Diffusive reaction dynamics on invariant free energy profiles

Sergei V. Krivov*‡ and Martin Karplus*‡§

*Laboratoiré de Chimie Biophysique, Institut de Science et d'Ingénierie Supramoléculaires, Université Louis Pasteur, 67000 Strasbourg, France; and
§Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

Edited by David Chandler, University of California, Berkeley, CA, and approved July 17, 2008 (received for review January 9, 2008)

A fundamental problem in the analysis of protein folding and other complex reactions in which the entropy plays an important role is the determination of the activation free energy from experimental measurements or computer simulations. This article shows how to combine minimum-cut-based free energy profiles (F_C), obtained from equilibrium molecular dynamics simulations, with conventional histogram-based free energy profiles (F_H) to extract the coordinate-dependent diffusion coefficient on the F_C (i.e., the method determines free energies and a diffusive preexponential factor along an appropriate reaction coordinate). The F_C, in contrast to the F_H, is shown to be invariant with respect to arbitrary transformations of the reaction coordinate, which makes possible partition of configuration space into basins in an invariant way. A "natural coordinate," for which F_H and F_C differ by a multiplicative constant (constant diffusion coefficient), is introduced. The approach is illustrated by a model one-dimensional system, the alanine dipeptide, and the folding reaction of a double β-hairpin miniprotein. It is shown how the results can be used to test whether the putative reaction coordinate is a good reaction coordinate.

diffusion | protein folding | one-dimensional free energy surfaces | variable diffusion coefficient

Free energy surface (FES) projected on a few progress variables (usually one or two) is often used to describe the equilibrium and kinetic properties of complex systems with a very large number (100 to 1,000 or more) of degrees of freedom. Studies of protein folding are an important example where this type of projected surface has been introduced and progress variables such as the number of native contacts and radius of gyration have been used (1–3). Most experimental analyses of protein folding have used a related approach; for example, if the distribution of folding times is exponential, it is assumed that there is a single free energy barrier along a generally unknown one-dimensional reaction coordinate. For a few systems that show more complex kinetics, the results have been interpreted in terms of projected FESs in two dimensions (4), although, again, the actual progress variables are not known. However, even when a one-dimensional single-barrier free energy projection seems adequate to describe the kinetics, there is a fundamental difficulty in determining the barrier height, because the measurements provide only one parameter (e.g., in protein folding, the rate constant of the corresponding unimolecular reaction is obtained). In such a standard "one-dimensional" analysis, the rate constant, k, is written as $k = k_0 e^{-\Delta F/kT}$, where k_0 is the preexponential factor and ΔF is the free energy of activation. Thus, there are two unknowns, k_0 and ΔF, to be determined from one measurement. For many small-molecule reactions, the entropic contribution to the barrier is negligible ($\Delta F \approx \Delta E$, the activation energy), so that a measurement of the temperature dependence of the reaction rate can be used to find ΔE and k_0, both assumed independent of temperature. However, for the protein-folding reaction and other reactions of complex systems, such as enzymatic reactions (5), the activation entropy plays an important role. As the protein folds, the loss of configurational entropy approximately cancels the stabilization of the native state by its lower energy (1), and the free energy barrier results from an imbalance between the two. Many discussions have been published concerned with the value of k_0 to use for obtaining an estimate of ΔF from rate measurements. In particular, for reactions in solution and the large motions of the polypeptide chain involved in protein folding, a Kramers-type equation (6) with diffusive prefactors is appropriate. Such prefactors are much smaller than the Eyring value of kT/h ($6 \cdot 10^{12}$ at 300 K), standardly used for gas-phase reactions. Values of k_0 on the order of 10^4 to 10^9 s^{-1} have been proposed (7–9). The "speed limit" for protein folding discussed by Kubelka *et al.* (10) essentially corresponds to a barrier-less reaction for which the rate is equal to the diffusion-limited rate coefficient. To summarize the experimental situation we quote from Yang and Gruebele (8): "Without sufficient knowledge of the critical reaction coordinate for describing the motion represented by ν^+ [here k_0] it is impossible to relate experimentally determined folding rates rigorously to computed free energy barriers." A major aim of this article is to propose a method for solving this problem.

Theoretical studies based on simulations of the reaction rate for complex systems, such as peptides and proteins, often show simple exponential kinetics (11–13). To be able to determine both the preexponential factor and the free energy barrier from simulations, it is necessary to have a method of constructing the one-dimensional projected FES in terms of an appropriate reaction coordinate, if such exists. Given this projected surface and the calculated rate from simulations one can extract the rate coefficient and the free energy barrier. In a previous article (14) we showed how to use the minimum-cut procedure (11, 15) for finding free energy barriers and constructing one-dimensional free energy profiles (FEPs). In that article, we considered the ballistic regime (i.e., the quenching interval was large enough that the number of recrossings of the transition state was negligible). The free energy of the barrier was actually determined only up to an arbitrary additive constant corresponding to a preexponential factor, which is set equal to unity; that is, the minimum-cut value used for the free energy is equal to the total number of transition (i.e., proportional to the rate). Here we focus on the diffusive regime, which is implicit in Monte Carlo (MC) simulations and is valid in many cases for molecular dynamics (MD) simulations, as in protein-folding studies. In this regime, as we show in what follows, the FEP and diffusion coefficient as a function of the reaction coordinate can be evaluated separately.

In what follows we first demonstrate the essential results for a one-dimensional system to avoid complexity. We then outline how the results are generalized to the multidimensional case; for the

Author contributions: S.V.K. and M.K. designed research; S.V.K. performed research; S.V.K. and M.K. analyzed data; and S.V.K. and M.K. wrote the paper.

The authors declare no conflict of interest.

[‡]To whom correspondence may be addressed. E-mail: krivov@isis-ulp.org or marci@tammy.harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0800228105/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0800228105

practically important case of clustered equilibrium trajectories, the method corresponds to the minimum-cut procedure for the corresponding network (11, 14, 15). Applications are made to a transition of the alanine dipeptide and to the folding reaction of a double β-hairpin miniprotein, both simulated by MD with implicit solvent.

Methodology

F_C and F_H. The conventional way to construct the projected FES is to perform equilibrium sampling of the configuration space (by MD or MC), select a progress variable (r), estimate the probability, $P(r)$, to be in particular region of (r) by binning (making a histogram of) the results, and calculate the free energy as $F(r) = -kT \ln P(r)$; an absolute reference for the free energy can be used if a unique ground is known, as in lattice simulations (12). We refer to such histogram-based free energy projections as F_H. By construction, the F_H shows the probability of the system to have particular values of the chosen variable (r), from which all of the equilibrium properties as functions of this variable can be obtained. With the further assumption that the chosen coordinate is a "good" reaction coordinate (i.e., that the projection on this coordinate preserves the system kinetics) and that the motion on this surface can be described as diffusive along the reaction coordinates (3, 16), one can obtain information about the system's dynamics from the F_H.

In many cases (11, 14), however, the standard progress variables (e.g., number of native contacts, radius of gyration) are not good reaction coordinates, because they do not preserve the barriers on the FES. Moreover, the diffusion coefficient is likely to vary in a complex way. Consequently, it is important for interpreting the simulated (or experimental) kinetics, as discussed in the introduction, to obtain the FEP as a function of a single coordinate that is a good reaction coordinate. One approach for doing this is to exploit an analogy between the system kinetics and equilibrium flow through a network (11, 14, 15). The essential element of this approach is to use the minimum-cut procedure for finding free energy barriers (11, 15) and, introducing the partition function as the reaction coordinate, to construct the FEP (14). The resulting free energy projections are referred to as "cut" FEPs (F_C), in contrast to the F_H. Other approaches for finding the reaction coordinates(s) for complex systems have been given in refs. 17–21.

On the basis of the calculated equilibrium trajectory, the partition functions of the bins of the F_H are equal, $Z_H(i) = \Sigma_j n_{ij} = N_i$, where n_{ij} is the number of transitions from bin j to bin i, and N_i is the number of times the system was found in bin i. The partition function of the cut used in the F_C between two neighboring bins i and $i + 1$ is equal to $Z_C(i, i + 1) = n_{i,i+1}$. If there are transitions between more distant bins (i.e., the quench interval at which we observe the system is so large that transitions from nonneighboring clusters occur), one has to sum over them, so that $Z_C(i, i + 1) = \Sigma_{j \le i < k} n_{jk}$. We note that although it is essential to bin the coordinate to construct the F_H, the F_C can, in principle, be obtained without binning. However, introduction of an equilibrium kinetic network (EKN), which does involve binning, is an efficient way to determine the F_C for a multidimensional system (see *The Multidimensional Case*).

F_C for Diffusive Motion. In our previous works (11, 14, 15) we estimated the reaction rate between two basins as $k_{ij} = Z_{ij}/Z_j$, where Z_{ij} is the detected number of transitions between two basins (found by the minimum-cut procedure) and Z_j is the time the system spent in basin j. This measure is valid if the quench interval dt is longer than the time to diffuse through the transition state so that the number of recrossing events is negligible ("ballistic" regime) and, at the same time, dt is shorter than the mean lifetime in the basin so that the number of transition events that are left undetected, attributable to the system going back to the original basin, is negligible. If there is a separation between the two time scales, as is often true, the quench interval can be chosen to be between the two scales, and a meaningful description of the kinetics can be obtained (22, 23). To extend the analysis to the cases in which the recrossing is essential, we consider a reaction involving diffusive motion. For this purpose we focus on a one-dimensional system and treat a region of the FEP that is flat to a good approximation; that is, $Z_H(x)$ and $D(x)$ are approximately independent of x, the reaction coordinate, in this region. This could be either a sufficiently small part of the FEP ($\Delta x \sim \sqrt{Ddt}$), so it does not change much, or an inherently flat part of the FEP, such as that in vicinity of a local maximum (e.g., a transition state) or minimum. The distance that the system moves during dt is distributed according to $P(y) = (4\pi Ddt)^{-1/2}\exp(-y^2/4Ddt)$, the free-diffusion result. The number of jumps that cross the cut at x in one direction is equal to $Z_C(x) = \int_{-\infty}^{0} yP(y)Z_H(x + y)dy = Z_H(x)(Ddt/\pi)^{1/2} = \langle |y| \rangle Z_H/2$, with $Z_H(x)$ the probability of finding the system at x, which is assumed to be constant in the interval of a few $\sqrt{Ddt}$; and $\langle |y| \rangle = \int_{-\infty}^{\infty} P(y)|y|dy$ is the mean length the system moves during dt. Thus, $D(x) = \pi/dt[Z_C(x)/Z_H(x)]^2$ together with $Z_H(x)$ give a complete description of the kinetics for a diffusive process. Because $D(y) \sim \langle \Delta y^2 \rangle/dt \sim \langle(\Delta x\ dy/dx)^2\rangle/dt \sim D(x)(dy/dx)^2$ and $Z_H(y) = Z_H(x)dx/dy$, we obtain that $Z_C(y) = Z_C(x)$ (i.e., F_C is invariant for diffusive motion; see also below).

The reaction rate between two basins in the diffusive regime is equal to the reciprocal of the mean first passage time (mfpt; $\langle t \rangle$) from one basin to the other. The analytic equation for the mfpt from A to B given by $\langle t \rangle = \int_A^B dxe^{\beta \mathrm{U}(x)}/D(x)\int_A^x dye^{-\beta U(y)}$, can be transformed to

$$\langle t \rangle = \int_A^B dx\, \frac{e^{-\beta U\,(x)}}{e^{-2\beta U\,(x)}\, D(x)} \int_A^x dye^{-\beta U(y)}$$

$$= dt/\pi \int_A^B Z_A(x)dZ_A(x)\, Z_C^{-2}\,(x), \qquad [1]$$

where $Z_A(x) = \int_A^x Z_H(y)dy$ (i.e., the partition function corresponding to the reactant region, basin A).

Invariance of F_C. The F_C is invariant, in contrast to the conventional F_H, with respect to an arbitrary continuous invertible transformation of the coordinate space. Although we showed this above for the specific case of diffusive motion, the result is true generally. In one dimension for the F_C, we have $F_C(y) = F_C(x(y))$, where $F'(y)$ is the FEP with respect to the new coordinate y, and $y(x)$ is the transformation. Thus, the transformation subjects the profile to arbitrary contraction or dilation along the coordinate axis in the one-dimensional case. Because the cut values are preserved (i.e., $n_{i,i+1}$ remain the same), the local maxima and minima of the profile are preserved, and the partition of the configuration space into free energy basins is also invariant. For the F_H, the total partition function, $Z_H(i)$ of the transformed image of the bin [i.e., the bin with borders $y_i = y(x_i)$ and $y_{i+1} = y(x_{i+1})$] is also preserved. However, the bin size dy_i changes so that the value of the mean partition function in the bin ($y_i < y < y_{i+1}$)$Z_H(y) = Z_H(i)/dy_i = Z_H(x)dx_i/dy_i$ is changed. Consequently, the free energy is transformed as $F_H(y) = F_H(x(y)) + kT \ln(|dy/dx|)$, which means that the set of local minima and maxima is not invariant in the F_H and that the partition of the configuration space into free energy basins can be altered.

This result has important practical consequences, because the FEPs are most commonly built by using putative reaction coordinates such as the number of native contacts or p_{fold} (19, 22), which are highly nonlinear functions of the Cartesian configuration space through which the kinetics proceeds. On the basis of the F_H in terms of such coordinates, the basins on the projected FES are determined, and the lowest pathways connecting them are obtained. This assumes implicitly that the diffusion coefficient is independent of the value of the reaction coordinate, which is not true in general. Because any coordinate system can be used to describe the kinetics (which is fully determined by F_C and F_H; see above), the changes in the F_H must be compensated by changes in the diffusion coefficient as a function of the reaction coordinate to keep the F_C constant; we give an example in *An Example: Comparison of* $\mathrm{F_H}$ *and* $\mathrm{F_C}$ *for a One-Dimensional Model System*. The above result indicates that an important attribute of the F_C, in contrast to the F_H, is that different putative reaction coordinates can be used to analyze the FES because the barriers and minima are preserved.

Natural Coordinate. One application of the invariance of the F_C is that, knowing both F_H and F_C, it is straightforward to make a continuous invertible transformation to a coordinate $y(x)$ such that they are proportional to each other; that is, that $Z_C(y)/Z_H(y) = \text{const}$ (independent of y) and the diffusion coefficient is constant. We call such a coordinate a "natural" coordinate. Because $Z_C(x) = Z_C(y)$ is invariant and $Z_H(y) = Z_H(x)dx/dy$, one obtains $dy/dx = \text{const} \times Z_H(x)/Z_C(x)$. A related but conceptually different approach to constructing such a coordinate on the basis of the F_H specific for the case of p_{fold} as the reaction coordinate was described recently by Rhee and Pande (16) [see supporting information (SI)].

Optimum One-Dimensional Projection. It is reasonable to assume that any "bad" projection that results in overlapping of different parts of the configuration space will result in faster kinetics (i.e., in a smaller mfpt). Clearly, the longest mfpt is obtained on the original FES or from a projection where no such overlapping occurs. Hence, the maximum value of the integral in Eq. **1** can serve as a definition of the best one-dimensional projection. Taking Z_A (the partition function of the reactant region) as the reaction coordinate (x) in Eq. **1** gives for the mfpt $\langle t\rangle = dt/\pi\int_A^B dZ_A Z_A Z_C^{-2}(Z_A)$. Assuming that the $Z_C(Z_A)$ for different values of Z_A are independent, the maximum $\langle t\rangle$ is attained when $Z_C(Z_A)$ takes the minimal value for each value of Z_A (i.e., the definition of the FEP introduced in ref. 14, which optimizes F_C, maximizes the barrier as a function of Z_A).

A referee pointed to a very interesting paper in a book (24) of which we were unaware. That paper also considers two types of FEPs: one is the standard one, identical with $F_H(q)$, and the other is related to, but different from, $F_C(q)$ except for a special case. We compare the two approaches in the SI.

An Example: Comparison of F_H and F_C for a One-Dimensional Model System

We use a simple model potential energy surface (PES), $U(x) = -\cos(x)$, with $U(x)$ in units of kT and x in radians. For this one-dimensional system, the PES is the same as the FES and the FEP. The dynamics were simulated by performing MC sampling at a temperature (in energetic units of kT) equal to 0.5 for 10^7 steps, with the steps selected from a Gaussian distribution with zero mean and root-mean-square deviation (rmsd) of 0.1 (i.e., the diffusion coefficient is $D = \langle\Delta x^2\rangle/2\Delta t = 0.1^2/2$). At quench interval of one step ($dt = 1$), the observed dynamics is in the diffusive regime because of the use of MC sampling. To compute the $F_H(x)$ we partitioned the x axis into bins of size of 0.01. The mean value of the partition function $Z_H(x)$ in bin i, associated with the point x, is equal to the number of times the system was found in this bin, divided by the size of the bin. The value of the partition function $Z_C(x)$ of the $F_C(x)$ at point x is equal to the number of times the system's trajectory crossed this point in one direction (or, for equilibrium sampling, in both directions divided by 2). With the relation $F(x) = -kT\ln(Z(x))$, one obtains the $F_H(x)$ and $F_C(x)$.

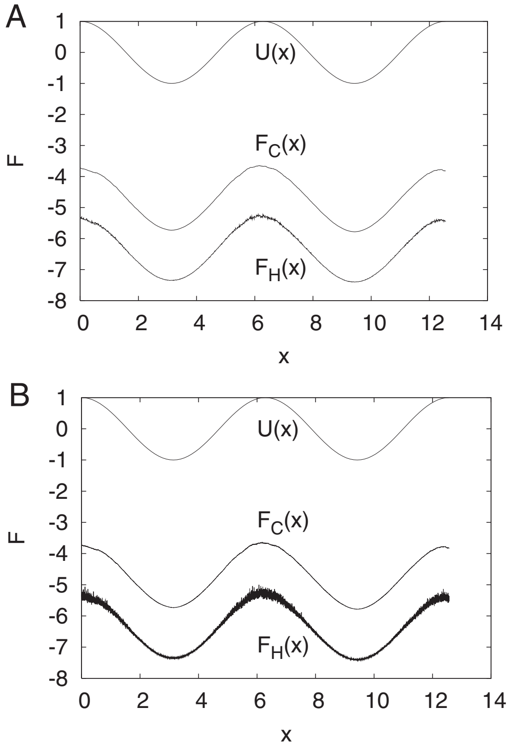

Fig. 1. Model PES (= FES) $U(x) = -\cos(x)$ (solid line) together with reconstructed F_C and F_H. The distance between F_C and F_H is equal to $kT\ln(\sqrt{Ddt/\pi}) = 0.25\ln(0.01/2/\pi) \approx 1.61$ (see below). (*A*) Bin size of 0.01. (*B*) Bin size of 0.001 (see text).

Fig. 1 shows that $F_H(x)$ and $F_C(x)$ are essentially identical to $U(x)$ except for the conventional additive constant. The $F_H(x)$ has slightly more noise in the regions of the maxima, where sampling is limited. We note that to build a histogram [$F_H(x)$], one specifies a bin width that provides the optimum tradeoff between good statistics (large width) and good spatial resolution (small width). For the $F_C(x)$ there is no such problem; one can put surfaces arbitrarily close to each other, and one still obtains meaningful results. With a bin size of 0.001 (instead of 0.01) based on the same MC simulation, the $F_H(x)$ shows increased fluctuation, whereas the $F_C(x)$ does not (see Fig. 1*B*).

To illustrate the invariance of $F_C(x)$, the highly nonlinear transformation $y(x) = x + \sin(4x)/4$ was chosen. Fig. 2*A* shows the result. Both the analytical transformation and the trajectory of the original MC simulation transformed to the new reaction coordinate were used to obtain the profiles; they are essentially identical. Although $U(y)$ and the $F_C(y)$ change shape, the important point is that both still have two minima separated by a barrier. By contrast, the $F_H(y)$ has eight minima separated by seven barriers. For the present case, use of the angular coordinate x satisfies the condition that the $F_C(x)$ and the $F_H(x)$ be identical [up to a constant $-kT\log(\sqrt{Ddt/\pi})$] as shown in Fig. 1, whereas they are not for the transformed coordinate $y(x)$. The diffusion constant obtained from $D(y) = D(x)(dy/dx)^2$ mirrors the transformed $F_H(y)$ (see Fig. 2*A*).

Fig. 2*B* shows the $F_C(z)$ [the $F_H(z)$ is not shown, because it coincides with $F_C(z)$ by construction] along the natural coordinate z, where $dz/dy = Z_H(y)/Z_C(y)$, and $Z_C(y)$ and $Z_H(y)$ correspond to the $F_C(y)$ and $F_H(y)$ shown in Fig. 2*A*. The $F_C(z)$ along the natural coordinate z is identical to that in Fig. 1 except that the x axis is more extended, whereas the $Z_C(x)$ and $Z_H(x)$ differ by a multiplicative constant $(D/\pi)^{1/2}$ in Fig. 1; they are identical in Fig. 2*B*.

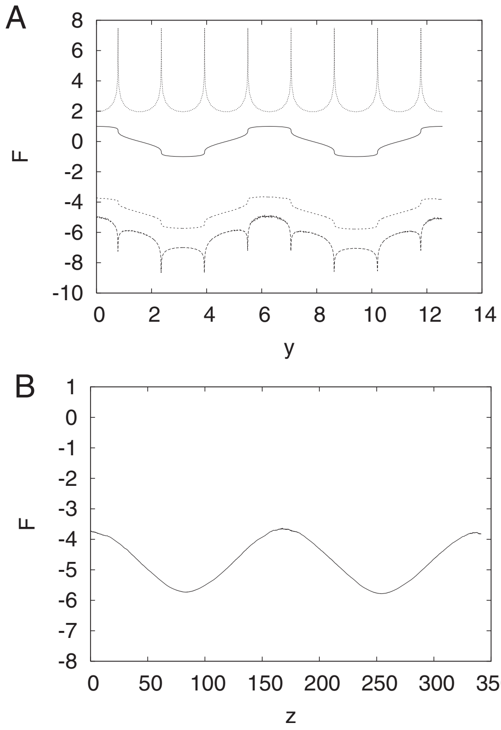

Fig. 2. Transformations of the FEPs. (*A*) Model PES and reconstructed FEPs along the coordinate $y(x) = x + \sin(4x)/4$. The lines are (top to bottom) $-kT \ln(D(y))$, where $D(y)$ is the diffusion constant as function of y, $U(x(y))$, F_C, and F_H. (*B*) F_C (and F_H) from *A* transformed "back" to natural coordinate *z*.

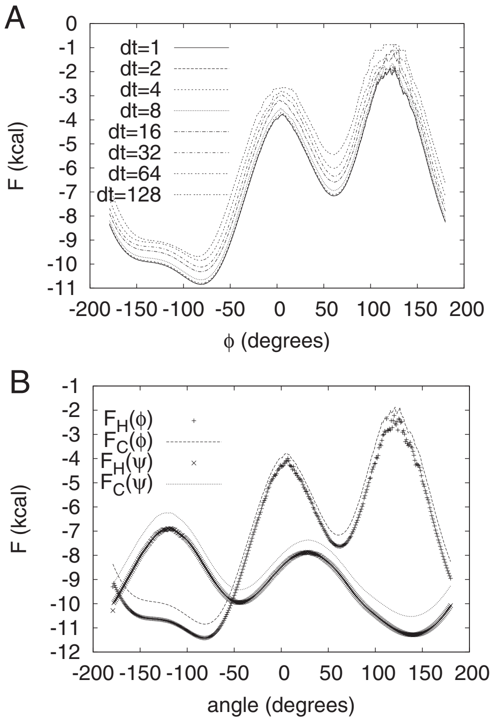

Fig. 3. FEPs for alanine dipeptide. (*A*) F_C along the ϕ dihedral angle for various quench time intervals *dt* (given in MD steps). (*B*) F_C and F_H along the ϕ and ψ dihedral angles.

For this model system, Eq. **1** gives values for the mfpt $\langle t\rangle$ between 49,781 and 46,100 steps for $dt = \{2^0 \ldots 2^8\}$, which is in agreement with value of 47,169 steps found from the MC simulation. Instead of comparing mfpts, one can compare the effective number of transitions between the basins, estimated via the mfpt as $n_{ij} = n_{ji} = Z_j/\langle t_{ij}\rangle$, and determined directly by counting the number of transitions between the basins in the trajectory. For $dt = 1$, the numbers are 95.6 and 106, respectively. The statistical uncertainty (variance) of the number of transitions estimated via the mfpt is half as large as the one obtained by actual counting; for example, the variances estimated with the trajectory divided into 10 pieces, 10^6 steps each, are 1.6 and 3.4, respectively.

The relation between F_C and F_H $[Z_C = Z_H(Ddt/\pi)^{1/2}]$, derived above, is valid for inherently diffusive motion, as described by MC dynamics. For MD the motion can also be diffusive (e.g., as it is in most cases of protein folding), but on a short time scale (when the stochastic approximation is not yet valid) the motions are essentially ballistic so that $\Delta x \sim vdt$ and $Z_C \sim Z_H vdt$. If one examines a trajectory and changes only dt during the analysis, then, because $Z_H \sim 1/dt$, $Z_C \sim (D/dt)^{1/2}$ in the diffusive regime. This is in contrast with $Z_C \sim$ const for the ballistic regime and can be used, therefore, to distinguish the two regimes.

The dependence of the F_C on dt for the model potential is discussed in the SI.

Alanine Dipeptide. The kinetics of the alanine dipeptide was simulated with the CHARMM program (25) by using the polar hydrogen force field (26) with the ACE2 implicit solvent model (27) for 10^8 steps; each step was 2 fs. A temperature of 400 K was used to ensure coverage of the important regions of the FES; the temperature was controlled by a Nose–Hoover thermostat.

Fig. 3*A* shows the $F_C(\phi)$ along the ϕ dihedral angle obtained with various quench time intervals. For $dt = 1$, 2, 4 MD steps, the F_C is approximately constant (i.e., the motion is ballistic). For $dt = 8$ MD steps, F_C changes by a constant (i.e., the motion starts to deviate from ballistic and can be approximately described as diffusive). This is supported by the fact that the difference between $F_C(\phi)$ for $dt = 8$ and $dt = 64$ is approximately constant and equal to 0.8, which is close to the exact value of $kT \ln(8)/2 = 0.83$ for diffusive motion. Nonmonotonic changes in the distances between the FEPs arise from the fact that motion is not completely stochastic and some correlations are still present.

Fig. 3*B* shows the F_C and F_H along the ϕ and ψ dihedral angles. There is a notable difference between the profiles, with the difference depending on the angle, which indicates that the diffusion coefficient is not constant. The difference between F_C and F_H for both angles (data not shown) can be approximated by $a + b\cos(\phi)$ $[a + b\cos(\psi)]$ with $b \approx 0.4$ kcal for ϕ and $b \approx 0.35$ kcal for ψ.

The analysis has shown that the dynamics of the dipeptide along the dihedral angles can be considered to be diffusive for time steps of ≥8 fs with a diffusion coefficient [determined as $D(x) = \pi/dtZ_C^2(x)/Z_H^2(x)$ for $dt = 8$ fs] ranging from 3 deg^2/fs (at angle values close to ±180) to 7 deg^2/fs (at angle values close to 0). The mfpt to go from C_{7eq} ($\phi = -79$, $\psi = 133$) to the C_{7ax} ($\phi = 63$, $\psi = -77$) as estimated with Eq **1**, based on the calculated profile for $dt = 8$ fs, is equal to $3.8 \cdot 10^6$ steps. The result is in good agreement with the number ($3.9 \cdot 10^6$ steps) obtained by direct counting (based on 24 transitions). It indicates that ϕ alone is a good reaction coordinate for describing transition between C_{7eq} and C_{7ex}. This behavior is somewhat surprising, because there are two major transition states on the alanine dipeptide FES for this transition: the associated (ϕ, ψ) values are ($\phi = 1$, $\psi = -71$) and ($\phi = 9$, $\psi = 89$) (28). Because the ϕ values are nearly the same, they match in the ϕ projection, resulting in the highest possible (correct) single barrier.

The usual harmonic approximation in the Kramers formulation (29) for the mfpt is $\langle t\rangle = 2\pi/(\beta\omega\omega^\dagger D^\dagger)\exp(\beta\Delta G) = 2\pi/(\beta\omega\omega^\dagger D^\dagger)Z_H/Z_H^\dagger$ (where † denotes the values at the transition state). By introducing the expression for the diffusion coefficient $D = \pi/dt(Z_C/Z_H)^2$, the equation can be transformed to

$$\langle t \rangle = \frac{2dt}{\beta\omega\omega\dagger}\frac{Z_H\, Z_H^\dagger}{Z_C^{\dagger 2}}. \qquad [2]$$

Using the values of $\omega = 0.04$ and $\omega^\dagger = 0.077$ obtained by fitting the potential in Fig. 3*B*, we obtain $\langle t \rangle = 3.2 \cdot 10^6$ steps. If one takes account of the anharmonicity of the ground state, $\langle t \rangle = 3.5 \cdot 10^6$ steps, which is in good agreement with the exact results.

The Multidimensional Case

Generalization of the F_C-based approach to the multidimensional case is straightforward, in principle, because $Z_C(S)$ is defined for every surface S in configurational space as the number of transitions through it. However, the specification of the surface in the multidimensional space is evidently more complex than that in the one-dimensional case. The invariance to an arbitrary continuous invertible transformation of configurational space remains valid, because the number of transitions through the surface is preserved. To find the ensemble of transition states between two points, one determines the surface with the minimal partition function that separates these two points (15, 30). However, when one approximates the flow by a finite trajectory, one has to take into consideration the fact that for a configuration space of more than two dimensions the trajectory essentially never crosses itself. Thus, one can always find a surface that separates any two points and crosses the trajectory only once. To avoid this problem one can coarse-grain the space and gather nearby points of the trajectory into clusters, based on an appropriate criterion (e.g., rmsd, secondary structure strings, number of native contacts). This leads to an EKN consisting of a set of states and the transitions between them (14). Instead of specifying the cutting surface, one then needs only list the edges of the network, which are cut by the surface. The flow over the surface is mapped onto the cuts of the network; the minimum cut is used to find the links corresponding to the transition-state ensemble (15).

Folding of the Beta3s Miniprotein. Folding of the 20-residue Beta3s double-hairpin miniprotein has been studied (31, 32), based on a 20-ms equilibrium trajectory calculated with the solvent-accessible surface area implicit solvent model (33). Detailed analyses of the folding behavior of this system and its folding network were made. Secondary structure clustering and rmsd clustering with an all-atom rmsd of 2.5 Å were compared, and it was shown that the basins on the FEP obtained from the two types of clustering were in good agreement. However, it was found that folding rate was faster by almost an order of magnitude in the case of secondary structure clustering than that obtained with rmsd clustering. To interpret this result we consider the dependence of FEPs obtained with the two types of clustering on the quench interval dt. Snapshots of trajectory taken with quench interval dt were clustered into an EKN and the FEPs were constructed with the pfoldf algorithm (14). Fig. 4*A* shows the FEPs obtained with rmsd 2.5-Å clustering. Increasing the quench interval generally leads to a less connected network; thus, the profile for $dt = 2$ has more noise. However, the $dt = 2$ profile is almost equidistant from the $dt = 1$ profile, with a spacing of $0.35(\ln(\sqrt{2}))$; that is, the profiles are proportional to $dt^{-1/2}$, consistent with the diffusive regime. Use of either profile gives the same value for the diffusion coefficient and leads to the same temporal behavior.

Fig. 4*B* shows the FEPs obtained with secondary structure clustering. The profiles are also almost equidistant with distance of $0.7(\ln(2))$; that is, the profiles are proportional to dt^{-1}, which is inconsistent with the diffusive regime. The profile obtained with a larger dt has a smaller diffusion coefficient ($D \sim Z_C^2/Z_H^2/dt \sim dt^{-2}/dt^{-2}/dt \sim dt^{-1}$) and, thus, exhibits slower kinetics. The dt^{-1} behavior can be explained if one supposes that with the secondary structure clustering, "shortcuts" between different parts of configuration space are possible (i.e., particular secondary structure strings correspond to significantly different configurations). Each such shortcut corresponds to a jump (with length independent of dt). Thus, $Z_C(x) = \langle |y| \rangle\, Z_H(x)/2 \sim dt^{-1}$ (see above), where $\langle |y| \rangle$ is the mean length of the jump. The analysis leads to the conclusion that rmsd clustering is appropriate for the kinetics of folding of Beta3s, whereas secondary structure clustering introduces a significant number of shortcuts, making the description of the system kinetics as a diffusive process inconsistent (i.e., profiles obtained at different quench intervals lead to different behavior).

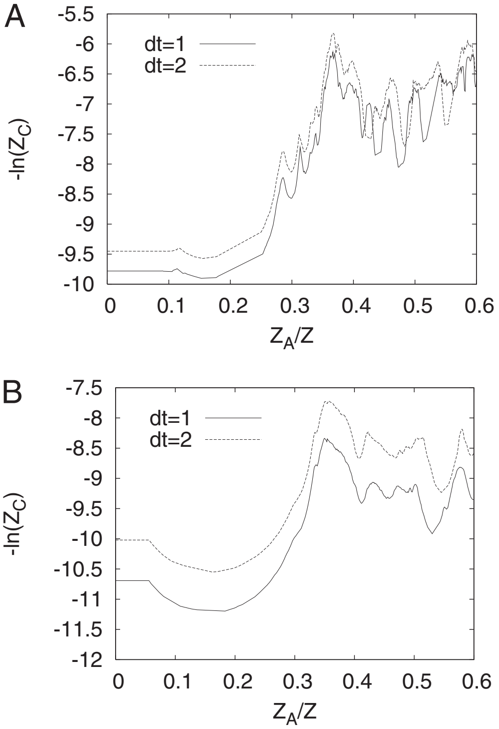

Fig. 4. $F_C/kT = -\ln(Z_C)$ for $dt = 1, 2$ for Beta3s. (*A*) rmsd clustering with a 2.5-Å cutoff radius shows $dt^{-1/2}$ behavior. (*B*) Secondary structure clustering shows dt^{-1} behavior. For brevity we show just the part of the FEP $0 < Z_A/Z < 0.6$. The native state occupies the region $0 < Z_A/Z < 0.35$, followed by the denatured state, which includes several enthalpic basins.

The advantage of the secondary structure clustering over the rmsd clustering is a small running time that increases linearly with trajectory size, whereas that for the latter grows quadratically. We suggest the following simple clustering method, which combines strong points of both methods. The configurations are in the same cluster only when they have equal secondary structures and their rmsd is less than the given threshold. Thus, rmsd is calculated only between configurations with equal secondary structures (i.e., the latter is used a hash function). Tests on the Beta3s miniprotein showed that the proposed clustering method is at least two orders of magnitude faster than rmsd and provides FEPs consistent with diffusive dynamics, unlike using secondary structure clustering alone.

Concluding Discussion

This article examines the properties of minimum-cut-based FEP (F_C) and shows, in particular, that in the diffusive regime the diffusion coefficient (possibly coordinate-dependent) can be obtained directly from the F_C and, together with the histogram-based free energy profile (F_H), provides a complete description of the kinetics and the equilibrium properties. This makes possible the decomposition of the calculated rate into a preexponential factor (diffusion coefficient) and a free energy of activation. Alternatively, one can obtain the preexponential factor from $k_0 = e^{\beta\Delta F}\langle t\rangle^{-1}$, where $\langle t\rangle$ is the mfpt from the analytical solution and ΔF is the F_C barrier height. An important property of the F_C is that they are invariant under an arbitrary invertible transformation of the reaction coordinate, which means that the F_C can be used, in contrast to F_H, to partition the FES into basins in an invariant way (i.e., the number of barriers and minima and their heights along any appropriate reaction coordinate should be the same). By comparing the calculated kinetics with that obtained directly from the simulation, one can test whether the reaction coordinate used to project the FES is appropriate. For example, one can compare the mfpt found from simulations with that obtained from the standard analytical solution for one-dimensional diffusion. Moreover, the F_C is less sensitive than the F_H to limited statistics (i.e., there is no "tradeoff" between accuracy and resolution for the F_C, in contrast to F_H).

The partition function is introduced as a reaction coordinate, because it is among the simplest and most flexible coordinates that increase monotonically as the system goes from the initial to the final state. If there are several well defined pathways, this reaction coordinate will adapt its shape to them and progress mainly along the pathways (see an example in figure 5 of ref. 14). If the FEP is accurate, it describes the essence of the reaction kinetics by showing the barriers and basins on the way from the initial to the final state. Because the chosen progress coordinate is very flexible, the obtained FEP is likely to be the best way of projecting the FES onto a one-dimensional coordinate (see about *Optimum One-Dimensional Projection* above). Moreover, although the partition function may seem abstract (as does p_{fold}, but see ref. 18), one can identify the structures associated with most important pathways by postprocessing the profiles.

Finally, we mention that recently there has been increasing discussion of the fact that reactions that in the past had been described in terms of a one-dimensional FES [e.g., enzymatic reactions (34) or the analysis of single-molecule experiments (35)] in fact require more than one dimension for a valid description. Although we have illustrated the present methodology by applying it to protein folding, we note that the approach is perfectly general. There may be practical limitations introduced by the difficulty of obtaining the necessary data. Nevertheless, the concept that it is possible to introduce a one-dimensional FES that contains all of the information necessary to describe the kinetics of reactions in complex systems should make the present approach of widespread interest. By considering time series of FRET efficiency, for example, one can obtain an invariant FEP together with the coordinate-dependent diffusion coefficient. The approach also suggests that the biasing potential in adaptive biased simulation (e.g., adaptive umbrella sampling) should be applied to "flatten" the invariant quantity F_C, instead of F_H, to speed up the kinetics of equilibration.

ACKNOWLEDGMENTS. We thank Amedeo Caflisch and his coworkers, who performed the long Beta3s simulation. The part of the research performed at Harvard was supported by a grant from the National Institutes of Health. S.V.K. is supported by the CHARMM Development Project.

1. Dobson CM, Sali A, Karplus M (1998) Protein folding: A perspective from theory and experiment. *Angew Chem Int Ed* 37:868–893.
2. Shea JE, Brooks CL (2001) From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu Rev Phys Chem* 52:499–535.
3. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG (1996) Protein folding funnels: The nature of the transition state ensemble. *Fold Des* 1:441–450.
4. Sabelko J, Ervin J, Gruebele M (1999) Observation of strange kinetics in protein folding. *Proc Natl Acad Sci USA* 96:6031–6036.
5. Karplus M (2000) Aspects of protein reaction dynamics: Deviations from simple behavior. *J Phys Chem B* 104:11–27.
6. Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7:284–304.
7. Hagen SJ, Hofrichter J, Szabo A, Eaton WA (1996) Diffusion-limited contact formation in unfolded cytochrome c: Estimating the maximum rate of protein folding. *Proc Natl Acad Sci USA* 93:11615–11617.
8. Yang WY, Gruebele M (2003) Folding at the speed limit. *Nature* 423:193–197.
9. Chan HS, Dill KA (1998) Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins* 30:2–33.
10. Kubelka J, Hofrichter J, Eaton WA (2004) The protein folding "speed limit." *Curr Opin Struct Biol* 14:76–88.
11. Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci USA* 101:14766–14770.
12. Palyanov AY, Krivov SV, Karplus M, Chekmarev SF (2007) A lattice protein with an amyloidogenic latent state: Stability and folding kinetics. *J Phys Chem B* 111:2675–2687.
13. Socci ND, Onuchic JN, Wolynes PG (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *J Chem Phys* 104:5860–5868.
14. Krivov SV, Karplus M (2006) One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J Phys Chem B* 110:12689–12698.
15. Krivov SV, Karplus M (2002) Free energy disconnectivity graphs: Application to peptide models. *J Chem Phys* 117:10894–10903.
16. Rhee YM, Pande VS (2005) One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J Phys Chem B* 109:6780–6786.
17. Best RB, Hummer G (2005) Reaction coordinates and rates from transition paths. *Proc Natl Acad Sci USA* 102:6732–6737.
18. Ma A, Dinner AR (2005) Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* 109:6769–6779.
19. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1998) On the transition coordinate for protein folding. *J Chem Phys* 108:334–350.
20. Das P, Moll M, Stamati H, Kavraki LE, Clementi C (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 103:9885–9890.
21. Mu Y, Nguyen PH, Stock G (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* 58:45–52.
22. Chandler D (1978) Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J Chem Phys* 68:2959–2970.
23. Chandler D (1987) *Introduction to Modern Statistical Mechanics* (Oxford Univ Press, New York), p 288.
24. E W, Vanden-Eijnden E (2004) Metastability, conformational dynamics, and transition pathways in complex systems. *Multiscale Modelling and Simulation*, eds Attinger S, Koumoutsakos P (Springer), pp 277–312.
25. Brooks BR, *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4:187–217.
26. Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. *J Chem Phys* 105:1902–1921.
27. Schaefer M, Bartels C, Karplus M (1998) Solution conformations and thermodynamics of structured peptides: Molecular dynamics simulation with an implicit solvation model. *J Mol Biol* 284:835–848.
28. van der Vaart A, Karplus M (2005) Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *J Chem Phys* 122:114903.
29. Levy RM, Karplus M, McCammon JA (1979) Diffusive Langevin dynamics of model alkanes. *Chem Phys Lett* 65:4–11.
30. Truhlar D, Garrett B, Klippenstein S (1996) Current status of transition-state theory. *J Phys Chem* 100:12771–12800.
31. Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc Natl Acad Sci USA* 97:10780–10785.
32. Rao F, Caflisch A (2004) The protein folding network. *J Mol Biol* 342:299–306.
33. Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins* 46:24–33.
34. Benkovic SJ, Hammes GG, Hammes-Schiffer S (2008) Free-energy landscape of enzyme catalysis. *Biochemistry* 47:3317–3321.
35. Min W, Xie X, Bagchi B (2008) Two-dimensional reaction free energy surfaces of catalytic reaction: Effects of protein conformational dynamics on enzyme catalysis. *J Phys Chem B* 112:454–466.