JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

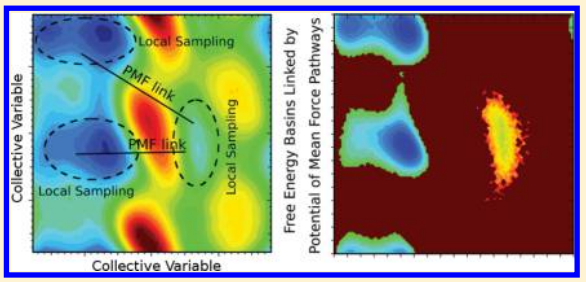
Computing Free Energy Differences of Configurational Basins

Edoardo Giovannelli,[†] Gianni Cardini,[†,‡] Cristina Gellini,[†,‡] Giangaetano Pietraperzia,[†,‡] and Riccardo Chelli*[,†,‡]

[†]Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Italy

[‡]European Laboratory for Nonlinear Spectroscopy (LENS), Via Nello Carrara 1, I-50019 Sesto Fiorentino, Italy

ABSTRACT: A simulation-based approach is proposed to estimate free energy differences between configurational states *A* and *B*, defined in terms of collective coordinates of the molecular system. The computational protocol is organized into three stages that can be carried on simultaneously. Two of them consist of independent simulations aimed at sampling, in turn, *A* and *B* states. In order to limit the evolution of the system around *A* and *B*, biased sampling simulations such as umbrella sampling can be employed. These simulations allow us to estimate local configuration integrals associated with *A* and *B*, which can be viewed as vibrational contributions to the free energy. Free energy evaluation is completed by the linking-path stage, in which the potential of mean force difference is estimated between two arbitrary points of the configurational surface, located the first around *A* and the second around *B*. The linking path in the space of the collective coordinates is arbitrary and can be computed with any method, starting from adaptive biasing potential/force approaches to nonequilibrium techniques. As an illustrative example, we present the calculation of free energy differences between conformational states of the alanine dipeptide in the space of backbone dihedral angles. The basic advantage of this method, that we term "path-linked domains" scheme, is to prevent accurate calculation of the whole free energy hypersurface in the space of the collective coordinates, thus limiting the statistical sampling to a minimum. Path-linked domains schemes can be applied to a variety of biochemical processes, such as protein−ligand complexation or folding-unfolding interconversion.



1. INTRODUCTION

In a system with many degrees of freedom, a conformational state can be classified as a subensemble of the phase-space states accessible to the particles of the system. The relative chemical stability of two conformational states can always be formulated in terms of free energy difference, whose estimate can be made at different approximation degrees. For example, the simplest way to estimate the free energy for the equilibrium of α and β anomers of monosaccharides is to calculate the difference between the lowest energies of the anomers. Of course, this approach ignores entropic effects due to the fact that, first, there are multiple configurations of both α and β anomers and, second, the individual conformations are not statically confined to the bottom of their energy well but exhibit large dynamic diversity in terms of conformational changes limited to the shape of that energy well. Note that glucose, for example, possesses literally hundreds of low-energy conformations of both anomeric states. We may proceed to a better approximation in conformational free energy estimate, including multiple levels in defining the state of a system. In what follows, we introduce the concept of system microstate, corresponding to a specific vector $\mathbf{x}$ in the $3N$-dimensional space of the coordinates relative to N atoms. A set of microstates sharing established chemical and structural properties will be referred to as a configuration. This may define a set of microstates relative to a given molecule which share specific structural features but may identify also the chemical entity itself (such a definition may be employed in the study of intramolecular and intermolecular reaction equilibria). Less generally, a configuration characterized by specific structural or geometrical features for a given chemical connectivity of the system will be called system conformation or conformational state. Note that such conformations can represent multiple configurations of stereoisomers, related to each other by some geometrical criterion, such as holding axial or equatorial substituents. Within this definition, a simple statistical mechanics calculation can then be used to estimate the free energy difference between the two conformational states.

In a classical molecular system, the ultimate approach for calculating the free energy difference ΔF between two conformational states at fixed temperature T and volume V involves the evaluation of two configuration integrals:

$$Q_A = \int_{\mathcal{V}_A} e^{-\beta U(\mathbf{x})} d\mathbf{x}, \quad Q_B = \int_{\mathcal{V}_B} e^{-\beta U(\mathbf{x})} d\mathbf{x} \tag{1}$$

$$\Delta F_{AB} = F_B - F_A = -\beta^{-1} \ln \frac{Q_B}{Q_A} \tag{2}$$

where $U(\mathbf{x})$ is the system potential energy, dependent on the atomic coordinates $\mathbf{x}$, and $\beta^{-1} = k_B T$, with k_B being the

Received: March 16, 2015
Published: July 7, 2015

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

Boltzmann constant. The symbols $\mathcal{V}_A$ and $\mathcal{V}_B$ represent the configurational volumes of the states A and B, which contain the microstates featuring the two states. We point out that eqs 1 and 2 are even more general than what discussed here. In fact, they hold for a generic equilibrium $A \rightleftharpoons B$ involving chemical species whose interconversion may occur through a chemical process; it does not matter if it is a conformational change or a chemical reaction. These concepts have been widely applied to protein−ligand binding free energy calculations.[1−3]

Direct evaluation of the ratio between configuration integrals for problems with high dimensionality is a challenge for most simulation methods also when developed for this specific aim. Indirect approaches utilizing various simulation techniques based on the free energy perturbation[4] and a smart Monte Carlo method termed jump-between-wells[5] have been used to calculate conformational free energy differences. The jump-between-wells method coupled with molecular dynamics (MD) involves the direct monitoring of populations of various configurations in the two conformation domains. This occurs in a simulation in which configurational interconversions are frequent, producing converged, Boltzmann-weighted ensembles of conformational states. Methods also emerged that involve the direct evaluation of the configuration integral as sums over conformational minima. Most notably, a method termed "mining minima"[6] has been introduced in which the configuration integral is evaluated over the "soft modes" identified by torsion angles. It should be stressed, however, that the exclusion of "hard modes" such as bond lengths and bond angles is, in general, a poor approximation. Cyclic structures, for example, undergo sufficient variation of their ring bond angles and even bond lengths during conformational interconversions to give a significant contribution to the conformational free energy. Recently, methods based on noninstantaneous Monte Carlo moves[7,8] have also been proposed to overcome high free energy barriers connecting different configurational/conformational states of a system.

In MD or Monte Carlo simulations, the estimate of the ratio of configuration integrals Q_A and Q_B requires the definition of the conformational states as two subspaces of the whole hypersurface of the relevant torsional coordinates. For the simple case of the alanine dipeptide, used here as an example of our approach, the relevant torsional coordinates are the ϕ and ψ angles of the peptide (Figure 1). In this space of coordinates, in spite of the relative complexity of the free energy surface, we may identify several subspaces which correspond to different conformational domains. The hard issue of the problem is that, in order to calculate the configuration integrals, the system must explore the whole torsional space, with the risk of being trapped in some deep free energy minima with dramatic consequences on the quality of sampling. Even if this formidable task could be accomplished with effective sampling schemes such as serial or parallel generalized ensemble methods,[9] it could be advantageous to restrain sampling (as much as possible) to the conformational volumes $\mathcal{V}_A$ and $\mathcal{V}_B$ for evaluating, separately, a sort of "site free energy contributions" for the two states. As results of independent simulations, such site free energies are not comparable each to the other directly. However, as we will show, it is possible "to connect" these site free energies through potential of mean force (PMF) differences calculated along an established collective coordinate connecting the volume $\mathcal{V}_A$ to $\mathcal{V}_B$, or vice versa. In such a way, it is possible to compute conformational free energy differences by sampling selectively the regions of interest and delineating a path from one to the other configuration domain.

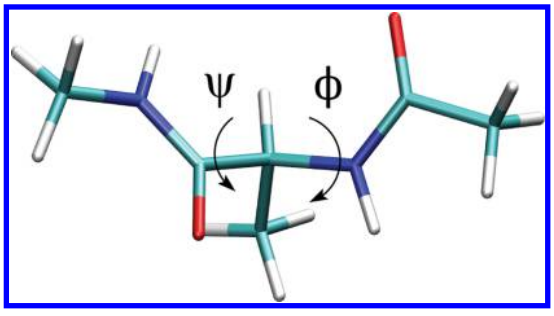


Figure 1. Representation of the ϕ and ψ dihedral angles in the alanine dipeptide.

This is in summary the scheme proposed in this article, that we term Path-Linked Domains (PLD) scheme. The methodology, though applied here to a system at constant volume and temperature, can be extended straightforwardly to the constant temperature and pressure ensemble, which is more relevant to application studies.

In Section 2, the method is presented. In Section 3, alanine dipeptide system and simulation details are described, while results of simulation tests are reported in Section 4. Concluding remarks are given in Section 5.

2. THEORY

Let us consider a dilute solution of a flexible molecule at constant volume and temperature, under the hypothesis that we are able to identify two conformational states of the molecule, say A and B, by means of some structural or energetical criterion or a combination of the two. The constant K_{eq} of the chemical equilibrium $A \rightleftharpoons B$ is defined as the ratio of the molar fractions n_A and n_B of the A and B states, i.e., $K_{eq} = n_B/n_A$. From the statistical point of view, n_A and n_B correspond to the probabilities of finding the solute molecule in A and B conformations and therefore are proportional to the configuration integrals Q_A and Q_B of eq 1. This allows us to write the equilibrium constant as $K_{eq} = Q_B/Q_A$ and the free energy difference of the two conformers as $\Delta F_{AB} = -\beta^{-1} \ln (Q_B/Q_A) = -\beta^{-1} \ln K_{eq}$. Therefore, the basic quantities needed to evaluate K_{eq}, or equivalently ΔF_{AB}, are the configuration integrals Q_A and Q_B. Calculating these integrals by, e.g., conventional equilibrium simulations or accelerated sampling techniques[9,10] may not be easy, especially due to the difficulty of reaching converging estimates in the high dimensional space of the atomic coordinates $\mathbf{x}$ of the system.[11] In practice, we cannot evaluate numerically the configuration integrals but only their ratio relative to subsets $\mathcal{V}_A$ and $\mathcal{V}_B$ of the atomic coordinates explored in the same simulation.

As conformational states A and B refer to the same molecule, it is generally easier to define configuration domains in the M-dimensional space of independent dihedral angles characterizing the conformers ($M = 2$ in the case of the alanine dipeptide; see Figure 1). Thus, defining the vector of the relevant dihedral angles as[12] $\boldsymbol{\zeta}(\mathbf{x}) = (\zeta_1(\mathbf{x}), \zeta_2(\mathbf{x}), ..., \zeta_M(\mathbf{x}))$, the configuration integral Q_A can be written as

$$Q_A = \int_{D_A} \int_{\mathcal{V}} \delta(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x})) e^{-\beta U(\mathbf{x})} d\mathbf{x}\, d\mathbf{z} \quad (3)$$

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

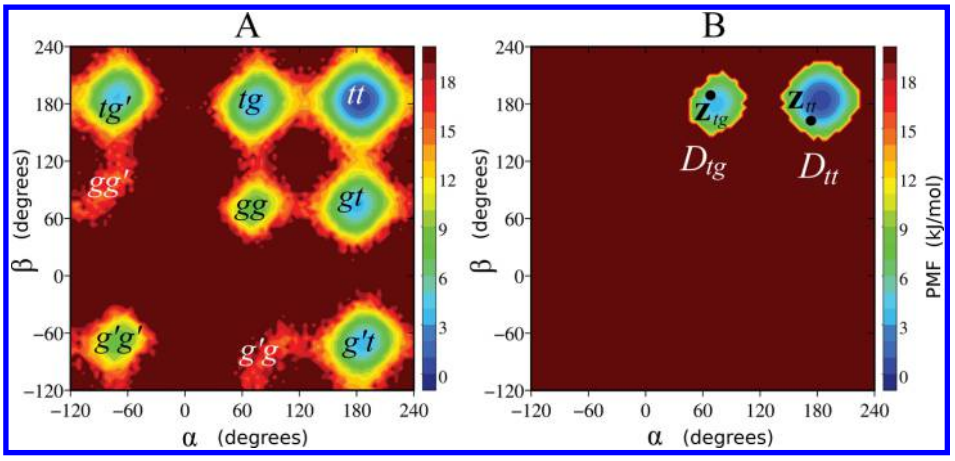


Figure 2. (A) $\Phi(\mathbf{z})$ surface of *n*-pentane as a function of the two inner C−C−C−C torsion angles α and β, obtained from a constant-pressure constant-temperature simulation of 64 molecules at room conditions. (B) Integration domains D_{tt} and D_{tg}, defined as the portions of the $\mathbf{z}$ space around the corresponding minima such that $\Phi(\mathbf{z})$ is less than 10 kJ mol^{-1} above the minimum inside the D_{tt} domain. Domains D_{tg}/D_{tt} and points $\mathbf{z}_{tg}/\mathbf{z}_{tt}$ correspond to D_A/D_B and $\mathbf{z}_a/\mathbf{z}_b$ of eq 6, respectively.

where the integral over $\mathbf{x}$ is extended to the whole space $\mathcal{V}$ of the atomic Cartesian coordinates, and the integral over $\mathbf{z}$ is extended to a limited domain D_A of the dihedral angle coordinates, $\zeta_1(\mathbf{x})$, $\zeta_2(\mathbf{x})$, etc., for which the molecule is classified to belong to the conformational state *A*. Note that the integral over $\mathbf{x}$ implicitly includes integration over the solvent coordinates. In eq 3, we recognize the PMF as a function of the dihedral angles:

$$\Phi(\mathbf{z}) = -\beta^{-1} \ln\left(\int_{\mathcal{V}} \delta(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x}))\mathrm{e}^{-\beta U(\mathbf{x})}\mathrm{d}\mathbf{x}\right) \tag{4}$$

In general, the PMF is a function of a multidimensional vector $\mathbf{z}$ representing the dihedral angles. Although the dimensions of $\mathbf{z}$ can be very large in dependence of the size and flexibility of the molecule, we are often interested to a limited set of dihedral angles, greatly simplifying the problem (e.g., the ϕ and ψ angles of a peptide). This reduction of dimensionality does not lie in some intrinsic property of the system but rather in the paradigm chosen to describe the physics of the system. Evaluation of the free energy difference between two states is a rather general problem,[2,13−15] which is related not only to sampling issues but also to the arbitrariness of the criterion adopted to define the "state of a system". In the following, we will not take care on how the $\mathbf{z}$ space is defined, but we will develop the treatment under the assumption that such a space can somehow be specified. Moreover, we also assume that some criterion can be found for defining and choosing the conformational domain D_A appearing into eq 3. These assumptions make the treatment quite general, so that referring to torsional degrees of freedom must only be viewed as a way of connecting the theory to the most important situation in which, we believe, this method could be applied.

Upon substitution of eq 4 into eq 3, we obtain

$$Q_A = \int_{D_A} \mathrm{e}^{-\beta\Phi(\mathbf{z})}\mathrm{d}\mathbf{z} \tag{5}$$

Therefore, the problem can be traced back to the calculation of $\Phi(\mathbf{z})$ and then to numerically integrate its exponential $\exp(-\beta\,\Phi(\mathbf{z}))$ over the domain D_A (or D_B, in the case of Q_B). The calculation of the integral of eq 5 could be computationally very expensive even in relatively simple molecules, basically due to the rapid increase in the number of torsional degrees of freedom with the number of atoms in the molecule. This leads to a manifold of minima in the $\Phi(\mathbf{z})$ hypersurface, which needs to be properly sampled in a computer simulation. For example, in *n*-alkanes, the PMF minima roughly increase with the number *c* of carbon atoms as 3^{c-3}, including symmetrically equivalent conformations. The representation of the $\Phi(\mathbf{z})$ surface of *n*-pentane as a function of the two inner C−C−C−C dihedral angles, obtained from a constant-pressure constant-temperature MD simulation of 64 molecules at room conditions, is reported in Figure 2A. Minima, corresponding to the conformers *tt*, *tg*, *gt*, *tg′*, *g′t*, *gg′*, *g′g*, *g′g′*, and *gg*, appear in well-defined regions of the PMF surface. Supposing that we are interested to the free energy difference between *tt* and *tg* conformers, we could select the integration domains D_{tt} and D_{tg} as the portions of the $\mathbf{z}$ space around the corresponding minima, such that $\Phi(\mathbf{z})$ < 10 kJ mol^{-1}. This would give the reduced conformational domains of the *tt* and *tg* states shown in Figure 2B. Of course, a geometrical criterion based on the accessible values of the α and β dihedral angles (e.g., $0° < \alpha < 120°$ and $120° < \beta < 240°$ for the *tg* state) would also be suitable to establish the conformational states of interest. However, independently of the criterion adopted to define the $\mathbf{z}$ domains, an estimate of the whole PMF should be provided to evaluate the configuration integrals, which could result in a too expensive calculation. Alternatively, one could compute the ratio Q_B/Q_A with separate MD simulations. The procedure is based on a manipulated expression of ΔF_{AB}, obtained by applying eq 5 into eq 2:

$$\Delta F_{AB} = \Delta\Phi_{AB} - \beta^{-1} \ln\left(\frac{\int_{D_B} \mathrm{e}^{\beta[\Phi(\mathbf{z}_b)-\Phi(\mathbf{z})]}\mathrm{d}\mathbf{z}}{\int_{D_A} \mathrm{e}^{\beta[\Phi(\mathbf{z}_a)-\Phi(\mathbf{z})]}\mathrm{d}\mathbf{z}}\right) \tag{6}$$

where $\Delta\Phi_{AB} = \Phi(\mathbf{z}_b) - \Phi(\mathbf{z}_a)$, with $\mathbf{z}_a$ and $\mathbf{z}_b$ being two any points of the conformational space located inside the D_A and D_B domains, respectively. In the *n*-pentane example, these points could be, for instance, those reported in Figure 2B. The integrals of eq 6 can be calculated by two independent MD simulations supplied with biasing potentials,[10] which enforce, in turn, the sampling of restricted regions, R_A and R_B, encompassing D_A and D_B domains. We will denote these MD simulations, as eq-R_A and eq-R_B. In practice, the integrals into eq 6 are approximated by sums over points taken on a regular

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

grid in the $\mathbf{z}$ space. Thus, for example, $\int_{D_A} e^{\beta[\Phi(\mathbf{z}_a)-\Phi(\mathbf{z})]}d\mathbf{z} \simeq \sum_{\mathbf{z}_i \in D_A} e^{\beta[\Phi(\mathbf{z}_a)-\Phi(\mathbf{z}_i)]}\Delta\mathbf{z}$, where the sum is limited to the grid points inside the D_A domain and $\Delta\mathbf{z}$ is the finite volume whose dimensions correspond to the resolution sizes employed to estimate $\Phi(\mathbf{z})$. The other integral of eq 6 is computed in analogous way. The possibility of computing the integrals of eq 6 with independent MD simulations relies on the fact that they involve exponential functions of relative PMFs rather than absolute PMFs and therefore do not depend on arbitrary additive constants. Choosing $\mathbf{z}_a$ and $\mathbf{z}_b$ inside D_A and D_B domains is necessary because it allows estimating the free energy differences $\Phi(\mathbf{z}_a) - \Phi(\mathbf{z})$ and $\Phi(\mathbf{z}_b) - \Phi(\mathbf{z})$ with single MD simulations that limit sampling to those domains. Actually, a less stringent condition is required in choosing $\mathbf{z}_a$ and $\mathbf{z}_b$; they must be placed inside the R_A and R_B domains, respectively, namely, where it is possible to evaluate the PMF via eq-R_A and eq-R_B simulations. In any case, in order to get statistically accurate outcomes, $\mathbf{z}_a$ and $\mathbf{z}_b$ should belong to intensively sampled regions of the conformational space.

The PMF difference $\Delta\Phi_{AB}$ can be calculated independently by any method to estimate the PMF along an established collective coordinate.[11,16] In this study, we have used nonequilibrium steered MD simulations,[17] but other approaches based on equilibrium or quasi-equilibrium schemes[10,18−20] are suitable as well. We will refer to these simulations as Linking Path (LP) simulations. A simplified representation of the PLD scheme is reported in Figure 3. We

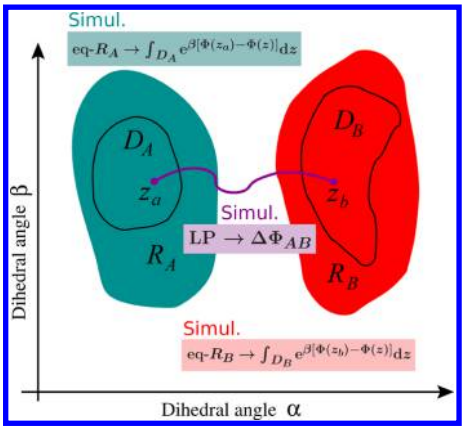


Figure 3. Schematic representation of the PLD protocol. The axes report the components α and β of a generic bidimensional vector $\mathbf{z}$ describing the degrees of freedom of interest. Green and red colors denote the configurational regions R_A and R_B, respectively, visited by the system during the equilibrium simulations eq-R_A and eq-R_B. The conformational domains D_A and D_B are enclosed by solid black lines. The magenta line represents the linking path, employed in LP simulations, connecting the points $\mathbf{z}_a$ and $\mathbf{z}_b$. eq-R_A, eq-R_B, and LP simulations are used to compute the quantities entering eq 6 (also reported in the picture).

stress that the actual advantage of expressing ΔF_{AB} as in eq 6 relies on the possibility of calculating configuration integrals without performing a complete sampling of the conformational space, i.e., without determining the whole $\Phi(\mathbf{z})$ surface.

We will apply eq 6 to the calculation of the free energy difference between conformational basins of the alanine dipeptide by the use of steered MD simulations to calculate $\Delta\Phi_{AB}$ and umbrella sampling[10] (US) simulations for the configuration integrals. A comparable simulation procedure has been devised to compute binding free energies of, e.g., protein−ligand complexes.[2,14]

3. SYSTEM AND SIMULATION DETAILS

3.1. System. Alanine dipeptide consists of an alanine residue terminated by acetyl and N-methyl capping groups (Figure 1). The small dimensions together with a quite complex conformational organization make this peptide perhaps the simplest model bearing most features of polypeptides. Its peculiar behavior is due to the presence of the flexible ϕ and ψ dihedral angles and of functional groups able to establish both intra- and intermolecular hydrogen bonds.[21,22] Although there are only two conformationally relevant degrees of freedom, i.e., the ϕ and ψ angles, the free energy surface $\Phi(\mathbf{z})$ as a function of these angles, $\mathbf{z} \equiv (\phi, \psi)$, is rather complex, presenting various local minima and maxima. The $\Phi(\mathbf{z})$ surface obtained from a US simulation taken from ref 19 is reported in Figure 4.

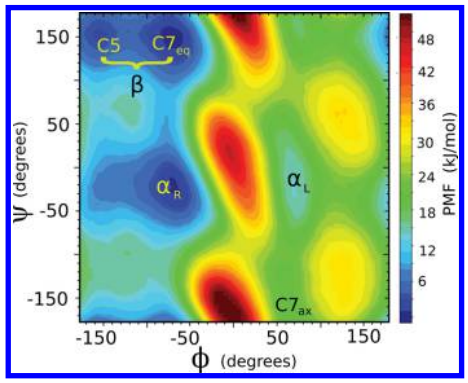


Figure 4. $\Phi(\mathbf{z})$ surface of the alanine dipeptide as a function of the ϕ and ψ dihedral angles. The relevant free energy minima are labeled. In this study, microstates into C5 and $C7_{eq}$ minima are considered to belong to a unique system configuration, labeled β. Calculations have been performed with a US simulation.[19]

Thanks to its modest size, alanine dipeptide has been often employed as a benchmark system[23] to verify sampling methods[19,21,24,25] or to evaluate the accuracy of force fields.[26] We apply the PLD scheme to calculate free energy differences between conformational states, or conformers, whose characterization in terms of ϕ and ψ angles is reported in Figure 4. Specifically, the target free energy differences are $F_{\alpha_L} - F_\beta$ and $F_{\alpha_L} - F_{\alpha_R}$. Note that the conformers C5 and $C7_{eq}$, well distinguishable as separate free energy basins in $\Phi(\mathbf{z})$, are considered here as a unique conformational state, indicated with β. A quantitative definition of the conformational states related to α_L, α_R, and β minima are given in Section 3.3.1. In order to get a feedback on the efficiency of the PLD scheme, the outcomes are compared to those obtained from a Serial Generalized Ensemble (SGE) simulation,[27] whose performances are almost equivalent to those of the popular replica-exchange method.[20,28]

3.2. Shared Simulation Setup. In this study, several MD simulations of equilibrium and nonequilibrium type have been carried out. The equilibrium simulations are (i) the eq-R_A and eq-R_B simulations, realized with conventional and US methods (some further details are given in Section 3.3.1), (ii) the simulations with restrained ϕ and ψ dihedral angles aimed at producing the initial microstates for the steered MD

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

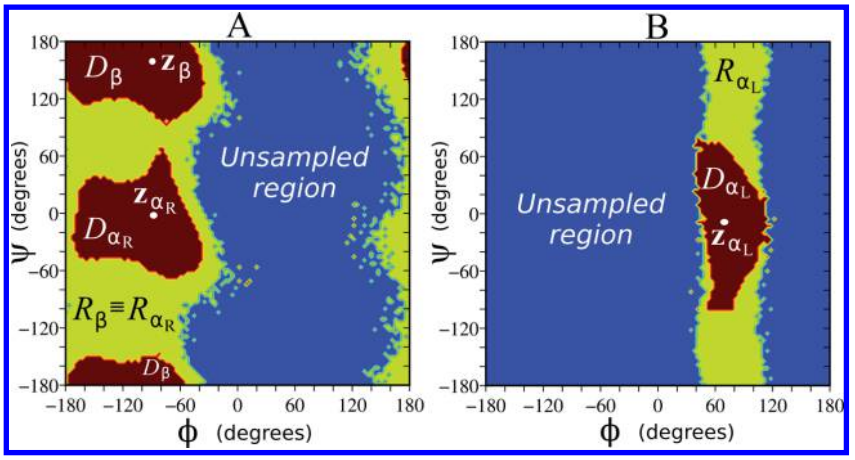


Figure 5. Green: R_x regions explored during eq-R_x simulations. Brown: D_x conformational domains resulting from applying criteria of eqs 7, 8, and 9. Blue: $\mathbf{z}$ configurations unexplored during the eq-R_x simulations. Note that the R_x regions encompass the D_x domains. The $\mathbf{z}_\beta$, $\mathbf{z}_{\alpha_R}$, and $\mathbf{z}_{\alpha_L}$ points connected through LP simulations are also shown. (A) Data related to the eq-R_{α_R} simulation (the same as eq-R_β). (B) Data related to the eq-R_{α_L} simulation.

simulations, and (iii) the SGE simulation performed for a comparative aim. Nonequilibrium simulations are the steered MD simulations (LP simulations) employed to compute the quantity $\Delta\Phi_{AB}$ into eq 6. All these MD simulations share the setup described as follows. The system consists of one alanine dipeptide and 288 water molecules simulated in the constant volume and temperature thermodynamic ensemble using the program ORAC.[29] A cubic box of 21 Å side-length with standard periodic boundary conditions has been adopted. The temperature control (298 K) has been achieved through a Nosé−Hoover thermostat.[30] The dipeptide is modeled by the AMBER03 force field,[26] while TIP3P potential has been used for water.[31] Electrostatics has been accounted for by the smooth particle mesh Ewald method,[32] adopting a fourth order B-spline interpolation polynomial for the charges, an Ewald parameter of 0.43 Å^{-1} and a grid spacing of 0.875 Å for the fast Fourier transform calculation of the charge weighted structure factor. A cutoff distance of 9.5 Å has been set for nonbonded interactions. A five time-step r-RESPA integrator[33] has been used for integrating the equations of motion.

3.3. Path-Linked Domains Scheme. Owing to the high barriers featuring the $\{\phi, \psi\}$ free energy surface of the alanine dipeptide and to the large free energy difference between the α_L and the β/α_R conformations[34] (Figure 4), statistical sampling generated from conventional MD simulations does not allow for quantitative free energy estimates.[20] The PLD approach to the free energy difference between α_L and β/α_R conformational states, i.e., $F_{\alpha_L} - F_{\beta/\alpha_R}$, consists of three independent and hence simultaneously affordable stages: (i) an equilibrium MD simulation in which ϕ and ψ are restrained within a region R_{β/α_R} encompassing the D_{β/α_R} conformational domain (the eq-R_{β/α_R} simulation in our terminology), (ii) the analogous eq-R_{α_L} simulation related to the α_L conformational state, and (iii) a bidirectional set of nonequilibrium steered MD simulations, the LP simulations, for estimating the difference between $\Phi(\mathbf{z})$ values computed at established points, $\mathbf{z}_{\beta/\alpha_R}$, and $\mathbf{z}_{\alpha_L}$, inside the R_{β/α_R} and R_{α_L} regions.

3.3.1. Definition of Conformational States and eq-R_x Simulations. The first problem to handle is to formulate operational definitions of the conformational states of interest, namely, to establish the exact meaning of β, α_R, and α_L conformational states. Obviously, such definitions identify the D_β, D_{α_R}, and D_{α_L} domains and are somehow arbitrary, because in general, $\Phi(\mathbf{z})$ varies smoothly with the set of coordinates $\mathbf{z} \equiv (\phi, \psi)$ chosen to describe the system states. We define the conformational state D_{α_R} as follows

$$D_{\alpha_R}\begin{cases} \mathbf{z} \equiv (\phi, \psi) \in \mathcal{M}_{\alpha_R} \cap \mathcal{N}_{\alpha_R} \\ \mathcal{M}_{\alpha_R} = \{\mathbf{z}: -180° \le \phi \le 0° \wedge -90° \le \psi \le 90°\} \\ \mathcal{N}_{\alpha_R} = \{\mathbf{z}: \Phi(\mathbf{z}) - \min(\Phi(\mathbf{z}')) < 10 \text{ kJ/mol} \wedge \mathbf{z}' \in \mathcal{M}_{\alpha_R}\} \end{cases} \quad (7)$$

Analogous definitions are adopted for D_β and D_{α_L}:

$$D_\beta\begin{cases} \mathbf{z} \equiv (\phi, \psi) \in \mathcal{M}_\beta \cap \mathcal{N}_\beta \\ \mathcal{M}_\beta = \{\mathbf{z}: -180° \le \phi \le 0° \wedge -180° \le \psi \le -100°\} \cup \\ \qquad \{\mathbf{z}: -180° \le \phi \le 0° \wedge 100° \le \psi \le 180°\} \cup \\ \qquad \{\mathbf{z}: 120° \le \phi \le 180° \wedge 100° \le \psi \le 180°\} \\ \mathcal{N}_\beta = \{\mathbf{z}: \Phi(\mathbf{z}) - \min(\Phi(\mathbf{z}')) < 10 \text{ kJ/mol} \wedge \mathbf{z}' \in \mathcal{M}_\beta\} \end{cases} \quad (8)$$

$$D_{\alpha_L}\begin{cases} \mathbf{z} \equiv (\phi, \psi) \in \mathcal{M}_{\alpha_L} \cap \mathcal{N}_{\alpha_L} \\ \mathcal{M}_{\alpha_L} = \{\mathbf{z}: 0° \le \phi \le 120° \wedge -100° \le \psi \le 100°\} \\ \mathcal{N}_{\alpha_L} = \{\mathbf{z}: \Phi(\mathbf{z}) - \min(\Phi(\mathbf{z}')) < 10 \text{ kJ/mol} \wedge \mathbf{z}' \in \mathcal{M}_{\alpha_L}\} \end{cases} \quad (9)$$

Denoting with "x" a generic label α_R, α_L, or β, it is worth noting that, for a given conformational state D_x, the definition of the region $\mathcal{N}_x$ does not depend on the (arbitrary) additive constant featuring $\Phi(\mathbf{z})$ because it relies on a PMF difference rather than an absolute PMF. Here, $\Phi(\mathbf{z})$ in a region R_x is computed through the corresponding eq-R_x simulation. In particular, to determine D_{α_R} and D_β, it was enough to perform a unique conventional MD simulation (i.e., without biasing potential) because the alanine dipeptide has been found to span a wide conformational basin, larger than the extended region $(\mathcal{M}_{\alpha_R} \cap \mathcal{N}_{\alpha_R}) \cup (\mathcal{M}_\beta \cap \mathcal{N}_\beta)$. Therefore, the simulations eq-R_{α_R} and eq-R_β are actually the same. Starting from a dipeptide conformation such that $\mathbf{z} \simeq (-45.7°, 152.7°)$, a simulation lasting 156 ns explored the region highlighted in green in Figure 5A, which corresponds to R_{α_R} or equivalently R_β (the R_A domain in Figure 3). The conformational states D_{α_R}

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

and D_β resulting from applying eqs 7 and 8 are shown in brown in Figure 5A. The PMF has been computed from the two-dimensional $g(\mathbf{z})$ distribution function as

$$\Phi(\mathbf{z}) = -\beta^{-1} \ln g(\mathbf{z}) \tag{10}$$

with angular resolution $\Delta\phi = \Delta\psi = 3.6°$. The D_{α_L} domain has been achieved through a US simulation (eq-R_{α_L}) applying the biasing potential $U_{us}(\mathbf{x}) = k_{us}(\phi(\mathbf{x}) - \phi_0)^2$, where $\phi(\mathbf{x})$ is the ϕ dihedral angle of the microstate $\mathbf{x}$, $\phi_0 = 75°$, and $k_{us} = 1.27 \times 10^{-2}$ kJ mol^{-1} deg^{-2}. This setup leads to sample the R_{α_L} region shown in green in Figure 5B. The corresponding D_{α_L} domain (from eq 9) is highlighted in brown in Figure 5B. The $\Phi(\mathbf{z})$ surface is recovered using eq 10, with $g(\mathbf{z})$ being computed through the reweighting formula[10]

$$g(\mathbf{z}) = \frac{\sum_{i=1}^{N_s} \delta(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x}_i)) e^{\beta U_{us}(\mathbf{x}_i)}}{\sum_{j=1}^{N_s} e^{\beta U_{us}(\mathbf{x}_j)}} \tag{11}$$

where $\boldsymbol{\zeta}(\mathbf{x}_i) \equiv (\phi(\mathbf{x}_i), \psi(\mathbf{x}_i))$ is the vector of the dihedral angles associated with the $\mathbf{x}_i$ microstate, and N_s is the number of microstates sampled in the US simulation. The PMFs limited to the conformational states D_{α_R} and D_β are drawn in Figure 6A,

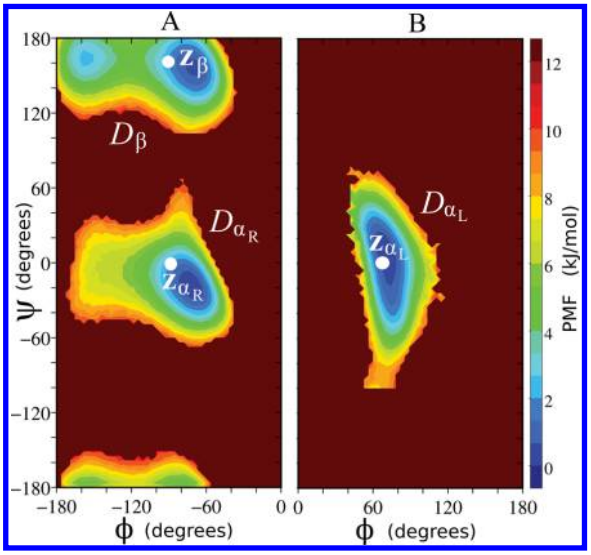


Figure 6. (A) D_{α_R} and D_β conformational domains, recovered from the eq-R_{α_R} simulation (the same as eq-R_β), are highlighted together with the corresponding $\Phi(\mathbf{z})$ free energy profiles. The $\mathbf{z}_\beta$ and $\mathbf{z}_{\alpha_R}$ points are also shown. (B) Same information as panel A recovered from the eq-R_{α_L} simulation. Different panels are used for data from eq-R_{β/α_R} and eq-R_{α_L} simulations because PMFs are shifted by an arbitrary constant.

while the PMF limited to the conformational state D_{α_L} is shown in Figure 6B. Hence, both Figures 6A and 6B display the free energy boundaries employed in eqs 7, 8, and 9 to define $\mathcal{N}_{\alpha_R}$, $\mathcal{N}_\beta$, and $\mathcal{N}_{\alpha_L}$, respectively. We notice that the latter PMF is shifted by an arbitrary constant with respect to the former ones since it is yielded by an independent MD simulation.

3.3.2. LP Simulations. As discussed in Section 2, free energy surfaces computed from independent MD simulations, i.e., the eq-R_x simulations, are quantitatively consistent each to the other only after determining the PMF difference between two any arbitrary points of such surfaces, (the $\mathbf{z}_a$ and $\mathbf{z}_b$ points of eq 6). For the three conformational states under consideration, we have chosen $\mathbf{z}_{\alpha_R} \equiv (-90°, 0°)$, $\mathbf{z}_\beta \equiv (-90°, 160°)$, and $\mathbf{z}_{\alpha_L} \equiv (70°, 0°)$. A view of their positions within the respective D_x domains is reported in both Figures 5 and 6. Once these points are defined, numerical computation of the integrals of eq 6 is straightforward. To calculate $\Delta\Phi_{\alpha_R\alpha_L} = \Phi(\mathbf{z}_{\alpha_L}) - \Phi(\mathbf{z}_{\alpha_R})$, namely, the $\Delta\Phi_{AB}$ quantity of eq 6, we have used steered MD simulations (LP simulations) linking $\mathbf{z}_{\alpha_L}$ to $\mathbf{z}_{\alpha_R}$ with a linear path. Analogous treatment has been adopted for $\Delta\Phi_{\beta,\alpha_L}$. Specifically, an external time-dependent potential $E(\mathbf{z},t)$ is applied in a series of nonequilibrium simulated trajectories to guide the coordinate $\mathbf{z}$ from an initial value $\mathbf{z}_i$ to the final value $\mathbf{z}_f$, according to a defined linear time schedule:

$$E(\mathbf{z},t) = k_{LP} \left| \mathbf{z} - \mathbf{z}_i - \frac{t}{\tau_{pull}} (\mathbf{z}_f - \mathbf{z}_i) \right|^2 \tag{12}$$

where $k_{LP} = 0.255$ kJ mol^{-1} deg^{-2}, and the pulling time τ_{pull} is 9 ps. Clearly, $\mathbf{z}_i$ and $\mathbf{z}_f$ may correspond, respectively, to $\mathbf{z}_{\beta/\alpha_R}$ and $\mathbf{z}_{\alpha_L}$ or vice versa, depending on the pulling direction. Initial system microstates of the guided trajectories have been picked at regular time intervals of 0.6 ps from two equilibrium MD simulations, one for each pulling direction, enforcing an external potential of $k_{LP}\,|\mathbf{z}-\mathbf{z}_i|^2$ type. Then, 6000 trajectories for each direction of the process have been carried out by using inverse time schedules. For each realization of the process, the work performed on the system is computed as

$$W = \int_0^{\tau_{pull}} \frac{\partial E(\mathbf{z},t)}{\partial t} dt \tag{13}$$

The two sets of works are exploited to estimate $\Delta\Phi_{\alpha_R\alpha_L}$ and $\Delta\Phi_{\beta,\alpha_L}$ according to the bidirectional PMF estimator by Minh and Adhib (eq 10 of ref 35; almost identical outcomes have been obtained by using the PMF estimator of ref 36).

3.4. Serial Generalized Ensemble Simulation. Although several implementations of SGE simulation techniques have been provided during the years,[37−41] in our comparative analysis, we have adopted the scheme proposed in refs 20 and 27, which is based on a "on the fly" update of ensemble free energies according to the Bennett acceptance ratio method.[42,43] The simulation run considered here results from extending in time the SGE simulation reported in ref 20 to which reference is made for a detailed description of the simulation setup. In brief, the SGE simulation has been performed with eight replicas of the system evolving independently through a generalized ensemble consisting of eight thermodynamic ensembles, which differ for the intramolecular potential energy of the alanine dipeptide, progressively scaled from 1 to 0.01 (for details on partitioning of the scaling factors among the thermodynamic ensembles, see Table 1 of ref 20). The simulation time per replica is 252 ns. Since the simulation allows exploring the whole $\{\phi, \psi\}$ space, it is possible to compute every possible free energy difference ΔF_{AB} by direct integration of an exponential function of $\Phi(\mathbf{z})$:

$$\Delta F_{AB} = -\beta^{-1} \ln \left(\frac{\int_{D_B} e^{-\beta\Phi(\mathbf{z})} d\mathbf{z}}{\int_{D_A} e^{-\beta\Phi(\mathbf{z})} d\mathbf{z}} \right) \tag{14}$$

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

Note however that, in the calculation of $\Phi(\mathbf{z})$, each microstate contributes according to a variable weight factor. Each weight is determined from the simulation by using the multistate Bennett acceptance ratio methodology[44] and depends on the thermodynamic ensemble in which the corresponding microstate is found.

4. SIMULATION TESTS

The purpose of the present simulation tests is to compare the performances of the PLD scheme to those of a SGE simulation,[20,27] in relation to estimates of conformational free energy differences of alanine dipeptide conformers in aqueous solution. In this respect, it is worth noting that the adopted SGE methodology has already been proved to be comparable in accuracy to the popular replica exchange method,[45−50] as the estimate of $\Phi(\mathbf{z})$ is concerned.[20] Moreover, we point out that it is not our aim here to present the PLD scheme as the best approach to study conformational distributions in peptides, or biopolymers in general, also because no systematic comparison is provided with other important methods for free energy calculations.[16] Rather, we limit our conclusions to observe that, in the treatment of small peptides, the PLD scheme outperforms the quite popular family of generalized ensemble simulations, offering interesting perspectives, alternative to methodologies already in use, for free energy calculations.

Specifically, we report on a comparative analysis of PLD and SGE methods concerning the calculation of free energy differences as a function of sampling times, assuming the outcomes of the US simulation reported in ref 19 as a reference. The computer time τ_{pld} needed to apply the PLD scheme is the sum of the times τ_A, τ_B, and τ_{LP} associated with the eq-R_A, eq-R_B, and LP simulations, respectively. The time τ_{LP} can, in turn, be viewed as a sum of various contributions: $\tau_{LP} = 2N(\tau_s + \tau_{pull})$, where $\tau_s = 0.6$ ps is the time needed to sample a single system microstate taken to initialize a pulling trajectory (the same τ_s is used for forward and backward directions), $\tau_{pull} = 9$ ps is the pulling time defined in eq 12, and N is the number of pulling trajectories in one direction; although not necessary, we have taken the same number of forward and backward trajectories.

In Figure 7, we show $\Delta\Phi_{\alpha_R,\alpha_L}$ and $\Delta\Phi_{\beta,\alpha_L}$ contributions to $\Delta F_{\alpha_R,\alpha_L}$ and $\Delta F_{\beta,\alpha_L}$ (eq 6), respectively, as a function of the number of pulling trajectories N. In both cases, good convergence appears to be reached with about few thousands of trajectories per direction. However, even adopting $N = 200$ (see arrows in Figure 7), estimates within only 0.2 kJ mol^{-1} about the limit values of $\Delta\Phi_{\alpha_R,\alpha_L} \simeq 12.0$ kJ mol^{-1} and $\Delta\Phi_{\beta,\alpha_L} \simeq 12.4$ kJ mol^{-1} are obtained. On the basis of these results, we may consider $N = 4000$ as the number of forward and backward trajectories beyond which convergent estimates are gained. Therefore, in order to simplify the analysis of the dependence of $\Delta F_{\alpha_R,\alpha_L}$ and $\Delta F_{\beta,\alpha_L}$ on the sampling time, we have fixed τ_{LP} to 76.8 ns, which is the time needed to realize 4000 forward and backward pulling trajectories. Note that τ_{LP}, τ_A, and τ_B, as well as τ_{sge}, i.e., the total SGE simulation time,[51] do not account for the equilibration time.

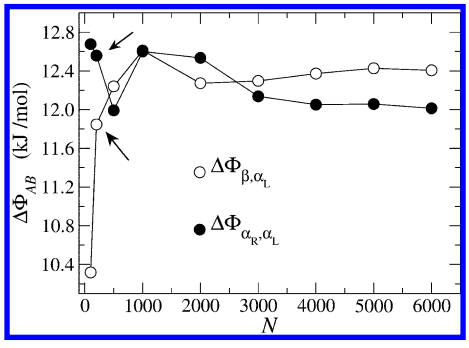


Figure 7. PMF differences $\Delta\Phi_{\alpha_R,\alpha_L}$ and $\Delta\Phi_{\beta,\alpha_L}$ as functions of the number N of pulling trajectories, estimated from LP simulations. The arrows indicate the data corresponding to $N = 200$. Lines are guides for the eyes.

A comparison of the efficiency of PLD and SGE methods is given in Figure 8, where we report $\Delta F_{\alpha_R,\alpha_L}$ and $\Delta F_{\beta,\alpha_L}$ as functions of the sampling times τ_{pld} and τ_{sge}. The free energy differences estimated through the US simulation[19] by using eq 14 are also shown in the figure as a reference. We note that, while free energy deviations of PLD from US do not exceed 0.2 kJ mol^{-1}, SGE estimates of $\Delta F_{\beta,\alpha_L}$ and $\Delta F_{\alpha_R,\alpha_L}$ deviate by about 0.5 and 1 kJ mol^{-1}, respectively. This clearly points to a better accuracy of the PLD scheme with respect to the SGE approach. Such a conclusion is supported from a further observation leading to infer poor convergence of SGE calculations. In fact, in spite of the large SGE sampling time (τ_{sge}) reached in our calculations, which is nearly 1 order of magnitude greater than τ_{pld}, the dependence of SGE free energies on τ_{sge} is clearly featured by a monotonically increasing trend. However, although SGE free energy estimates appear poorly convergent from a comparative standpoint, they are indeed satisfactory, as the deviations from the reference values are of the order of the chemical accuracy.

A detailed view on the reason why the PLD scheme outperforms the SGE method for the system under study can be gained from the analysis of the single terms of eq 6 (PLD) and eq 14 (SGE) contributing to $\Delta F_{\beta,\alpha_L}$ and $\Delta F_{\alpha_R,\alpha_L}$. To simplify the discussion, we will use the following notation: $f_A = -\beta^{-1} \ln(\int_{D_A} e^{\beta[\Phi(\mathbf{z}_a) - \Phi(\mathbf{z})]} d\mathbf{z})$ and $F_A = -\beta^{-1} \ln(\int_{D_A} e^{-\beta\Phi(\mathbf{z})} d\mathbf{z})$, with $A \equiv \alpha_R$, α_L, β. With these definitions, eq 6 becomes $\Delta F_{AB} = \Delta\Phi_{AB} + f_B - f_A$ and eq 14 becomes $\Delta F_{AB} = F_B - F_A$. The contributions f_β, f_{α_R}, and f_{α_L} to PLD free energy differences and F_β, F_{α_R}, and F_{α_L} to SGE free energy differences are reported in Figure 9. All f_x quantities have a modest dependence on time, their spread ranging around 0.1−0.2 kJ mol^{-1}. The SGE outcomes show a different pattern. In fact, while the spreads of F_β and F_{α_R} are comparable to those obtained with the PLD scheme, F_{α_L} exhibits an evident increasing trend from ∼8.3 to ∼9 kJ mol^{-1}, which is at the origin of poor convergence of $\Delta F_{\beta,\alpha_L}$ and $\Delta F_{\alpha_R,\alpha_L}$ in Figure 8.

To understand the reason for the low accuracy in evaluating F_{α_L} by using SGE schemes, we compare in Figure 10 estimates of $\Phi(\mathbf{z})$, limited to D_β, D_{α_R}, and D_{α_L} domains, obtained through PLD and SGE methods. PMFs in Figure 10A have been determined from eq-R_x simulations lasting 156 ns, while the PMF in Figure 10B is obtained by the SGE simulation of 2016 ns. A simple visual inspection of Figure 10 allows us to notice the noisy sampling of the D_{α_L} domain obtained with the SGE simulation in comparison to the eq-R_{α_L} simulation. Consistent with the data of Figure 9, which show comparable deviations with time of f_{β/α_R} and F_{β/α_R}, less remarkable sampling differences

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

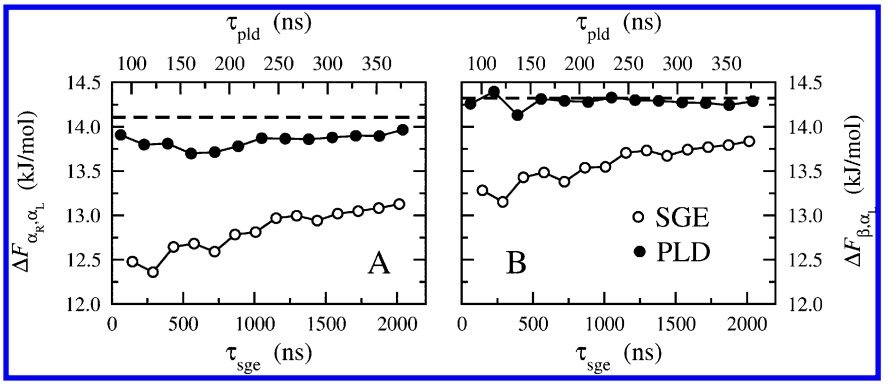


Figure 8. (A) Free energy difference $\Delta F_{\alpha_R,\alpha_L}$ as a function of the simulation time estimated from both PLD and SGE simulations (full and open circles, respectively). τ_{sge}, reported in the bottom axis, indicates the total SGE simulation time obtained by summing all replica times. τ_{pld}, reported in the top axis, is the total PLD time including contributions from eq-R_{α_R}, eq-R_{α_L}, and LP simulations. The dashed line marks the value obtained from the US simulation of ref 19. (B) Free energy difference $\Delta F_{\beta,\alpha_L}$ as a function of the simulation time. Lines are guides for the eyes.

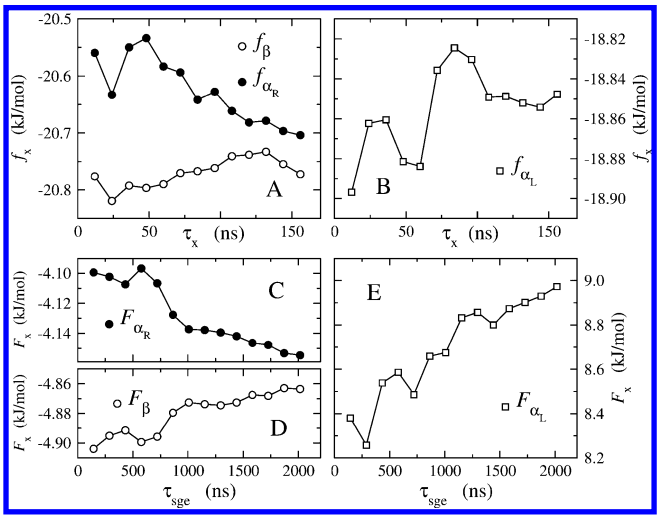


Figure 9. (A and B) Free energy contributions f_β, f_{α_R}, and f_{α_L} (see legend) to $\Delta F_{\alpha_R,\alpha_L}$ and $\Delta F_{\beta,\alpha_L}$, computed via PLD scheme (see eq 6), as functions of the eq-R_x simulation time τ_x. (C, D, and E) Free energy contributions F_β, F_{α_R}, and F_{α_L} (see legend) to $\Delta F_{\alpha_R,\alpha_L}$ and $\Delta F_{\beta,\alpha_L}$, computed via SGE scheme (see eq 14), as functions of the SGE simulation time τ_{sge}. Lines are guides for the eyes.

between PLD and SGE are observed in Figure 10 for the domains D_{α_R} and D_β. These results point to identify the cause of SGE inaccuracy in the low statistical weights of microstates featuring the D_{α_L} domain, ultimately due to the large free energy difference between the D_{α_L} free energy basin and the D_{α_R} and D_β basins (~15 kJ mol^{-1}; see Figure 10B). In fact, owing to this free energy difference, the D_{α_L} basin can be populated significantly during a SGE simulation only when replicas visit ensembles with downscaled intramolecular potential energy, whose microstates are featured by low weight factors. Borrowing the terminology from simulated tempering[37,52] or temperature replica exchange[45,47] methods, these downscaled energy ensembles correspond somehow to high temperature thermodynamic states.

A detailed analysis of the statistical error in reweighting-based simulations was reported by Shen and Hamelberg in ref 53. Inspired by those authors, we define the number of effective samples of a generic ensemble of statistically weighted elements as

$$N_e = \sum_{i=1}^{N_s} \frac{w_i}{\max(w)} \tag{15}$$

where w_i is the weight associated with the ith element, $\max(w)$ is the maximum weight in the series, and N_s is the number of elements. In our calculations, the elements of the ensemble correspond to the system microstates sampled in the simulation. More simply, N_e provides an estimate of the number of elements with significant weight, the value of which ranges from the limit of 1, if only one element is strongly dominant ($\max(w) \simeq 1$), to N_s, if weights are all equal ($\max(w) = N_s^{-1}$). Clearly, for a normalized set of weights, i.e., $\sum_{i=1}^{N_s} w_i = 1$, we can rewrite eq 15 as $N_e = 1/\max(w)$. Expressing N_e as in eq 15 can however result useful for evaluating the number of effective samples of a specific subset of the ensemble. To this purpose, we may limit the sum of eq 15 to the weights of that subset, while keeping $\max(w)$ as the maximum weight of the subset itself. In our context, we split the conformational space into many subsets, one for each bin in the $\mathbf{z}$ space. The bin around $\mathbf{z} \equiv (\phi, \psi)$ is defined as the volume element $\Delta\phi\ \Delta\psi$ centered into $\mathbf{z}$ with $\Delta\phi = \Delta\psi = 3.6°$. Thus, we introduce the number of effective samples as a function of $\mathbf{z}$

$$n_e(\mathbf{z}) = \sum_{i=1}^{N_s} \frac{\delta_{bin}(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x}_i))w(\mathbf{x}_i)}{\max_{\mathbf{z}}(w)} \tag{16}$$

where $w(\mathbf{x}_i)$ is the weight of the generic microstate $\mathbf{x}_i$, $\boldsymbol{\zeta}(\mathbf{x}_i)$ is the vector (ϕ,ψ) of dihedral angles, N_s is the number of microstates sampled during the simulation, $\max_{\mathbf{z}}(w)$ is the maximum weight in the subset of microstates which satisfy the condition $\boldsymbol{\zeta}(\mathbf{x}) = \mathbf{z}$, and $\delta_{bin}(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x}_i))$ holds 1 if $\boldsymbol{\zeta}(\mathbf{x}_i)$ is inside the $\mathbf{z}$ bin and 0 otherwise. In the SGE simulation, weight factors are determined by applying the multistate Bennett acceptance ratio method.[44] The resulting $n_e(\mathbf{z})$ distribution as a function of $\mathbf{z}$ is reported in Figure 11A for the D_{α_L}, D_{α_R}, and D_β domains. The number of effective samples is very large in D_{α_R} and D_β domains, while it is 2 orders of magnitude smaller in the D_{α_L} domain. This accounts for the noisy behavior of $\Phi(\mathbf{z})$ in the D_{α_L} domain, estimated from the SGE simulation (Figure 10B).

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

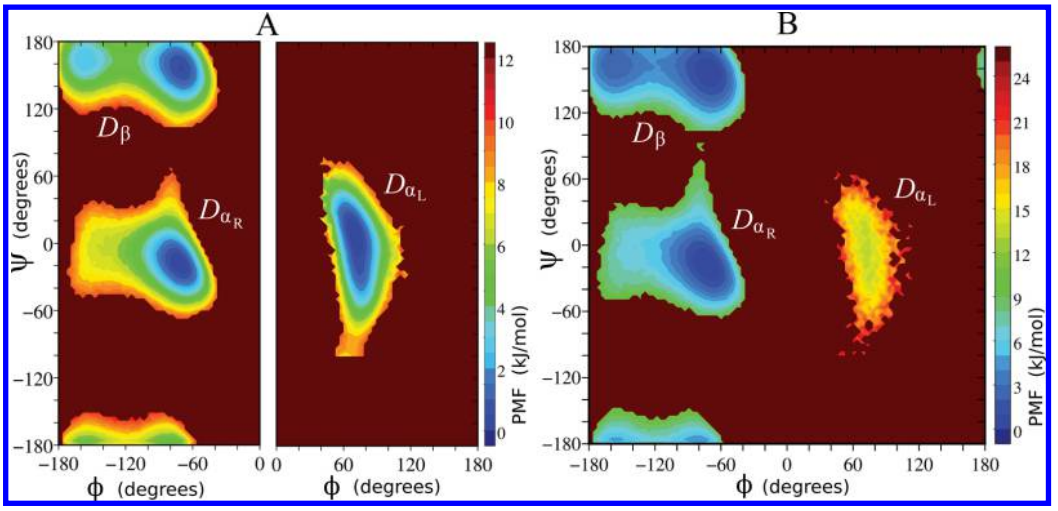


Figure 10. (A) $\Phi(\mathbf{z})$ obtained via PLD scheme with eq-R_x simulations lasting 156 ns. The plot, limited to D_β, D_{α_R}, and D_{α_L} conformational domains, is taken from Figure 6. Different panels are used for data from eq-R_{β/α_R} and eq-R_{α_L} simulations because PMFs are shifted up to an arbitrary constant. (B) $\Phi(\mathbf{z})$ obtained from the SGE simulation lasting 2016 ns. The same angular resolution has been used in all plots ($\Delta\phi = \Delta\psi = 3.6°$).

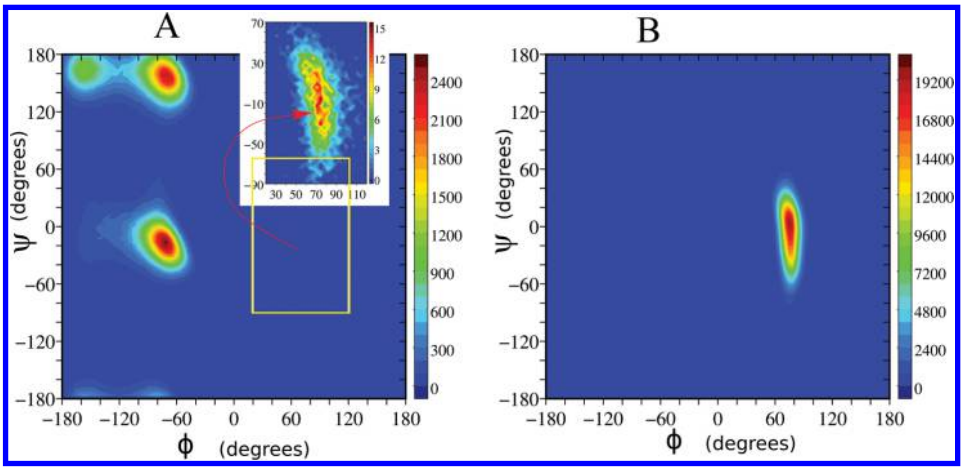


Figure 11. (A) Number of effective samples $n_e(\mathbf{z})$ as a function of $\mathbf{z} \equiv (\phi, \psi)$ computed from the SGE simulation (eq 16). Data are limited to the D_β, D_{α_R}, and D_{α_L} domains. The inset reports on a scaled view of the yellow marked rectangle around the D_{α_L} domain. (B) Number of effective samples $n_e(\mathbf{z})$ as a function of $\mathbf{z}$ computed from the eq-R_{α_L} US simulation (eq 18). Data are limited to the D_{α_L} domain.

Concerning the PLD procedure, reweighting is applied in the eq-R_{α_L} US simulation. According to eq 11, weight factors are expressed as

$$w(\mathbf{x}_i) = \frac{e^{\beta U_{us}(\mathbf{x}_i)}}{\sum_{j=1}^{N_s} e^{\beta U_{us}(\mathbf{x}_j)}} \tag{17}$$

Substituting eq 17 into eq 16 and considering that $U_{us}(\mathbf{x})$ depends only on $\boldsymbol{\zeta}(\mathbf{x})$ through the dihedral angle ϕ (see the expression of $U_{us}(\mathbf{x})$ below eq 10), we can write the number of effective samples for the eq-R_{α_L} simulation as

$$n_e(\mathbf{z}) = \sum_{i=1}^{N_s} \delta(\mathbf{z} - \boldsymbol{\zeta}(\mathbf{x}_i)) \tag{18}$$

This implies that the number of effective samples in the US simulation depends only on our ability to limit sampling around the domain of interest, namely, on the number of times a given conformation is visited during the simulation. The number $n_e(\mathbf{z})$ limited to the D_{α_L} domain obtained from the eq-R_{α_L} simulation is reported in Figure 11B. At variance with the SGE simulation, the most significant result is that values as large as 10^3−10^4 are obtained in a wide region around the domain of interest. This means that in the US simulation microstates contribute quite homogeneously to the $\mathbf{z}$ bins within D_{α_L}, whereas only few microstates give a detectable contribution in SGE simulation. In this observation one can ultimately recognize the basic differences in accuracy between SGE and PLD. We may thus conclude that limiting sampling around specific domains of interest, rather than to extend it to the whole free energy surface, may allow for a significant statistical enhancement.

5. CONCLUSIVE REMARKS

A simulation protocol, called Path-Linked Domains (PLD) scheme, is proposed to estimate free energy differences between configurational states, defined in terms of the hypersurface of (arbitrary) collective coordinates chosen to describe the molecular system. Although the methodology is illustrated by the analysis of conformational states of a small

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

peptide, nothing prevents from applying it in wider contexts, including chemical and biochemical problems involving complexation processes and drug-receptor interactions.[3,14,15]

The basic purpose of the PLD simulation scheme is to tackle the difficulty of conventional equilibrium molecular dynamics and Monte Carlo simulations in exploring free energy hypersurfaces featured by manifold barriers and minima. With respect to other simulation approaches, PLD allows us to limit sampling to defined subsets in the space of the collective coordinates chosen to describe the configurational states of interest (in our calculations, the ϕ and ψ dihedral angles of the alanine dipeptide). This restrained sampling is realized by means of two independent simulations that allow us to compute local configuration integrals associated with the two states. These integrals correspond to a sort of vibrational contributions to the free energy. Even if their difference cannot be directly related to the free energy difference between the two states, it is possible to "make a link" in the space of collective coordinates, determining the difference of potential of mean force between two arbitrary points within the domains featuring the configurational states. The linking path in the space of collective coordinates can be chosen arbitrarily and computed with any method available in the literature, starting from adaptive biasing potential/force methods[16] to nonequilibrium techniques, such as those employed in this study.

The main advantage of the method when computing free energy differences of established configurational states is therefore to prevent an accurate calculation of the *whole* free energy hypersurface, limiting sampling to bare essentials. On the other side, a drawback that we envisage in the PLD scheme, with respect to methods based on full sampling of the free energy hypersurface, is the prior knowledge of the two target configurational states. Sometimes, it can be difficult to gain such a knowledge from simple intuition. In these situations, one may however resort to short equilibrium simulations or to some accelerated sampling technique[11] to roughly probe the free energy landscape in the space of collective coordinates.

AUTHOR INFORMATION

Corresponding Author

*E-mail: riccardo.chelli@unifi.it.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by European Union Contract RII3-CT-2003-506350 and by the Italian Ministero dell'Istruzione, dell'Università e della Ricerca.

REFERENCES

(1) Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6285−6830.
(2) Nicolini, P.; Frezzato, D.; Gellini, C.; Bizzarri, M.; Chelli, R. *J. Comput. Chem.* **2013**, *34*, 1561−1576.
(3) Luo, H.; Sharp, K. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 10399−10404.
(4) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.
(5) Senderowitz, H.; Guarnieri, F.; Still, W. C. *J. Am. Chem. Soc.* **1995**, *117*, 8211−8219.
(6) Head, M. S.; Given, J. A.; Gilson, M. K. *J. Phys. Chem. A* **1997**, *101*, 1609−1618.
(7) Nilmeier, J. P.; Crooks, G. E.; Minh, D. D. L.; Chodera, J. D. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, E1009−E1018.
(8) Giovannelli, E.; Gellini, C.; Pietraperzia, G.; Cardini, G.; Chelli, R. *J. Chem. Theory Comput.* **2014**, *10*, 4273−4283.
(9) Park, S. *Phys. Rev. E* **2008**, *77*, 016709.
(10) Torrie, G.; Valleau, J. *J. Comput. Phys.* **1977**, *23*, 187−199.
(11) Frenkel, D.; Smit, B. *Understanding Molecular Simulations: From Algorithms to Applications*; Academic Press: San Diego, CA, 2002.
(12) Each component of $\boldsymbol{\zeta}(\mathbf{x})$ depends on the Cartesian coordinates of four atoms. However, to simplify the notation, here we take the general dependence on the coordinates $\mathbf{x}$ of all the atoms.
(13) Ytreberg, F. M. *J. Chem. Phys.* **2009**, *130*, 164906.
(14) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.
(15) Kolossvary, I. *J. Phys. Chem. A* **1997**, *101*, 9900−9905.
(16) Chipot, C., Pohorille, A., Eds.; *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer: Berlin, 2007; Vol. *86*.
(17) Park, S.; Schulten, K. *J. Chem. Phys.* **2004**, *120*, 5946−5961.
(18) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(19) Marsili, S.; Barducci, A.; Chelli, R.; Procacci, P.; Schettino, V. *J. Phys. Chem. B* **2006**, *110*, 14011−14013.
(20) Chelli, R.; Signorini, G. F. *J. Chem. Theory Comput.* **2012**, *8*, 830−842.
(21) Smart, J. L.; Marrone, T. J.; McCammon, J. A. *J. Comput. Chem.* **1997**, *18*, 1750−1759.
(22) Brooks, C. L.; Case, D. A. *Chem. Rev.* **1993**, *93*, 2487−2502.
(23) Feig, M. J. *J. Chem. Theory Comput.* **2008**, *4*, 1555−1564.
(24) Vymětal, J.; Vondrášek, J. *J. Phys. Chem. B* **2010**, *114*, 5632−5642.
(25) Tazaki, K.; Shimizu, K. *J. Phys. Chem. B* **1998**, *102*, 6419−6424.
(26) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.
(27) Chelli, R. *J. Chem. Theory Comput.* **2010**, *6*, 1935−1950.
(28) Chelli, R.; Signorini, G. F. *J. Chem. Theory Comput.* **2012**, *8*, 2552.
(29) Marsili, S.; Signorini, G. F.; Chelli, R.; Marchi, M.; Procacci, P. *J. Comput. Chem.* **2010**, *31*, 1106.
(30) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695.
(31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(32) Essmann, U.; Perera, M. L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, G. L. *J. Chem. Phys.* **1995**, *103*, 8577.
(33) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990−2001.
(34) From now on, the symbol β/α_R will be used to refer to β or α_R conformations indifferently.
(35) Minh, D. D. L.; Adib, A. B. *Phys. Rev. Lett.* **2008**, *100*, 180602.
(36) Chelli, R.; Procacci, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1152−1158.
(37) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsov-Velyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776.
(38) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1997**, *18*, 920.
(39) Mitsutake, A.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *332*, 131.
(40) Huang, X.; Bowman, G. R.; Pande, V. S. *J. Chem. Phys.* **2008**, *128*, 205106.
(41) Park, S.; Pande, V. S. *Phys. Rev. E* **2007**, *76*, 016703.
(42) Bennett, C. H. *J. Comput. Phys.* **1976**, *22*, 245.
(43) Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S. *Phys. Rev. Lett.* **2003**, *91*, 140601.
(44) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *129*, 124105.
(45) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140.
(46) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
(47) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.
(48) Tesi, M. C.; Janse van Rensburg, E. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82*, 155.
(49) Spill, Y. G.; Bouvier, G.; Nilges, M. *J. Comput. Chem.* **2013**, *34*, 132−140.

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571

(50) Signorini, G. F.; Giovannelli, E.; Spill, Y. G.; Nilges, M.; Chelli, R. *J. Chem. Theory Comput.* **2014**, *10*, 953−958.
(51) The total SGE simulation time τ_{sge} is the sum of the times related to all replicas.
(52) Marinari, E.; Parisi, G. *Europhys. Lett.* **1992**, *19*, 451.
(53) Shen, T.; Hamelberg, D. *J. Chem. Phys.* **2008**, *129*, 034103.

DOI: 10.1021/acs.jctc.5b00248
J. Chem. Theory Comput. 2015, 11, 3561−3571