
[25] S. Cheluvaraja and H. Meirovitch, J. Chem. Phys. **125**, 024905 (2006).
[26] M. Mihailescu and H. Meirovitch, J. Phys. Chem. B **113**, 7950 (2009).
[27] L. Chaiet and F. J. Wolf, Arch. Biochem. Biophys. **106**, 1 (1964).
[28] N. M. Green, Adv. Protein Chem. **29**, 85 (1975).
[29] E. A. Bayer and M. Wilchek, Methods Enzymol. **184**, 49 (1990).
[30] S. Freitag, I. L. Trong, L. Klumb, P. S. Stayton, and R. E. Stenkamp, Protein Sci. **6**, 1157 (1997).
[31] P. C. Weber, D. H. Ohlendorf, J. J. Wendoloski, and F. R. Salemme, Science **243**, 85 (1989).
[32] P. C. Weber, M. W. Pantoliano, and L. D. Thompson, Biochemistry **31**, 9350 (1992).
[33] V. Chu, S. Freitag, I. L. Trong, R. E. Stenkamp, and P. S. Stayton, Protein Sci. **7**, 848 (1998).
[34] W. A. Hendrickson, A. Pahler, J. L. Smith, Y. Satow, E. A. Merritt, and R. P. Phizackerley, Proc. Natl. Acad. Sci. USA **86**, 2190 (1989).
[35] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[36] J. W. Ponder, TINKER—Software Tools for Molecular Design, Version 5.0, 2009, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis.
[37] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, and P. A. Kollman, AMBER 10, 2008, University of California, San Francisco.
[38] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarenden, Oxford, 1987).
[39] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[40] R. P. White and H. Meirovitch, J. Chem. Theory Comput. **2**, 1135 (2006).
[41] D. S. Cerutti, I. L. Trong, R. E. Stenkamp, and T. P. Lybrand, J. Phys. Chem. B **113**, 6971 (2009).
[42] N. Gō and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969).
[43] N. Gō and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[44] M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).
[45] H. Meirovitch and Z. Alexandrowicz, J. Stat. Phys. **15**, 123 (1976).
[46] H. Meirovitch, J. Chem. Phys. **111**, 7215 (1999).
[47] H. Meirovitch, Int. J. Mod. Phys. **1**, 119 (1990).
[48] H. Meirovitch, Phys. Rev. A **32**, 3709 (1985).
[49] L. W. Yang, A. J. Rader, X. Liu, C. J. Jursa, S. C. Chen, H. A. Karimi, and I. Bahar, Nucleic Acids Res. **34**, 24 (2006).
[50] C. E. Chang, W. Chen, and M. K. Gilson, J. Chem. Theory. Comput. **1**, 1017 (2005).
[51] N. Singh and A. Warshel, Proteins **78**, 1724 (2010).
[52] M. Zacharias, T. P. Straatsma, and J. A. McCammon, J. Chem. Phys. **100**, 9025 (1994).
[53] I. J. General, R. Dragomirova, and H. Meirovitch, J. Phys. Chem. B (in press).
[54] T. Lazaridis, A. Masunov, and F. Gandolfo, Proteins **47**, 194 (2002).
[55] M. Mihailescu and H. Meirovitch, Entropy **12**, 1946 (2010).
[56] N. Singh and A. Warshel, Proteins **78**, 1705 (2010).
[57] M. A. Olson, Proteins **57**, 645 (2004).