Methods 52 (2010) 115–121

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Computing free energies of protein conformations from explicit solvent simulations

Pavel I. Zhuravlev [1], Sangwook Wu [1], Davit A. Potoyan, Michael Rubinstein, Garegin A. Papoian *

Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3290, United States

ARTICLE INFO

Article history:
Accepted 5 May 2010
Available online 20 May 2010

Keywords:
Protein conformational free energy
Path coordinate
Explicit solvent simulations
Native state ensemble

ABSTRACT

We report a fully general technique addressing a long standing challenge of calculating conformational free energy differences between various states of a polymer chain from simulations using explicit solvent force fields. The main feature of our method is a special mapping variable, a path coordinate, which continuously connects two conformations. The path variable has been designed to preserve locality in the phase space near the path endpoints. We avoid the problem of sampling the unfolded states by creating an artificial confinement "tube" in the phase space that prevents the molecule from unfolding without affecting the calculation of the desired free energy difference. We applied our technique to compute the free energy difference between two native-like conformations of the small protein Trp-cage using the CHARMM force field with explicit solvent. We verified this result by comparing it with an independent, significantly more expensive calculation. Overall, the present study suggests that the new method of computing free energy differences between polymer chain conformations is accurate and highly computationally efficient.

1. Introduction

Within the energy landscape paradigm, the protein native state is naturally viewed as a multitude of nested conformational basins, that are dynamically explored during protein function [1–7]. This functional landscape represents only a small fraction of the larger folding landscape – which includes denatured conformations [2] (see Fig. 1). On the scale of the whole folding landscape, it is possible to describe folding dynamics through the statistical properties of the landscape. However, in the case of protein functional motions and native dynamics, the specific details of the functional landscape play an important role, necessitating detailed characterization of the landscape at a relatively high energetic resolution, corresponding to the structural resolution of ~ 1 Å. For example, such topographical maps [2] may be needed, to investigate transitions in allosteric proteins, which undergo global conformational rearrangements upon local perturbation such as ligand binding. In some cases, allosteric switching is thought to modulate enzymatic rates [8]. Thus, elucidation of functional landscapes may help to understand how targeted point mutations influence catalytic activities [9] and may shed light on large scale phenomena, such as molecular motor functioning [10].

The energy landscape is a function of a large number of conformational and solvent degrees of freedom. In practical applications, the landscape is projected into one or several collective degrees of freedom, to allow physically meaningful interpretation of the chain dynamics. The present work provides the solution for a simpler problem: how to calculate the free energy difference between two specific conformations, *A* and *B*, of a polymer chain in a simulation with explicit solvent? Solving this problem is a step towards building a reduced representation of energy landscape and would not only help shed light on the biological processes mentioned above, but it would also aid in the development of atomistic and coarse-grained force fields, by allowing researchers to compare the free energy differences among the same conformations computed with different force fields and representations.

The attempts for addressing the challenge of calculating conformational free energies of molecules and macromolecules have a long history. A popular molecular mechanics/Poisson–Boltzmann/surface area (MM/PBSA) technique is based on generating a representative set of conformations with explicit solvent and then removing solvent and estimating free energy as a sum of several terms [11]. This technique is based on several uncontrolled approximations that may potentially limit its applicability [12], such as reliance on continuum electrostatics calculations to estimate part of polymer's solvation free energy, where these types of estimates can sometimes be quantitatively inaccurate [13]. A similar method, ES/IS, avoids using the Poisson–Boltzmann equation, and instead collects statistical averages computed from explicit simulations [14]. However, some of the terms in the free energy ansatz are still estimated by employing implicit continuum models of the solvent [14]. Some newer techniques like the deactivated morphing method [15], which is based on using a series of unphysical intermediates states between conformations *A* and *B*, use fully explicit solvent. The deactivated morphing method has been used to calculate the free energy difference between folded and

* Corresponding author.
E-mail address: gpapoian@unc.edu (G.A. Papoian).
[1] These authors contributed equally to this work.

1046-2023/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2010.05.003

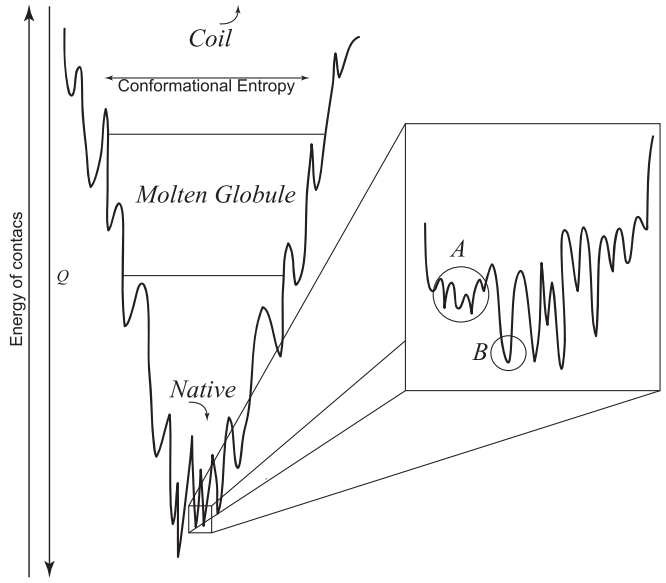

Fig. 1. The protein folding energy landscape is schematically shown in the shape of a funnel [19]. The bottom of the funnel, containing the native, functional landscape, is zoomed in on the right. The new technique presented in this work allows to choose two conformations, *A* and *B*, from the native ensemble and calculate the free energy difference between them.

misfolded states of human Pin1 WW domain [16]. However, high (thousands of k_BT) free energy differences separating the unphysical states require extremely thorough sampling, pointing to a potentially very fast growth of computational cost with the system size. Structurally based umbrella sampling techniques have also been employed [17,18], however, certain technical problems elaborated below significantly limit their domain of applicability.

To the best of our knowledge, the technique presented in this paper is devoid of any of the drawbacks mentioned above. It is fully general, takes the solvent into account explicitly, does not rely on any uncontrolled approximation (other than those intrinsic to any particular force field) and allows tuning of the conformational resolution. The method is accurate and computationally efficient. In certain cases, the full free energy profile for a transition may be obtained rather than just the free energy difference between two conformations.

Let point *A* in the phase space of a polymer chain be defined by precise coordinates of all the atoms of the chain. Point *A* has a finite entropy, and therefore statistical weight, due to solvent degrees of freedom. Point *A* only represents a point in polymer's conformational phase space, but it is expanded to a small region in the full phase space of the system. Furthermore, physically meaningful questions most often imply that conformation *A* includes not only the point *A* in conformational phase space, but also some finite size locale around point *A*. The latter is often called a conformational basin. The basin size depends on the question of interest and relevant physical considerations: for example, it could be defined by extent of atomic vibrations or by the experimental error in determining the structure of *A*. It may also be defined by the features of the particular local minimum on the energy landscape, such as its width and depth compared with thermal energy. The method that we report here does not provide explicit constraints on structural similarity within a basin, as these will vary between studies. Instead, it provides means to calculate the free energy difference between the *A* and *B* basins, once those are defined based on other physical considerations. Thus, by "conformation A" we mean some well-defined neighborhood of point *A*, its conformational basin, and by "free energy of *A*" we mean the logarithm of statistical weight of this basin.

Many techniques for calculating free energy differences (such as umbrella sampling) require the free energy of the system to be computed as a function, $F(\xi)$, of a dynamical variable ξ, where common examples of ξ include density, magnetization, and radius of gyration. It is defined by a set of phase space variables and reflects the state of the system at any moment in time. Umbrella sampling is a way to sample low-populated regions of the free energy profile $F(\xi)$ by restricting trajectories to the narrow regions of the profile with parabolic potential $U = k(\xi - \xi_0)^2$. These regions are called umbrella windows, and the name comes from parabolic shape of the potential [20]. Coming up with an appropriate scalar variable ξ (that we will refer to as a *path coordinate*) for the problem of a transition between two polymer conformations is a non-trivial task. In this paper we present a method that solves this problem. The path coordinate has to keep most of the relevant information contained in the multitude of conformational and solvent degrees of freedom and simultaneously discriminate between *A* and *B*. In addition, both conformations *A* and *B* must correspond to finite segments of the ξ space, which means that conformations similar to *A* must have ξ close to $\xi(A)$ and vice versa, a region of $\xi \approx \xi(A)$ must only contain conformations similar to *A* (the statement has to be true for conformation *B* as well). In some cases, the topology of the landscape or allosteric motions themselves can provide a good, physically meaningful collective coordinate [21,22]. The new path coordinate that we propose works for general case. It is local in the conformational phase space near points *A* and *B* and is highly resolved in discriminating them. We tested the method on two conformations from the native ensemble of a 20-residue protein Trp-cage [23] at temperature 282 K. The resulting free energy difference between them was found to be 0.43 kcal/mol(0.77 k_BT). We also calculated the same free energy difference with an independent, more computationally expensive technique resulting in 0.45 kcal/mol(0.81 k_BT) and confirming the accuracy of the method within 5%.

Any path coordinate is destined to have a range of values that contains all the unfolded states which are equally unrelated to either *A* or *B*. The phase space volume corresponding to this region is huge. Nevertheless, in general it must be sampled to obtain a free energy profile between *A* and *B*, unless energy landscape of the molecule has some intricate self-averaging property. Under specific circumstances, this may be the case in proteins, for example, when *A* and *B* correspond to an actual allosteric transition, but in general the sampling of this region might be problematic, particularly considering the computational cost of all-atom explicit solvent simulation techniques on currently available computational resources. We have conceived a solution of this problem by creating an artificial confinement in the phase space that prevents the molecule from unfolding without affecting the calculation of the desired free energy difference. Along with introducing the new path coordinate, this constitutes a new technique which is the main result of the current work.

Summarizing, in this article, we report a fully general and computationally efficient technique for finding conformational free energy differences between various states of a protein chain from all-atom explicit solvent molecular dynamics simulations. We applied our technique to compute the free energy difference between two conformations of Trp-cage (NLYIQWLKDGGPSSGRPPPS) native ensemble, using the CHARMM force field with explicit solvent. We compared the results with those derived from an alternative, independent method, which is computationally much more expensive, revealing the remarkable efficiency and accuracy of the proposed technique.

2. The path coordinate and confining of the trajectories

To chart the native state in high resolution it is necessary to use some distance measure, $s(X, Y)$, between the points of

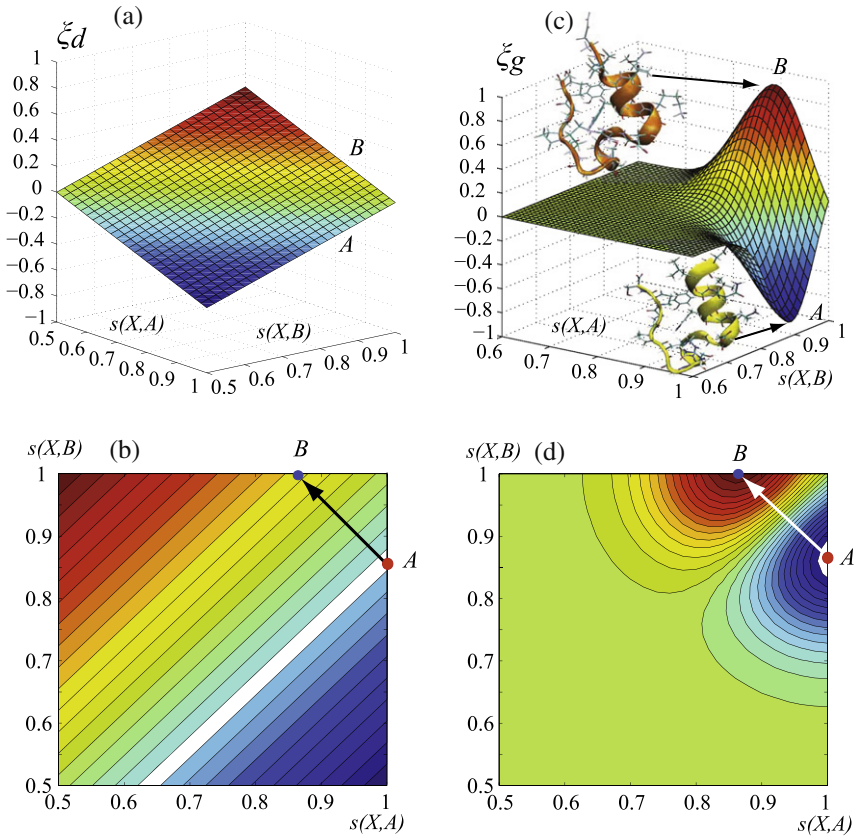

Fig. 2. The path coordinate $\xi_d = s(X,A) - s(X,B)$ is shown in 3D (a) and as a contour plot (b). The path coordinate ξ_g (Eq. (2)) is shown in 3D (c) and as a contour plot (d). The labeled points correspond to conformations A and B – two of the states detected by an NMR study of Trp-cage native state [23] A one-dimensional dynamical variable necessarily partitions the phase space into multidimensional iso-surfaces. In both cases on this figure iso-surfaces around $\xi_d \approx 0$ (a green stripe) and $\xi_g \approx 0$ (the large green area) contain all the unfolded states. This is the case with any one-dimensional path coordinate. Our method solves this problem with the confinement potential "tube".

conformational phase space, to quantify similarity between any two conformations X and Y. Examples of such measures include root–mean–square-deviation of corresponding atomic coordinates (RMSD), contact order, fraction of shared contacts (q), and fraction of shared dihedral angles. With such measure it is possible to map the whole conformational state on a single variable, i.e. to define the variable for an arbitrary conformation X. This variable is the similarity $s(X,N)$ between X and a preliminary chosen specific conformation N. This idea is used in protein folding with N being the native state [24]. The $s(X,N)$ is then the coordinate that describes folding. Since in our problem the two states may be very similar, as it happens in the native basin, we need a much higher resolution, and the one taking into account the conformational changes transverse to folding (i.e. transverse to $s(X,N)$) [2]. One way to increase the resolution is to use two variables instead of one, $s(X,A)$ and $s(X,B)$, the similarities to two specific conformations [25], mapping now the conformational phase space onto a 2D-plane (see Fig. 2). Depending on particular choice of $s(X,Y)$ this variable may have different ranges. In many cases it changes from 0 (X and Y are totally different) to 1 (X is the same as Y). Fig. 2 assumes such a case: both $s(X,A)$ and $s(X,B)$ can change from 0 to 1, mapping thus the whole conformational phase space onto a square. Area near the origin corresponds to conformations highly dissimilar to both A and B ($s(X,A) \approx s(X,B) \approx 0$). If A and B are both folded states belonging to the native ensemble (for instance, two allosteric states), the origin will contain all the unfolded states (since these states are dissimilar from the folded states). A and B then will be similar ($s(A,B) \approx 1$) and close to the upper right corner of the square. Then the diagonal of the square ($s(X,A) = s(X,B)$) will correspond to the folding coordinate line, and motions perpendicular to this diagonal will be transverse to folding. Fig. 2b illustrates such a case: A and B are two structures from the native ensemble of a small protein Trp-cage. The structural resolution of the native region (upper right corner) is much higher than that of unfolded region near the origin.

The most obvious path coordinate would be

$$\xi_d(X) = s(X,A) - s(X,B), \tag{1}$$

but it lacks the aforementioned property of locality near A and B which is compulsory. We want $\xi(X) = \xi(A) \pm \delta\xi$ to only include conformations similar to A, so that $s(X,A) = s(A,A) \pm \delta s$ with δs being small [26]. However, the difference based definition of path coordinate ξ_d (Eq. (1)) permits arbitrary large changes to both terms in the difference as long as the difference itself stays the same. In other words, the whole strip that is highlighted in white in Fig. 2b will contribute to free energy of conformation A, including the unrelated unfolded conformations near the origin.

We propose a path coordinate that remains local around the conformations of interest:

$$\xi_g(X) = \exp\left[-\frac{(s(X,A) - s(A,B))^2 + (s(X,B) - 1)^2}{2\sigma_g^2}\right] - \exp\left[-\frac{(s(X,B) - s(A,B))^2 + (s(X,A) - 1)^2}{2\sigma_g^2}\right]. \tag{2}$$

If this coordinate is visualized as elevation above the 2D-plane defined by $s(X,A)$ and $s(X,B)$, it corresponds to a positive gaussian peak of width σ_g centered on conformation B (with coordinates on the 2D plane $s(A,B)$ and $s(B,B) = 1$) and a negative gaussian peak

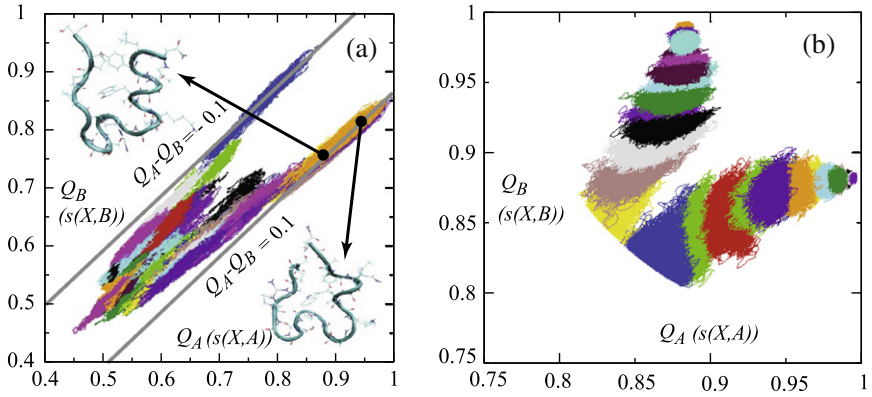

Fig. 3. Mapping of umbrella windows onto the phase space square. A patch of the same color corresponds to a trajectory in a single window. (a) $\xi_d = s(X,A) - s(X,B) = Q_A - Q_B$ (Eq. (1)) performs poorly as a path coordinate. Two very dissimilar structures (shown in the corners) are in the same (orange) window, that also includes conformation *A*. (b) ξ_g (Eq. (2)) performs much better with window patches covering the conformations *A* and *B* very compactly.

centered on conformation *A* (with coordinates on the 2D plane $s(A,A) = 1$ and $s(A,B)$) (Fig. 2c and d). In this coordinate constant elevation strips form local regions of width σ_g near points *A* and *B*. Note that it is not necessary for both gaussians to have the same width σ_g.

We chose a small protein, Trp-cage, to test our method. Trp-cage is one of the smallest known proteins (20 residues) with a set of native structures reported by an earlier NMR study [23]. We chose the two most dissimilar structures in this set as points *A* and *B*. Despite the fact that we had allosteric states in mind while developing the method, these Trp-cage states are not expected to represent deep minima, but are simply used to test our approach. Furthermore, although allosteric states are typically minima with a barrier separating them, our method is more general and can be applied to any two arbitrarily defined conformational states, even if they are not minimum energy structures. Our technique allows computation of the correct ratio of thermal probabilities to find the system in either of these states, or the free energy difference. Free energy differences between conformations which are not deep minima may be used for example to gauge the accuracy of coarse-grained force-fields, by comparing with the corresponding results from atomistic simulations. Likewise, our technique may be used with many different similarity measures, $s(X,Y)$. In this work, we chose the fraction of common contacts $q(X,Y)$ to quantify similarity between structures *X* and *Y*, or more precisely

$$s(X,Y) = q(X,Y) = \frac{1}{N}\sum_{ij} \exp\left[-\frac{(r_{ij}^X - r_{ij}^Y)^2}{2\sigma^2}\right], \tag{3}$$

where r_{ij}^X and r_{ij}^Y indicate the distances between *i*th and *j*th atom in conformations *X* and *Y*, respectively, and normalization factor *N* is equal to the number of atom pairs used to compare structures *X* and *Y*. In the example with Trp-cage we included carbons C_α, C_β, C_γ, C_δ, C_ϵ, and C_z (78 atoms total) in the summation. Gaussian function in Eq. (3) smoothes the boundary between a "contact" and "no contact". Further in the text we use the following notation:

$$Q_A \equiv q(X,A), Q_B \equiv q(X,B), \xi_g \equiv \xi_g(Q_A,Q_B) \equiv \xi_g(X).$$

The comparison of locality between the previous, difference based path coordinate ξ_d and the newly proposed gaussian based path coordinate ξ_g is shown in Fig. 3. Patches of different colors correspond to different windows of umbrella sampling (that keep the path coordinate localized). It can be seen from Fig. 3a that the previously reported coordinate ξ_d indeed forces the trajectories to sample a stripe-like region of the 2D (Q_A, Q_B) plane. The ξ_d windows that contain conformations *A* and *B* also group with them unrelated, partially unfolded structures. In comparison, when using the new path coordinate ξ_g, the window that contains *A* is local, as shown in Fig. 3b, and the trajectory in this windows does not stray far from *A*, keeping the conformations unrelated to *A* from this window.

The new path coordinate ξ_g constitutes the essence of the technique reported here. However, there is another major feature that might be needed for efficient calculations under specific circumstances. Note that the conformational phase volume as a function of ξ_g is not constant: it is much larger in the region $\xi_g \approx 0$, which contains all the unfolded states and decreases rapidly towards the endpoints $\xi_g = 1$ and $\xi_g = -1$. In principle, the $\xi_g \approx 0$ umbrella windows have to be thoroughly sampled as well which may represent a problem with large proteins and explicit solvent force fields. However, if one is interested only in free energy difference between conformations *A* and *B* and not in the full free energy profile between them, this problem can be side-stepped. It is possible to confine the sampling trajectories inside an artificial "tube" that envelops a presumptive path between the states. This can be done by adding an appropriate confinement potential V_c to the Hamiltonian, $H' = H + V_c$. V_c should be chosen in such a way that the conformational basins of A and B are not affected ($V_c(A) \approx V_c(B) \approx 0$). Using the modified Hamiltonian, the free energy difference between the conformations *A* and *B* is

$$F'_A - F'_B = -\frac{1}{\beta}\ln\frac{\int_{\Gamma_A} e^{-\beta H'} d\Gamma}{\int_{\Gamma_B} e^{-\beta H'} d\Gamma} = -\frac{1}{\beta}\ln\frac{\int_{\Gamma_A} e^{-\beta (H+V_c)} d\Gamma}{\int_{\Gamma_B} e^{-\beta (H+V_c)} d\Gamma}, \tag{4}$$

where $\beta = 1/kT$, Γ represents the whole conformational space and Γ_A and Γ_B indicate the phase volumes of conformations *A* and *B*. If we set $V_c = 0$ (below desirable marginal error, e.g. less than 0.01 kcal/mol) everywhere in Γ_A and Γ_B, then $F_A - F_B = F'_A - F'_B$, thus, V_c will not affect the free energy difference we are calculating. In our test example with Trp-cage we chose a V_c that won't allow the trajectories to unfold. On the (Q_A, Q_B) square this would mean preventing the trajectories from going towards the origin, keeping them in the upper right corner, corresponding to the native region. Thus V_c can be visualized as a wall of cylindrical shape surrounding the upper right corner of the phase space square (Q_A, Q_B). Keeping the radius of the cylinder small would aid computational efficiency, but it should be large enough to not touch the conformational basins of *A* and *B* and to allow sufficient overlap between umbrella windows (see Section 5 and Fig. 5).

3. Results

For the two conformations of Trp-cage that we chose to test the method $q(A,B) = 0.88$ (Fig. 4). We chose the value of parameter σ_g in Eq. (2) to be 0.23. The free energy profile as a function of ξ_g is

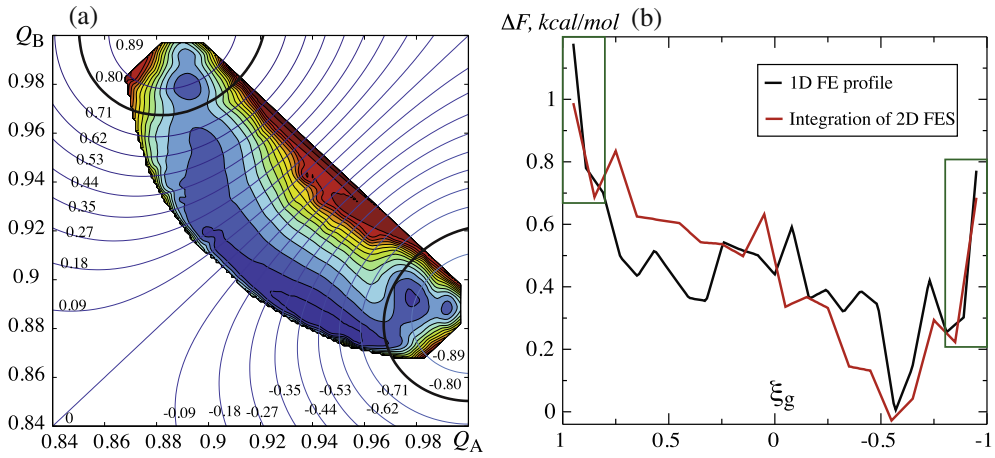

Fig. 4. (a) 2D free energy surface as a function of Q_A and Q_B; (b) 1D free energy profiles obtained by integration of a FES in (a) and by umbrella sampling of ξ_g are very similar showing the consistency of the methods. In all simulations the confinement potential (5) was used. The basins of A and B are marked by thicker contour lines in (a) and by rectangles in (b).

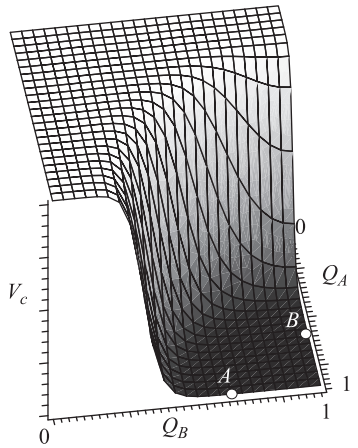

Fig. 5. The confinement phase space "tube" in our case is a wall of cylindrical shape, surrounding the native region of the phase space (corresponding to the upper right corner of the $((Q_A, Q_B)$ square).

shown as a black solid curve in Fig. 4b. This 1D profile was calculated using 108 umbrella windows that are each 1.2 ns long. High conformational entropy in the region $\xi_g \approx 0$ is the thermodynamic factor which tends to lower that region's free energy. However, we are mainly interested in obtaining the free energy difference between the conformations A and B (marked by rectangles). As discussed in introduction, these correspond to finite segments of the path coordinate ξ_g line. The size of these segments is not set by our method and must be chosen based on other considerations, such as magnitude of atomic vibrations. If the structures for A and B are obtained from experiment, the size may be defined by experimental precision. For a given segment size, though, our procedure provides a definite answer. It makes physical sense to choose a segment size that corresponds to a conformational basin, the local minimum on the free energy landscape, providing there is one. We chose the sizes to be $\Delta\xi_g = 0.2$. To calculate the free energy of a segment, we sum the partition functions of all the states within it: $Z_A = \int_{0.8}^{1} \exp(-\beta F(\xi_g))d\xi_g$ and $Z_B = \int_{-1}^{-0.8} \exp(-\beta F(\xi_g))d\xi_g$. The free energy difference $F_A - F_B = -kT\ln(Z_A/Z_B)$, turns out to be 0.43 kcal/mol (0.77 k_BT).

To independently verify this result we also constructed a 2D free energy surface as a function of Q_A and Q_B (Fig. 4a). 2D FES calculations are much more expensive computationally [27–29]. For these calculations, 922 umbrella windows are needed, 1.2 ns each resulting in over 1 μs of total simulation time (compared to ~ 100 ns simulation time for the 1D profile). To compare free energies from 1D free energy profile $F(\xi_g)$ and from 2D free energy surface $F(Q_A, Q_B)$ we integrated the 2D FES numerically using

$$e^{-\beta F(\xi'_g)} = \int e^{-\beta F(Q_A,Q_B)}\delta(\xi'_g - \xi_g(Q_A, Q_B))dQ_A dQ_B.$$

The profile obtained from the 2D surface $F(Q_A, Q_B)$ yields 0.45 kcal/mol (0.81 k_BT) free energy difference between the basins, to be compared with 0.43 kcal/mol (0.77 k_BT) obtained from 1D $F(\xi_g)$ calculations. The difference is within 5%, indicating that the method is highly accurate.

The best way to estimate the error bars for the graph in Fig. 4b would be to perform several independent simulation runs. Since that is computationally expensive in explicit solvent force field, instead we carried out five independent runs of the same system in simple implicit solvent, modeled by dielectric medium with $\epsilon = 80$. The standard error for the mean estimated from these five runs turned out to be less than 0.05 kcal/mol.

4. Discussion

The new method presented for computing free energy differences between polymer chain conformations has several advantages compared to previous approaches. The technique is general and does not involve calculations with unphysical states of the molecule (other than forcing the system to visit states that are poorly accessible thermally). It has adjustable structural resolution, that can be changed depending on the nature of the two conformations of interest. The resolution can be changed by increasing or decreasing the set of atoms whose positions enter into the definition of $s(X, Y)$. For instance, partially unfolded states of proteins would require coarser treatment, than the one in the example here: side-chain rearrangements should be considered as not changing the conformation, because they would occur on the same timescale as solvent motions.

The enveloping "tube" for the pathway between A and B is not as artificial construct as it may seem at first glance, at least regarding proteins. In real allosteric transitions proteins do not unfold (or do so only partially), which means a natural tendency to stay in the region of the energy landscape we are trying to sample. Thus, many proteins would naturally sample a well-defined path if the umbrella window sampling times are not too long, which would allow escape over kinetic barriers surrounding the dominant path. In this case, an externally introduced confining tube would serve only as a

"guard rail" for the trajectory rather than a wall that cuts a part of phase space off. In the example presented in the paper we used the following considerations to find a functional form for V_c that would approach such ideal case. First, the tube has to allow sufficiently many pathways connecting A and B, so that there is enough overlap between the umbrella windows. Then, the rate of transition between neighboring windows (with V_c on) multiplied by the simulation time inside a window must be larger than unity. However, this should not be an excessively large number, since it is a gauge for the sampling problem that the tube is meant to solve. Thus, some optimal width tube needs to be devised around the steepest descent path between the end points. Since this path is not known *a priori*, we probed the phase space with short time umbrella windows, thus "seeing" where the trajectory "prefers" to go (Fig. 3b), until we observe a continuous path between A and B. The areas where the system spends most of its time are grouped around the steepest descent path, forming the shape of the tube (see Section 5). A more general procedure for constructing an adjustable confinement tube and making sure that its main purpose is to be a "guard rail" would allow the calculation of realistic transition pathways between the states, not just the free energy difference, with high computational efficiency. Thus, if the system naturally wants to go from A to B, the tube acts as a "guard rail", and we can recover the real transition pathway with a free energy profile along the pathway. In the opposite case, when system does not want to go from A to B, we may need to force it with the confinement tube, and then we recover just the free energy difference.

Root–mean–square-deviation of atom positions (RMSD) is widely used as a similarity measure between two conformations, s. Our method is formulated in terms of general $s(X,Y)$, so RMSD could also be used. The plot corresponding to Fig. 2 would look somewhat differently (the whole quadrant instead of square with the native region near the origin) but with the same main features. We preferred to use fraction of native contacts $q(X,Y)$ for the following reasons. q is a two-point parameter, comparing the distances between pairs of atoms, while RMSD compares the coordinates of each atom individually (after proper alignment). Use of q does not require this preliminary alignment of the structures. As the contact energy plays crucial role in proteins, q is more correlated with protein physics than RMSD. For example, if one imagines two conformations having two α-helices, that are close in one of them and apart in the other one, then RMSD between them will be very large, suggesting no structural similarity, while q will still show the similarities of individual helices.

In fact, umbrella sampling with the path coordinate ξ_d (Eq. (1)) and RMSD in the role of s has been used to calculate free energy difference between A- and B-forms of DNA [17] and free energy mapping of allosteric switching between the open and the closed forms of adenylate kinase upon ligand binding [18]. As discussed above, this technique has the non-locality problem, that we have solved by introducing a new path coordinate in this paper. Along with the confinement idea, our technique is fully general, adjustable to high resolution, computationally efficient and does not operate on unphysical states of the system. The method can be used for an arbitrary pair of states, without the naturally existing transition path between them, like in the current work, to compute free energy differences between two protein conformations. In addition, a straightforward generalization should allow computing the whole free energy profile for the transition path between two states, when this transition occurs in nature.

5. Methods

All atomistic molecular dynamics (MD) simulations were carried out using LAMMPS [30] (large-scale atomic/molecular massively parallel simulator) using CHARMM27 protein–lipid force field with explicit solvent [31]. Trp-cage was placed in a $50 \times 50 \times 50$ Å^3 box with 2275 water molecules (TIP3P model) and the counterions, $5Na^+$ and $6Cl^-$ in order to mimic the physiological conditions. The system was prepared in NAMD using the standard protocol [32]. The system was heated up to 282 K and equilibrated for 800 ps using targeted MD to keep the innate NMR structures. Next, NPT simulations were carried out in LAMMPS for 60 ps with targeted MD to bring the conformation to a specific umbrella window. Finally, for each of the windows, 1.2 ns long NPT simulations were carried out, where the last 1 ns was used for data analysis.

The confinement potential for the "tube" enveloping the trajectory is constructed as follows. In the case of Trp-cage, the upper right corner of the (Q_A, Q_B) square is naturally devoid of states, thus, we only need to confine the trajectory from the side of lower values of Q_A and Q_B. We placed a cylindrical "wall" around upper right corner of conformation space square (Fig. 4a), that confines the trajectories to the upper right corner. The "wall" (actually, a sharp step with finite width) was implemented by a hyperbolic tangent of a distance from the upper right corner of the phase space square $(Q_A = 1, Q_B = 1)$:

$$V_c = \varepsilon\Big(1 + \tanh\Big(k\Big[(Q_A - 1)^2 + (Q_B - 1)^2 - \mu^2\Big]\Big)\Big). \tag{5}$$

The parameters ε, k and μ, that were chosen as 10 kcal/mol, 5×10^3 and 0.135, respectively, satisfy the conditions of good overlap between umbrella windows, but, eliminate a huge number of intermediate states.

Acknowledgments

We would like to acknowledge financial support of the Camille and Henry Dreyfus Foundation, National Science Foundation under Grants CHE-0846701, CHE-0616925, and CHE-715225, National Institutes of Health under Grant 1-R01-HL0775486A.

References

[1] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, Science 254 (5038) (1991) 1598–1603.
[2] P.I. Zhuravlev, G.A. Papoian, Curr. Opin. Struct. Biol. 20 (1) (2010) 16–20.
[3] K.A. Henzler-Wildman, M. Lei, V. Thai, S.J. Kerns, M. Karplus, D. Kern, Nature 450 (7171) (2007) 913–916.
[4] P.W. Fenimore, H. Frauenfelder, B.H. McMahon, R.D. Young, Proc Natl Acad Sci USA 101 (40) (2004) 14408–14413.
[5] H. Frauenfelder, F.G. Parak, R.D. Young, Annu. Rev. Biophys. Biophys. Chem. 17 (1988) 451–479.
[6] Y. Levy, S.S. Cho, J.N. Onuchic, P.G. Wolynes, J. Mol. Biol. 346 (4) (2005) 1121–1145.
[7] P.I. Zhuravlev, C.K. Materese, G.A. Papoian, J. Phys. Chem. B 113 (26) (2009) 8800–8812.
[8] D. Kern, E.R.P. Zuiderweg, Curr. Opin. Struct. Biol. 13 (6) (2003) 748–757 (a review of preexisting equilibrium cases).
[9] B.J. Grant, A.A. Gorfe, J.A. McCammon, PLoS Comput. Biol. 5 (3) (2009) e1000325.
[10] C. Hyeon, J.N. Onuchic, Proc. Natl. Acad. Sci. USA 104 (44) (2007) 17382–17387.
[11] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, Acc. Chem. Res. 33 (12) (2000) 889–897.
[12] A. Weis, K. Katebzadeh, P. Söderhjelm, I. Nilsson, U. Ryde, J. Med. Chem. 49 (22) (2006) 6596–6606.
[13] A. Savelyev, G.A. Papoian, J. Am. Chem. Soc. 129 (19) (2007) 6060–6061.
[14] Y.N. Vorobjev, J. Hermans, Biophys. Chem. 78 (1–2) (1999) 195–205.
[15] S. Park, A.Y. Lau, B. Roux, J. Chem. Phys. 129 (13) (2008) 134102.
[16] P.L. Freddolino, S. Park, B. Roux, K. Schulten, Biophys. J. 96 (9) (2009) 3772–3780.
[17] N.K. Banavali, B. Roux, J. Am. Chem. Soc. 127 (18) (2005) 6866–6876.
[18] I.F. Thorpe, C.L. Brooks, Proc. Natl. Acad. Sci. USA 104 (21) (2007) 8821–8826.
[19] J.N. Onuchic, P.G. Wolynes, Curr. Opin. Struct. Biol. 14 (1) (2004) 70–75.
[20] G. Torrie, J. Valleau, J. Comput. Phys. 23 (2) (1977) 187–199.
[21] P.C. Whitford, S. Gosavi, J.N. Onuchic, J. Biol. Chem. 283 (4) (2008) 2042–2048.
[22] A. Pisliakov, J. Cao, S. Kamerlin, A. Warshel, Proc. Natl. Acad. Sci. USA 106 (41) (2009) 17359–17364.

[23] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Nat. Struct. Biol. 9 (6) (2002) 425–430.
[24] S.S. Plotkin, J. Wang, P.G. Wolynes, Phys. Rev. E 53 (6) (1996) 6271–6296.
[25] S. Wu, P.I. Zhuravlev, G.A. Papoian, Biophys. J. 95 (12) (2008) 5524–5532.
[26] Or more formally $\forall \delta s > 0$, $\exists \delta \xi > 0 : \forall x : |\xi(x) - \xi(a)| < \delta\xi \rightarrow |s(x,a) - s(a,a)| < \delta s$.
[27] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman, J. Comput. Chem. 13 (8) (1992) 1011–1021.
[28] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 314 (1999) 141–151.
[29] A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. 63 (12) (1989) 1195–1198.
[30] S. Plimpton, J. Comput. Phys. 117 (1995) 1–19.
[31] A.D. MacKerell, N.K. Banavali, N. Foloppe, Biopolymers 56 (4) (2000) 257–265.
[32] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé, K. Schulten, J. Comput. Chem. 26 (16) (2005) 1781–1802.