JCTC Journal of Chemical Theory and Computation

Article

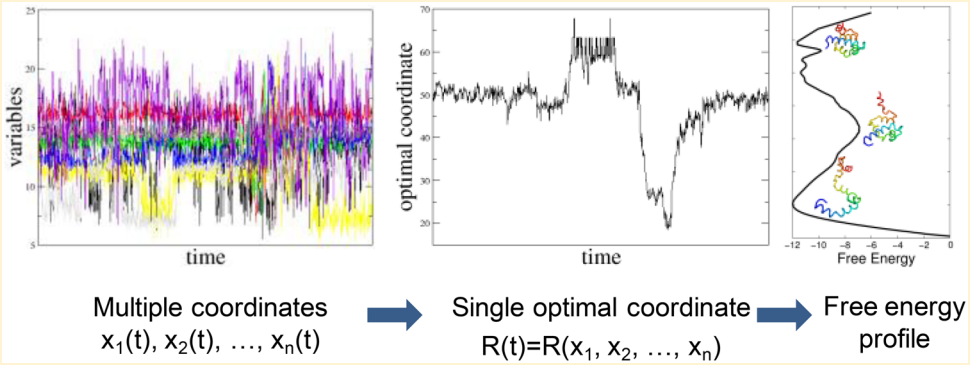
Cite This: J. Chem. Theory Comput. 2018, 14, 3418−3427

pubs.acs.org/JCTC

Protein Folding Free Energy Landscape along the Committor - the Optimal Folding Coordinate

Sergei V. Krivov*

Astbury Center for Structural Molecular Biology, University of Leeds, Leeds LS2 9JT, United Kingdom



ABSTRACT: Recent advances in simulation and experiment have led to dramatic increases in the quantity and complexity of produced data, which makes the development of automated analysis tools very important. A powerful approach to analyze dynamics contained in such data sets is to describe/approximate it by diffusion on a free energy landscape - free energy as a function of reaction coordinates (RC). For the description to be quantitatively accurate, RCs should be chosen in an optimal way. Recent theoretical results show that such an optimal RC exists; however, determining it for practical systems is a very difficult unsolved problem. Here we describe a solution to this problem. We describe an adaptive nonparametric approach to accurately determine the optimal RC (the committor) for an equilibrium trajectory of a realistic system. In contrast to alternative approaches, which require a functional form with many parameters to approximate an RC and thus extensive expertise with the system, the suggested approach is nonparametric and can approximate any RC with high accuracy without system specific information. To avoid overfitting for a realistically sampled system, the approach performs RC optimization in an adaptive manner by focusing optimization on less optimized spatiotemporal regions of the RC. The power of the approach is illustrated on a long equilibrium atomistic folding simulation of HP35 protein. We have determined the optimal folding RC - the committor, which was confirmed by passing a stringent committor validation test. It allowed us to determine a first quantitatively accurate protein folding free energy landscape. We have confirmed the recent theoretical results that diffusion on such a free energy profile can be used to compute exactly the equilibrium flux, the mean first passage times, and the mean transition path times between any two points on the profile. We have shown that the mean squared displacement along the optimal RC grows linear with time as for simple diffusion. The free energy profile allowed us to obtain a direct rigorous estimate of the pre-exponential factor for the folding dynamics.

1. INTRODUCTION

Due to advances in computer hardware and simulation methodology, it is becoming increasingly easier to generate large simulation data sets of complex molecular systems, with a prominent example being the long equilibrium trajectories of fast folding proteins.[1,2] Because of the complexity of dynamics and high-dimensionality of the resulting trajectories, the generation of many trajectories *per se* is not sufficient to provide full scientific insight. Eventually it becomes necessary to synthesize the data into as faithful as possible a picture of the process of interest. Given the growing size and complexity of simulations, analysis and interpretation of such data are widely recognized as fundamental bottlenecks in the application of atomistic simulations.[3−6]

A fundamental way to analyze a simulation is to determine the underlying free energy landscape, i.e., the free energy as a function of one or more reaction coordinates (RCs), collective variables (CV), or order parameters.[1,5,7−10] Generally, one is interested in finding free energy minima or metastable states, pathways, transitions states (TS), and free energy barriers. The major difficulty in such an analysis is the selection of appropriate RCs. A poorly chosen RC can result in a misleadingly simple free energy landscape with missing minima and the absence or underestimation of barriers.[3,7] Experience has shown that RCs chosen based on intuition or using common methods such as principal component analysis (PCA) are usually suboptimal. Hence a large number of methods have been suggested to determine good RCs or CVs in an

Received: January 31, 2018
Published: May 23, 2018

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

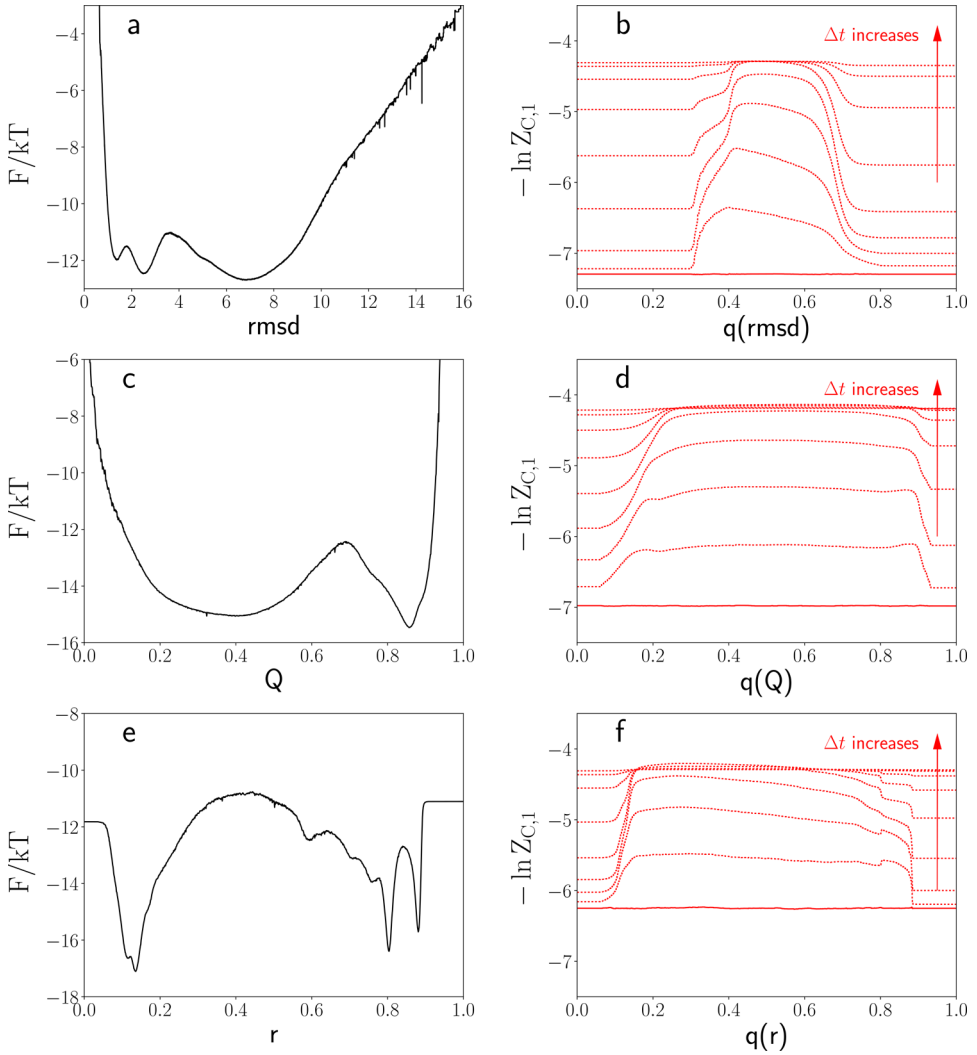


Figure 1. Free energy profiles (a, c, e) and $Z_{C,1}$ profiles (b, d, f) for different RCs. To apply the committor validation test, coordinate r is first transformed to the committor as a function of this coordinate $r \rightarrow q(r)$ using the Markov state model formalism.[5,27] Then $Z_{C,1}(q(r), \Delta t)$ profiles are computed for $\Delta t/\Delta t_0 = 1, 4, 4^2, \cdots, 4^8$. The closer $Z_{C,1}(q(r), \Delta t)$ to $N_{AB} = 73$ is, the more optimal coordinate r is. For the optimal coordinate or committor $Z_{C,1}(q, \Delta t) = N_{AB}$. For suboptimal coordinates $Z_{C,1}(q(r), \Delta t_0) > N_{AB}$ or $-\ln Z_{C,1}(q(r), \Delta t_0) < -\ln N_{AB}$. As Δt increases $-\ln Z_{C,1}(q(r), \Delta t)$ increases as well and reaches the limiting values of $-\ln N_{AB}$ for large Δt. a and b) the C_α root-mean-square deviation from the native structure. c and d) the fraction of native contacts. e and f) putative RC obtained via the nonparametric optimization after 200 iterations.

automated and unbiased way.[4,8,9,11−17] For recent reviews see refs 5, 6, 10, and 18.

Optimal RCs are an important class of RCs, which are selected in an optimal way so that the corresponding diffusive model can be used to compute some properties of the dynamics exactly.[5] For equilibrium dynamics between two states (e.g., protein folding) such an optimal RC is known as the committor or p_{fold}, in the context of protein folding dynamics. p_{fold} equals the probability to fold before unfolding, starting from the current position. To define it explicitly, consider a system where the evolution of probability density $P(\boldsymbol{X}, t)$ is described by the Fokker−Planck (diffusion) equation corresponding to the overdamped Langevin equation

$$\partial P(\boldsymbol{X}, t)/\partial t = \nabla\cdot[e^{-\beta U(\boldsymbol{X})}D(\boldsymbol{X})\nabla(e^{\beta U(\boldsymbol{X})}P(\boldsymbol{X}, t))]$$

where $\boldsymbol{X}$ denotes the position in multidimensional configuration space, U is the potential energy, D is the diffusion tensor, $\beta = 1/(kT)$, k is the Boltzmann constant, and T is temperature. Given two boundary states A and B, the committor, $q(\boldsymbol{X})$, is the solution of the adjoint equation[19,20]

$$\nabla\cdot[e^{-\beta U(\boldsymbol{X})}D(\boldsymbol{X})\nabla q(\boldsymbol{X})] = 0 \quad (1)$$

with boundary conditions $q = 0$ for $\boldsymbol{X} \in \partial A$ and $q = 1$ for $\boldsymbol{X} \in \partial B$. The committor is thus a complex, high-dimensional function, which is the solution to the high-dimensional partial differential equation. It has been determined accurately only for a small number of low-dimensional model systems.[21] Determining the committor for a realistic complex system of interest is a very difficult unsolved problem. Moreover, in practice, one needs to determine the committor from a long equilibrium trajectory, rather than from $U(\boldsymbol{X})$ and $D(\boldsymbol{X})$. While a number of approaches have been suggested to determine the committor,[5] they all have serious drawbacks. In particular, putative RCs determined for realistic systems cannot pass proposed committor validation tests.[8,22] Here we present a solution to this important problem. We describe an approach that accurately determines the committor, so it can pass the validation tests, and illustrate its performance by analyzing a long equilibrium protein folding trajectory.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

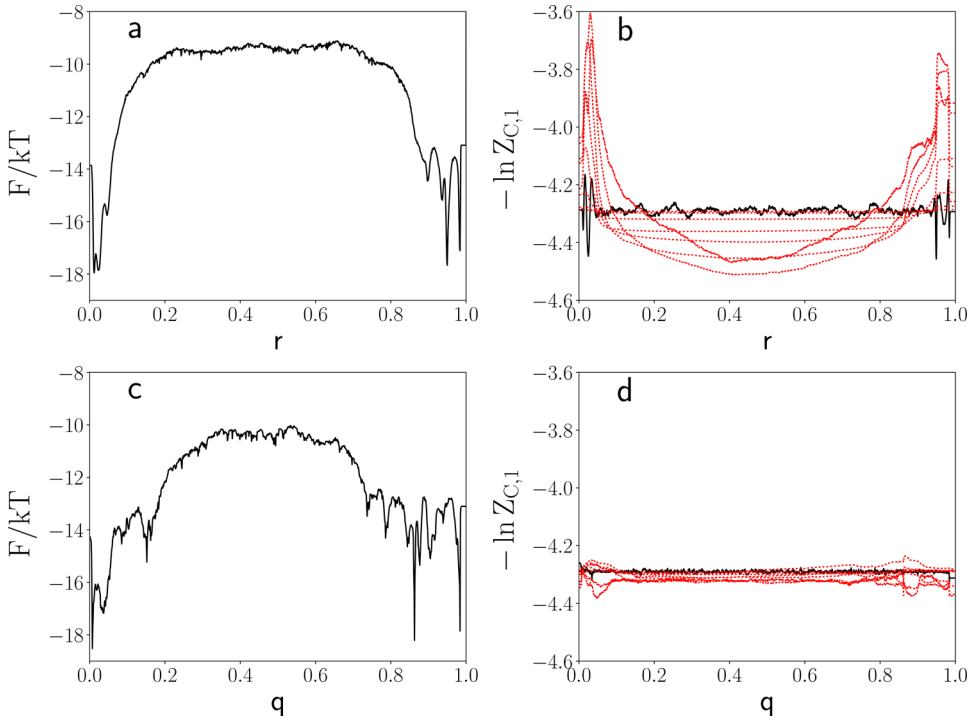


Figure 2. Optimization of RC for HP35 double mutant. The free energy profile (a) and $Z_{C,1}$ profiles (b) for RC r obtained with the nonparametric approach. While RC r is close to the committor it still deviates from it. The free energy profile (c) and $Z_{C,1}$ profiles (d) for RC q obtained with the adaptive nonparametric approach. RC q closely approximates the committor. $Z_{C,1}(r, \Delta t)$ are computed for $\Delta t/\Delta t_0 = 1, 4, 4^2, \cdots, 4^8$. Solid black lines on (b) and (d) show $Z_{C,1}(r, \Delta t_0)$.

2. METHOD

2.1. Nonparametric Variational Optimization of RCs.

Variational approaches appear to be most promising for RC optimization.[5] A functional form (FF) with many parameters $R(X, \alpha_i)$ is suggested as an approximation to an RC. One numerically optimizes the parameters α_i by optimizing a particular functional, for example, the probability of being on a transition path,[8,23] the likelihood functional,[17] the cut profiles,[9,24,25] or the total squared displacement.[22,26] Here we consider the last one. Given a long equilibrium multidimensional trajectory $X(k\Delta t_0)$, where Δt_0 is the trajectory sampling interval, one computes the reaction coordinate time-series $r(k\Delta t_0) = R(X(k\Delta t_0), \alpha_i)$. Here and below r defines an arbitrary reaction coordinate, while q is reserved for the committor. Given two boundary states A and B, the optimal coordinate between them (the committor) is the one that provides a minimum to the total squared displacement $\Delta r^2 = \sum_k [r(k\Delta t_0 + \Delta t_0) - r(k\Delta t_0)]^2$, under the constraints that $r(k \in A) = 0$ and $r(k \in B) = 1$, i.e., the boundary states A and B map to 0 and 1, respectively. It is straightforward to prove this principle by assuming that the system dynamics is described by a Markov state model. The total squared displacement equals $\Delta r^2 = N\sum_{ij} P_{ji}(\Delta t) P_i (r_j - r_i)^2$, where $P_{ji}(\Delta t)$ is the transition probability matrix from state i to state j after Δt, P_i is the equilibrium probability, N is the total number of snapshots in the trajectory, and r_i is the position of microstate i on the RC. Differentiating with respect to r_k and assuming the detailed balance $P_{ji}(\Delta t)P_i = P_{ij}(\Delta t)P_j$, one obtains the following equation for committor q[5,22]

$$\sum_j P_{jk}(\Delta t)(q_j - q_k) = 0 \tag{2a}$$

$$q_A = 0, \quad q_B = 1 \tag{2b}$$

Note that the assumption that systems dynamics is described by a Markov state model is used only for the derivation of equations. One does not need to perform the actual construction of such a model, which means that this assumption does not restrict the applicability of the algorithm.

The theoretical minimum value of the functional, attained for $r = q$, equals $\Delta q^2 = 2N_{AB}$,[22] where N_{AB} is the total number of transitions from state A to B. Thus, if during RC optimization $\Delta r^2/2$ reaches N_{AB}, it follows that the putative RC equals the committor. During optimization of an RC for a finitely sampled system it is possible to obtain $\Delta r^2/2 < N_{AB}$, i.e., a value of the functional that is lower than the theoretical lower bound. In this

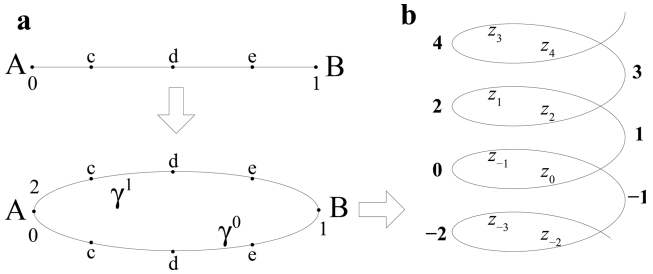


Figure 3. a) Two identical copies of the RC are joined into a circle in order to eliminate boundaries without modifying the RC. The idea is similar in spirit to the method of images in electrostatics. A and B denote boundary states, while c, d, e denote some states on the RC. Numbers show the values of the RC for the boundary states A and B. b) Schematic representation of the multivalued character of the optimal RC. Its value increments by 2 every time one goes around the full circle. It is analogous to a multivalued angle, whose value increments by 2π, i.e., $z \sim \phi/\pi$.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

case we say that the putative RC *overfits* the trajectory. Because of the usage of many fitting parameters the RC starts to approximate the statistical noise due to limited sampling rather than the actual dynamics.

A major weakness of these approaches is that it is difficult to suggest a good FF approximating the RC. The difficulty becomes apparent if one remembers that such an RC should be able to accurately project a few million snapshots of a very high-dimensional trajectory. In particular, it implies an extensive knowledge of the system, and that such a FF is likely to be system specific.

Recently we have suggested a nonparametric approach, which bypasses the difficult problem of finding an appropriate FF.[12] Since Δr^2 depends explicitly only on the RC time-series $r(k\Delta t_0)$, one may directly optimize the values of $r(k\Delta t_0)$ rather than the parameters α_i of the FF $R(\boldsymbol{X}, \alpha_i)$.

However, $r(k\Delta t_0)$ values cannot be varied independently of each other,[12] because points close in the original multidimensional space should have close projections ($R(\boldsymbol{X})$ is a continuous function), i.e., if $\boldsymbol{X}(i\Delta t_0) \sim \boldsymbol{X}(j\Delta t_0)$, then $r(i\Delta t_0) \sim r(j\Delta t_0)$. To vary $r(k\Delta t_0)$ in an appropriate, concerted way, one improves $r(k\Delta t_0)$ in the following iterative manner: $r'(k\Delta t_0) = f(r(k\Delta t_0), y(k\Delta t_0))$, where $r'(k\Delta t_0)$ is the updated values of the RC time-series, $y(k\Delta t_0)$ is the time-series of a randomly chosen coordinate of the original multidimensional space $\boldsymbol{X}$, and $f(x, y)$ is a low degree polynomial. If the system obeys some symmetry (e.g., the rotational and translational symmetries for biomolecules), then the RC should obey the same symmetry. A simple way to ensure this is to use as $y(k\Delta t)$ collective variables that respect the symmetry. For analysis of protein dynamics, one can use distance time-series between randomly chosen pairs of atoms. Another possibility is to use time-series of sin or cos of a randomly selected dihedral angle. The flowchart of the algorithm is provided in Figure 7 in the Appendix.

The idea has been successfully tested on an extensively sampled 50 dimensional model system, with a trajectory of 10^6 steps containing 989 transitions (N_{AB}).[12] After 9100 iterations $\Delta r^2/2$ reached 988.9. Continuation of the optimization for 100000 iterations in total insignificantly decreased $\Delta r^2/2$ to 986.5, indicating that no notable overfitting is possible and that the putative RC should be very close to the committor, which was confirmed by applying the committor validation tests.[22,27] The optimization has improved the seed RC, even though the difference between the corresponding free energy profiles at the top of the TS was only 0.05 kT, i.e., the approach is very sensitive.

However, it is not always possible in practice to perform an extensive sampling of a system of interest. A typical example is the simulation of protein folding, where it is very computationally expensive to obtain a handful of folding−unfolding events.[1,2] In this case the direct application of the simple approach described above leads to $\Delta r^2/2 \ll N_{AB}$, which is an indication of severe overfitting. Here we describe the approach, which allowed us to determine the optimal RC or committor for a typical realistic system of interest, namely the atomistic folding trajectory of a double mutant of HP35.

Briefly, the idea is as follows. As we show below, an RC is likely to be optimized in a nonuniform manner: it is easier to optimize TSs rather than free energy minima. Consequently, some parts of the RC may be optimized more than others. For an extensively sampled system, where overfitting is not possible, this does not present a problem. As some, more optimized parts of the RC reach their optimal values, they cannot be improved further, and the only way to decrease the functional is to optimize the suboptimal parts of the RC. For a system with limited sampling, where overfitting is possible, one needs to find a way to make the optimization uniform and stop it as soon as all the parts of the RC are optimal. Using the protein folding trajectory as an example, we first describe how one can detect the time scales and the regions of the putative RC which are most suboptimal. Then we describe how one can improve uniformity of optimization by focusing on these suboptimal regions and time scales.

2.2. Identification of Suboptimal Spatiotemporal Regions. Suboptimal spatiotemporal regions can be detected by using $Z_{C,1}(r, \Delta t)$ cut-profiles, an important quantity for RC analysis, which can be straightforwardly computed from RC time-series $r(k\Delta t_0)$ and whose properties we briefly summarize below (more details are provided in the Appendix).[5,22] The integral $\int Z_{C,1}(r, \Delta t)\, dr$ equals $\Delta r^2(\Delta t)/2$, hence $Z_{C,1}(r, \Delta t)$ can be interpreted as a position dependent density of the $\Delta r^2(\Delta t)/2$ functional. To optimize the entire RC one needs to minimize the average of $Z_{C,1}(r, \Delta t)$, and to optimize the RC in a particular region one needs to minimize $Z_{C,1}(r, \Delta t)$ in that region. For a suboptimal RC, $Z_{C,1}(r, \Delta t)$ values generally decrease to the limiting value of N_{AB}, as Δt increases. The larger the difference between $Z_{C,1}(r, \Delta t_1)$ and $Z_{C,1}(r, \Delta t_2)$ the less optimal the RC around r. For the optimal RC or the committor $Z_{C,1}(q, \Delta t) = N_{AB}$, which can be used as a committor validation test.[27] Thus, our aim is to determine an RC time-series $r(k\Delta t_0)$, such that $-\ln Z_{C,1}(r, \Delta t) \approx -\ln N_{AB}$ up to statistical uncertainty, roughly estimated as $1/\sqrt{2N_{AB}}$.

We consider a long equilibrium folding−unfolding trajectory of HP35 Nle/Nle double mutant consisting of 1509392 snapshots at 380 K.[28] The boundary states are defined using rather stringent criteria to ensure that only the configurations from the respective basins are obtained: node B in the native state is defined by the C_α root-mean-square deviation (rmsd) from the native 2f4k pdb structure[29] smaller than 1.0 Å, and node A in the denatured state is defined by the C_α rmsd greater than 10.5 Å (Figure 1a). The total number of transitions between these nodes, determined from the trajectory, is $N_{AB} = 73$.

Figure 1 shows the free energy profiles and $Z_{C,1}$ profiles (the committor validation test) for two popular conventional RCs, the rmsd and the fraction of native contacts (Q),[30] and compares them with the putative RC obtained with the nonparametric approach starting from the coordinate initialized to zero. $Z_{C,1}$ profiles show that the conventional RCs are far from being optimal and are worse than the putative RC already after 200 iterations. The FEP $F(r)$ (Figure 1e) shows the main transition state (TS) barrier separating the denatured and native states and that the native state contains two basins.

After 33700 iterations $\Delta r^2/2$ has reached the stopping value of N_{AB}. Figures 2a-b show the FEP and the $Z_{C,1}(r, \Delta t)$ profiles, respectively. As one can see $Z_{C,1}$ shows relatively large variations, especially in the regions around the free energy minima, i.e., the difference between $-\ln Z_{C,1}(r, \Delta t)$ and $-\ln N_{AB}$ is significantly larger than $1/\sqrt{2N_{AB}} \approx 0.08$. It means that while the putative RC is close to the committor (cf. Figure 1), it still deviates from it. This is due to the following reasons. First, variability of $Z_{C,1}(r, \Delta t_0)$ along r indicates that the RC is optimized in a nonuniform way. It is well optimized in the TS

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

region, where $Z_{C,1}(r, \Delta t_0)$ is constant, and much less so around the minima.

Consider an analytical equation for the committor along a single coordinate: $dq/dx \sim D(x) \exp[F(x)/kT]$. Assuming that $D(x)$ is relatively constant, the equation shows that regions with high free energy (barriers) get exponentially magnified compared to the regions with low free energy (minima). When one tries to update the RC using a low degree polynomial $r' = f(r, y)$, it is difficult to simultaneously update the entire RC on two very different scales. Increasing the polynomial degree might help; however, a more efficient solution is to update the segments with vastly different scales separately. For example one may use different low degree polynomials for different segments. Second, the fact that $-\ln Z_{C,1}(r, \Delta t_0)$ is higher than the other $-\ln Z_{C,1}(r, \Delta t)$ around the TS indicates that the latter are less optimized than the former, e.g., the RC is optimized in a temporally nonuniform way.

2.3. Improving Spatial Uniformity. To identify such less and more optimized spatiotemporal segments in an automatic way, i.e., without user intervention, we suggest the following procedure. During optimization the variability of $Z_{C,1}(r, \Delta t)$ is monitored to determine the regions of RC which are less optimized. Namely, the larger the difference between $-\ln Z_{C,1}(r, \Delta t')$ and $-\ln Z_{C,1}(r, \Delta t)$ for some $\Delta t' > \Delta t$ the less optimal is the coordinate in the region around r. One finds such $\Delta t' > \Delta t$, for which the nonuniformity of the distance between the profiles $\frac{\max_r \xi(r)}{\min_r \xi(r)}$ is the largest, where $\xi(r) = Z_{C,1}(r, \Delta t)/Z_{C,1}(r, \Delta t')$. Then segments where $\xi(r) > (1-0.02)\max_r \xi(r)$ are considered to be less optimized. The less and more optimized segments of the RC are updated using different polynomials of higher and lower degree, respectively. Here we used polynomials of fifth and second degree. A polynomial of lower, fourth, degree was not sufficiently flexible to improve suboptimal regions. One can also just update the less optimized segments while keeping the rest of the RC fixed. This simple procedure improves the spatial uniformity of optimization.

2.4. Improving Temporal Uniformity. Temporal uniformity can be improved by optimizing the RC with longer sampling intervals. However, one cannot simply optimize $\Delta r^2 = \sum_k [r(k\Delta t + \Delta t) - r(k\Delta t)]^2$, with, e.g., $\Delta t = 2\Delta t_0$. The optimal RC corresponding to $\Delta t > \Delta t_0$ differs from that corresponding to $\Delta t = \Delta t_0$.[22] An intuitive way to understand the difference is to note that when a trajectory is observed with a longer sampling interval, one may miss the events when the system visits a boundary node and quickly comes back, thus underestimating the probability to end up at the boundary. More formally, if the RC r_i satisfies eq 2a for $\Delta t = \Delta t_0$, then it is straightforward to show that r_i satisfies the same equation for $\Delta t = k\Delta t_0$, where $P_{ji}(k\Delta t) = P^k(\Delta t)_{ji}$. However, the equation is not satisfied by the boundary nodes, which satisfy eq 2b. In particular, eq 2a means that the average displacement from every point is zero, and for a boundary point it is not true: all points are either smaller or larger than it.

One way to overcome this problem is to eliminate boundaries without modifying the RC by joining two identical copies of the RC into a circle as shown in Figure 3a.[31,32] When at states A or B, the system can follow either of the RC copies with equal probability. Then the average displacement from points A and B is zero due to symmetry, i.e., eq 2a is valid for all points. The optimal RC (the solution of eq 1 or eq 2a) on a circle is a multivalued function.[31] For example, consider diffusion on a circle with constant $U(\phi)$ and $D(\phi)$, where ϕ is the angle. Eq 1 reads $\partial^2 q/\partial\phi^2 = 0$, with the solution $q = \phi/\pi$, meaning q is multivalued similar to ϕ: after making the full circle q is incremented by 2 (2 RCs of length 1).

In practice this multivalued RC on a circle, denoted by z (Figure 3b), is constructed from single valued RC r as follows. Consider a circle with a perimeter of 2 (radius $1/\pi$) with coordinate z along the perimeter (Figure 3a). States A and B correspond to points with angle $\phi = \pi$ and $\phi = 0$, correspondingly (Figure 3a). The points divide the circle into two segments: lower half γ^0 for $\phi \in [-\pi: 0]$, where $z(r) = r$ and upper half γ^1 for $\phi \in [0: \pi]$, where $z(r) = 2 - r$. These points, correspondingly, divide the multivalued RC into different segments or branches z_m (Figure 3b). For each segment z_m one has the following correspondence: $z_m(r) = m + r$ for even $m = 2l$ i.e., for segments on γ^0, and $z_m(r) = m + 1 - r$ for odd $m = 2l + 1$ i.e., for segments on γ^1. The segment number time-series $m(k\Delta t_0)$ is determined from seed RC time-series $r(k\Delta t_0)$ as follows. Whenever the trajectory visits a boundary node, it selects with equal probability which of the two adjoint segments it will follow. For example, when the trajectory, currently on segment z_0, visits node B, it selects with equal probability z_0 or z_1, if it visits node A, it selects between z_0 and z_{-1}. Once determined, $m(k\Delta t_0)$ are kept fixed during optimization (the boundary states snapshots do note move), only $r(k\Delta t_0)$ can change. Such constructed multivalued function z satisfies eq 2a for any value of Δt, and thus one can optimize $\Delta z^2(\Delta t) = \sum_k [z_{m(k\Delta t+\Delta t)}(r(k\Delta t + \Delta t)) - z_{m(k\Delta t)}(r(k\Delta t))]^2$ for any value of Δt. Note that $\Delta z^2(\Delta t_0) = \Delta r^2(\Delta t_0)$.

2.5. Adaptive Nonparametric Optimization. The ideas described above are combined into a simple optimization algorithm. Starting with RC initialized to zero, one iteratively improves the RC by nonparametrically minimizing $\Delta z^2(\Delta t)$, using polynomials of fifth and second degrees for less and more optimized segments of the RC, respectively. Every 5 iterations $Z_{C,1}$ profiles are scanned to identify these segments. The sampling interval Δt is changed randomly every 50 iterations as $\Delta t = 2^{[10\eta]}\Delta t_0$, where η is a uniformly distributed random number and $[\cdots]$ denotes an integer part. For $\Delta t > \Delta t_0$ optimization continues while $\Delta z^2(\Delta t)/2 > 1.15N_{AB}$, while for Δt_0 it continues while $\Delta z^2(\Delta t_0)/2 > N_{AB}$. The flowchart of the algorithm is provided in Figure 8 in the Appendix. Figures 2c-d show the results obtained with the approach (cf. Figure 1). The variability of $\ln Z_{C,1}$ is uniformly decreased; it is roughly bounded by ± 0.08, which means that we are close to the inherent statistical noise, and further improvement of the results makes little sense.

3. RESULTS

3.1. The FEP as a Function of the Optimal RC. Using the committor for the analysis and description of the folding dynamics may not be very convenient as the diffusion coefficient varies significantly along the coordinate $D(q) = J_{AB}/P_{eq}(q) = Z_H(q)^{-1}N_{AB}/\Delta t$, where $P_{eq}(q)$ is the equilibrium probability or $Z_H(q)$ is the corresponding histogram density computed from $q(k\Delta t_0)$.[5,19,20,22] It is more convenient to use a "natural" coordinate, which we denote as $\tilde{q}$, where the diffusion coefficient is constant $D(\tilde{q}) = 1$ and that is related to the committor by the following monotonous transformation $d\tilde{q}/dq = D(q)^{-1/2}$.[24] Since the transformation is monotonous, the new coordinate is as good as the committor for the description of the dynamics. Figure 4 shows the free energy profile $F(\tilde{q})$ as a function of $\tilde{q}$. Note that $D(\tilde{q}) = 1$ in units where time is measured in timesteps of 0.2 ns.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

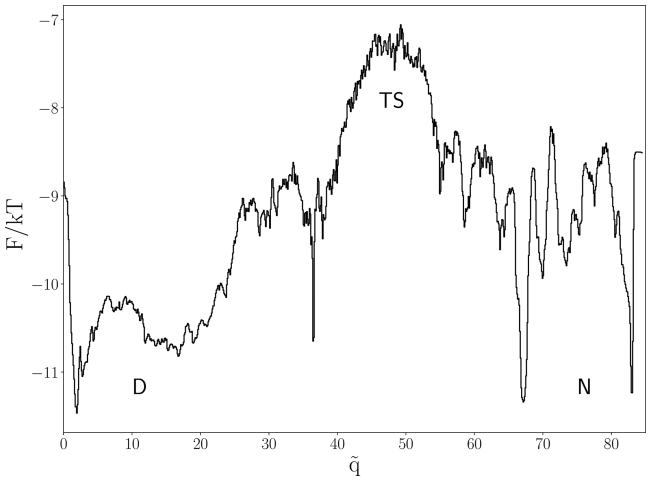


Figure 4. FEP of HP35: $F(\tilde{q})$, where $\tilde{q}$ is the optimal RC or the committor rescaled so that the diffusion coefficient is constant, $D(\tilde{q}) = 1$.

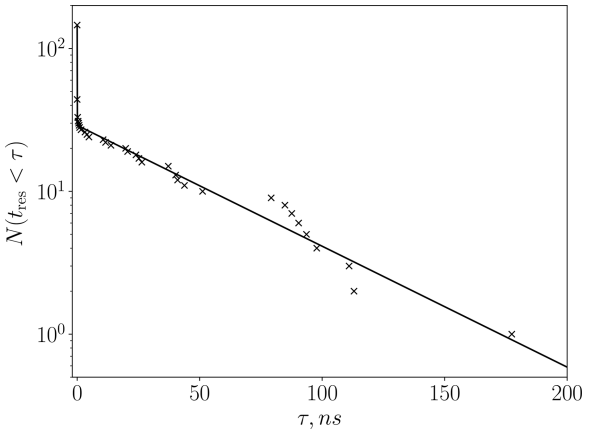


Figure 5. Cumulative distribution of residence times t_{res} in the basin $q = 0.863$ for different transition paths (symbols). The distribution can be accurately approximated by a sum of two exponents with two very different time scales $117e^{-\tau/0.00778} + 29e^{-\tau/51.3}$ (line), which suggests two different pathways.

The free energy profile $F(\tilde{q})$ is relatively smooth in the denatured state (D) and the TS, while the native state (N) has many deep minima and high barriers. It is consistent with experimental observations that the native state has many conformational substates.[33] The substates differ structurally only locally. The high barriers are likely due to the compact structure of the native state: to perform any local conformational change, the protein needs to partially unfold first.

A single reaction coordinate does not show multiple pathways explicitly. However, if free energy basins that belong to different pathways do not overlap, they can serve as

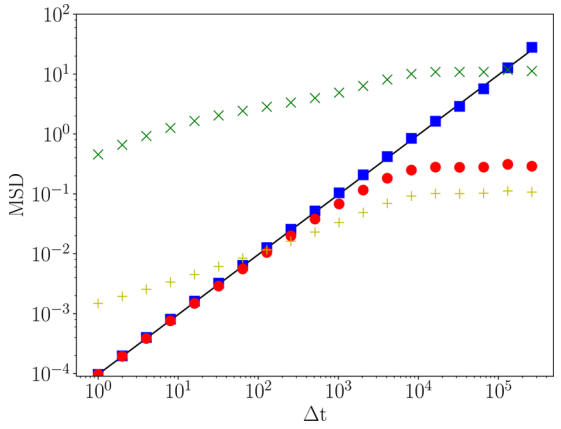


Figure 6. MSD of various RCs as a function of time: $\langle \Delta z^2 \rangle$ is shown by blue squares, $\langle \Delta q^2 \rangle$ is shown by red circles, the MSD of the fraction of native contacts is shown by yellow crosses, and the MSD of the C_α rmsd from the native structure is shown by green x's. The line shows the diffusive dependence $2N_{AB}\Delta t/(N\Delta t_0)$. For small Δt, when the system does not yet feel the boundaries, $\langle \Delta q^2 \rangle \sim \Delta t$.

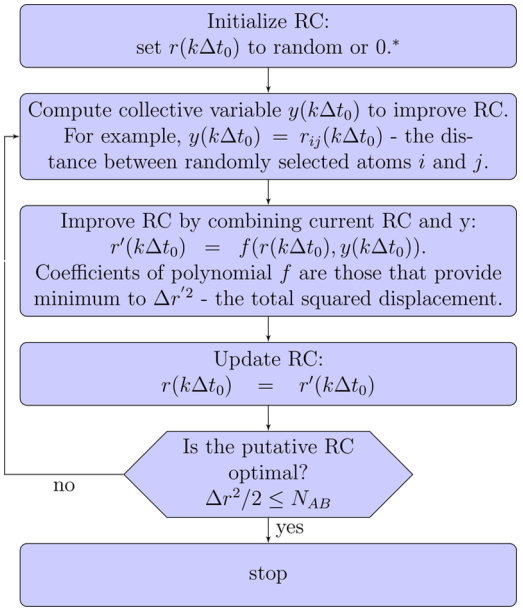


Figure 7. Flowchart outlines the nonparametric RC optimization algorithm. The algorithm computes the putative RC time-series $r(k\Delta t_0)$. * points that belong to the boundary states A and B are initialized to 0 and 1, respectively. They do not change during RC optimization.

fingerprints to distinguish different pathways. For example, there is a clear separation between distributions of residence times on each transition path in the deep basin defined by $|q-0.863| < 0.0002$ (Figure 5). About 29 of a total of 146 pathways

Table 1. Comparison of the Dynamical Quantities Computed from the Diffusive Model and Directly from the Trajectory[a]

$\tilde{q}(a)$	$\tilde{q}(b)$	N_{ab}	mfpt_{ab}	mfpt_{ba}	mtpt_{ab}
0	84.5	73 (0.1%)	3034 (0.1%)	1101 (−0.1%)	234 (−4%)
1.7	83	75 (2%)	3032 (−3%)	1102 (−2%)	208 (−8%)
17	68	89 (3%)	2547 (−3%)	962 (−3%)	66 (−13%)
36.5	58.6	115 (8%)	2072 (−7%)	750 (−7%)	10.7 (−1%)
38	55	127 (12%)	1959 (−11%)	712 (−11%)	7.8 (−13%)

[a]The numbers show the latter, while percentages in the brackets show the relative difference between the two. Times are given in ns.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

belong to the second distribution with much longer mean residence times. It suggests that this basin is an intermediate state, that belongs to a minor pathway.

3.2. The Diffusive Model Reproduces Important Dynamical Quantities. The profile $F(q)$ with the diffusion coefficient $D(q)$ or $F(\tilde{q})$ and $D(\tilde{q})$ define a diffusive model of the dynamics projected on the committor. According to the theory this model should provide a rather accurate description of folding dynamics of the protein.[5,19,20,22] In particular, the following important dynamical quantities can be computed exactly: the equilibrium flux $J_{AB} = N_{AB}/(N\Delta t_0)$ where N is the total number of trajectory snapshots, the mean first passage times (mfpt), and the mean transition path times (mtpt). Also, the mean squared displacement grows linearly with time as for simple diffusive motion.

The quantities were computed as[5]

$$\frac{1}{N_{AB}} = \int_{\tilde{q}(A)}^{\tilde{q}(B)} \frac{d\tilde{q}}{Z_{C,1}(\tilde{q})} = \int_{\tilde{q}(A)}^{\tilde{q}(B)} \frac{d\tilde{q}}{e^{-F(\tilde{q})} D(\tilde{q})\Delta t_0} \tag{3}$$

$$\mathrm{mfpt}_{AB} = \frac{N\Delta t_0}{N_{AB}} \int_0^1 dq P_{eq}(q)(1-q) = \langle 1-q\rangle / J_{AB} \tag{4}$$

$$\mathrm{mtpt}_{AB} = \frac{N\Delta t_0}{N_{AB}} \int_0^1 dq P_{eq}(q) q(1-q) = \langle q(1-q)\rangle / J_{AB} \tag{5}$$

By selecting two points a and b ($a < b$) on the committor q, one can define two new boundary states: A' contains all the points with $q < a$, and B' contains all the points with $b < q$. The optimal RC for the new boundary states can be obtained by simple rescaling of the original RC:[5] $q' = (q-a)/(b-a)$ for $a < q < b$; $q' = 0$ for $q < a$; and $q' = 1$ for $b < q$. Hence, the equilibrium flux, the mfpt, and the mtpt can be computed exactly between any two such states.

Table 1 compares these quantities computed from the diffusive model using eqs 3−5 and directly from the trajectory for different boundary states along the RC. We selected the original boundary states and free energy minima. The relative differences are around the expected statistical error for q of 8%. The differences can be reduced even further, if one removes non-negligible systematic bias due to the finite value of the trajectory sampling interval Δt_0 (see Appendix, Table 3).

Figure 6 shows the mean squared displacement (MSD) as a function of time computed for different RCs. The MSD of the multivalued optimal RC (z) follows a simple diffusive law, i.e., it grows linearly with time $\langle \Delta z^2 \rangle = 2N_{AB}\Delta t/(N\Delta t_0)$. The MSD of q follows that of z for small Δt, when the system does not yet feel the boundaries, and approaches the limiting value for large Δt. The MSD of the other two popular suboptimal RCs, the fraction of native contacts and the C_α rmsd from the native structure, shows subdiffusive behavior: $\langle \Delta r^2 \rangle \sim \Delta t^\alpha$ with $\alpha \sim 0.45$ and $\alpha \sim 0.35$, respectively. This illustrates that one of the reasons that the dynamics of various protein degrees of freedom is subdiffusive is because these degrees are not optimal RCs.[5,22,34−36]

3.3. The Pre-Exponential Factor. A fundamental problem in the analysis of protein folding dynamics, and an active area of research, is the determination of the free energy barrier ΔF and the pre-exponential factor k_0, which are related to the folding rate as $k_f = k_0 e^{-\Delta F/kT}$. *Direct* determination of these quantities from experiment has been hampered by very limited spatial and temporal resolution. The situation has significantly improved recently[37−40] e.g., one can now directly estimate the transition path times by counting single photons. However, the interpretation of the experiments still assumes a particular shape of the folding free energy landscape, which cannot be established in a direct manner. In this sense, the following quote from Yang and Gruebele[41] summarizes the experimental situation: "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.".

Having determined the optimal RC $\tilde{q}$ and the corresponding FEP $F(\tilde{q})$, which provide a quantitative description of the folding dynamics, we are now in a position to rigorously determine these quantities in a direct manner. We first note that to uniquely determine the height of the barrier, which is not invariant to monotonous transformations of RC (compare $F(q)$ with $F(\tilde{q})$), one needs to impose the constraint that the diffusion coefficient is constant. We estimate k_0 in three different ways. Taking the folding barrier of $4kT$ (Figure 4) and $k_f^{-1} = \tau_f = 3054$ ns (Table 1) one finds $k_0^{-1} = 55$ ns.

Applying the harmonic approximation to the Kramer's equation for the mfpt and assuming that the curvature at the denatured state and the unfolding basin are approximately equal, one can derive the following estimate $k_0^{-1} = 2\pi\tau_{corr}$, where $\tau_{corr} = kT/(D\omega)^2$ is the autocorrelation decay time at the TS.[42] The TS is approximated by a parabola with $(\omega^2/2)/kT = 0.023$, leading to $k_0^{-1} = 27$ ns.

Assuming diffusive dynamics over the parabolic TS (with the barrier height over $2kT$), Szabo derived the following relation between the mtpt and k_0[43]

$$\mathrm{mtpt} = (2\pi k_0)^{-1} \ln[2e^{\gamma} \ln(k_0\tau_f)] \tag{6}$$

where $\gamma = 0.577$ is Euler's constant. Taking points with $\tilde{q} = 38$ and $\tilde{q} = 55$ or $\tilde{q} = 36.5$ and $\tilde{q} = 58.6$ as boundaries for computing the mtpt (Table 1) one finds $k_0^{-1} = 18$ ns or $k_0^{-1} = 24$ ns, respectively. For boundaries at $\tilde{q} = 17$ and $\tilde{q} = 68$ or $\tilde{q} = 1.7$ and $\tilde{q} = 83$ one finds $k_0^{-1} = 186$ ns or $k_0^{-1} = 888$ ns, respectively. As one notes the estimate strongly depends on the choice of the boundaries.

We argue that the correct choice of boundaries is $\tilde{q} = 38$ and $\tilde{q} = 55$ or $\tilde{q} = 36.5$ and $\tilde{q} = 58.6$. Inside these boundaries the free energy profile is (approximately) parabolic, and thus the assumptions used to derive eq 6 are satisfied. While the boundaries are closer to the TS than to the minima of the denatured and native states, the mfpt between them captures 65% of the folding time. In other words, eq 6 is rather accurate, if applied to the parabolic part of the TS.

FEPs along suboptimal RCs are rather smooth with no apparent barriers in the native basin (see, e.g., Figures 1c and e), which may lead one to the erroneous conclusions that an estimate based on eq 6 is valid for boundaries taken far from the TS, e.g., at the bottoms of the basins. The diffusive model cannot be used for a quantitative description of the dynamics projected on such an RC, because the projected dynamics is subdiffusive. For example, the difference between the mtpt computed from the diffusive model and from the trajectory for the number of native contacts RC (Figure 1c) is 1145%.

4. CONCLUDING DISCUSSION

We have presented an approach to determine the optimal RC or committor for realistic systems with limited sampling. The approach is nonparametric and can approximate any RC with

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

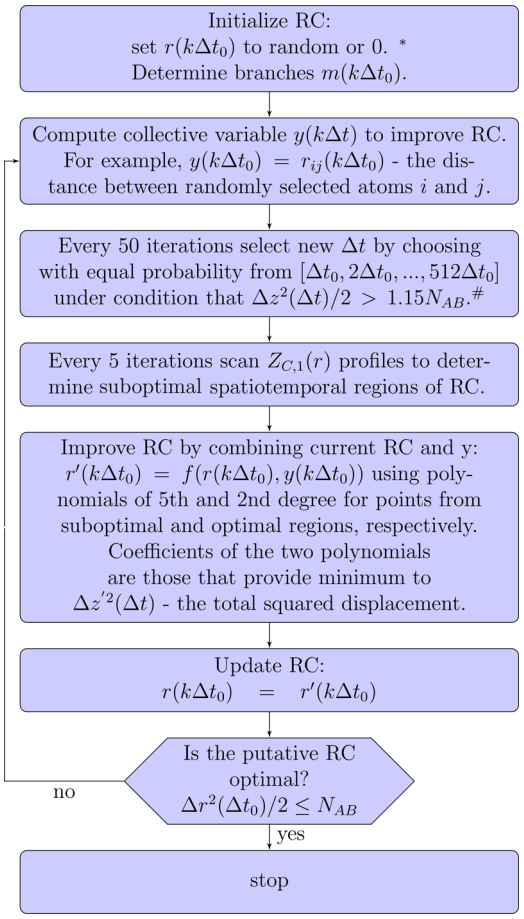


Figure 8. Flowchart outlines the adaptive nonparametric RC optimization algorithm. The algorithm computes the putative RC time-series $r(k\Delta t_0)$. * points that belong to the boundary states A and B are initialized to 0 and 1, respectively. They do not change during RC optimization. # for $\Delta t = \Delta t_0$ the condition is $\Delta z^2(\Delta t_0)/2 > N_{AB}$.

high accuracy. It can be readily applied to any system, avoiding prior analyses required to suggest a good system-specific functional form approximating the RC. In order to optimize the RC in a uniform manner we introduced adaptive optimization over different spatiotemporal regions, which required the introduction of a multivalued RC.

The approach was applied to the equilibrium folding trajectory of the HP35 double mutant. The determined RC closely approximates the committor as was validated by the optimality criterion – $Z_{C,1}$ is constant up to the expected statistical noise. We have demonstrated that important dynamical properties - the equilibrium flux, the mean first passage times, and the mean transition path times between any two regions on the RC can be computed exactly, up to statistical uncertainty. The mean squared displacement of the optimal RC grows linearly with time as for simple diffusion. We emphasize that no fitting of the parameters of the diffusive models was employed and that an accurate description is achieved at the trajectory time scale of 0.2 ns. Using this RC we obtained a direct rigorous estimate for the pre-exponential factor of $k_0^{-1} \sim 30$ ns.

To determine the optimal RC one needs to specify the boundary states A and B, which is often done by using order parameters, e.g., the root-mean-square deviation from the native structure here. However, such a definition may lead to poor results in more complex systems, where conventional order parameters may not be sensitive enough. One possibility to properly define boundary states in such systems is to use dominant eigenvectors, determined by the nonparametric approach.[12] While the process of nonparametric optimization of eigenvectors is not stable, its initial stable phase could be sufficient to define the boundary states.

The problem of finding an optimal RC for the description of complex dynamics is not unique to protein folding or molecular dynamics in general. Consider, for example, the problem of accurate description, monitoring and prognosis of disease dynamics. We assume that disease dynamics should be described stochastically, e.g., due to inherent randomness or coarse grained/incomplete description. In that case the best coordinate that describes the progress of the disease (the best biomarker) between two end states, e.g., healthy and abnormal, is the committor.[5,26] In particular, it should accurately predict the odds of positive outcome and the mean time to achieve that. The proposed approach can be used to construct such a coordinate from an ensemble of patient trajectories in an automated way without any disease specific information.

APPENDIX

Properties of $Z_{C,1}(r, \Delta t)$

Given a long RC time-series $r(k\Delta t_0)$, $Z_{C,1}(r', \Delta t)$ equals half the total length the trajectory makes, when it transits through a point r' on the RC

$$Z_{C,1}(r', \Delta t) = 1/2 \sum_k^{r'} |r(\Delta t + k\Delta t) - r(k\Delta t)| \quad (7)$$

where $\sum_k^{r'}$ *denotes the sum over such k when* r' *is between* $r(\Delta t + k\Delta t)$ *and* r $(k\Delta t)$. This quantity can be computed by considering every timestep $\Delta t = \Delta t_0$ of the time-series, every second timestep $\Delta t = 2\Delta t_0$, third, and so forth.

If the RC satisfies eq 2a, then $Z_{C,1}(r, \Delta t_0) = const$.[22] If RC satisfies 2a, then it satisfies 2a with $P_{ij}(2\Delta t_0) = \sum_k P_{ik}(\Delta t_0) P_{kj}(\Delta t_0)$ and $Z_{C,1}(r, 2\Delta t_0) = Z_{C,1}(r, \Delta t_0) = const$, and so forth.[22]

Boundary nodes satisfy eq 2b rather than eq 2a, and if transitions over r' visit boundary nodes, then $Z_{C,1}(r', \Delta t > \Delta t_0) \neq Z_{C,1}(r', \Delta t_0)$. To overcome this problem at the boundaries, a special counting method using the ensemble of transition path segments has been suggested which restores driftlessness at boundaries and makes $Z_{C,1}$ constant everywhere.[22] Alternatively one can combine two identical copies of the RC into a circle in order to eliminate boundaries,[31,32] as described in the main text (Figure 3). In this case the RC is a multivalued function, denoted as z. Below, for brevity, $Z_{C,1}(r', \Delta t)$ denotes the $Z_{C,1}$ profile as a function of the original RC r, while $Z_{C,1}(z')$ denotes a different $Z_{C,1}$ profile as a function of the multivalued RC z. Analogously eq 7, given time-series $z(k\Delta t_0)$, $Z_{C,1}(z', \Delta t)$, equals half the total length the trajectory makes, when it transits through a point z' on the RC

$$Z_{C,1}(z', \Delta t) = 1/2 \sum_k^{z'} |z(\Delta t + k\Delta t) - z(k\Delta t)| \quad (8)$$

where $\sum_k^{z'}$ *denotes the sum over such k when* z' *is between* $z(\Delta t + k\Delta t)$ *and* $z(k\Delta t)$. $Z_{C,1}(r', \Delta t)$ is obtained by summing up over all segments or branches $z_m(r')$ of the multivalued function z, i.e., by projecting z back to r:

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

Table 2. Comparison of the Dynamical Quantities Computed with the Sampling Intervals of $0.001\Delta t_0$ and Δt_0 by Simulating Diffusion on the Free Energy Profile[a]

$\tilde{q}(a)$	$\tilde{q}(b)$	N_{ab}	mfpt_{ab}	mfpt_{ba}	mtpt_{ab}
0	84.5	130958 (2%)	3284 (−2%)	1159 (−2%)	245 (−12%)
1.7	83	134511 (2%)	3205 (−3%)	1121 (−2%)	201 (−3%)
17	68	157507 (0.7%)	2705 (−3%)	990 (−3%)	64 (−7%)
36.5	58.6	206167 (2%)	2087 (−2%)	736 (−2%)	12 (−10%)
38	55	224917 (2%)	1910 (−2%)	677 (−2%)	8 (−13%)

[a]The numbers show the latter, while percentages in the brackets show the relative difference between the two. Times are given in ns.

$$Z_{C,1}(r', \Delta t) = \sum_m Z_{C,1}(z_m(r'), \Delta t) \quad (9)$$

For the optimal RC or the committor such computed $Z_{C,1}(r', \Delta t) = N_{AB}$ is constant, i.e., is independent of r' and Δt for Δt much less than the trajectory length.

In the limit of very small Δt, $Z_{C,1}(r, \Delta t) = Z_H(r)D(r)\Delta t$, where $Z_H(r) \sim e^{-F(r)/kT}$ is the density of trajectory points around r, F is the free energy, and D is the diffusion coefficient (5). The relation can be used to determine the diffusion coefficient for arbitrary RC. For very large Δt, $Z_{C,1}(r, \Delta t) = N_{AB}$. Since for the committor coordinate $Z_{C,1}(q, \Delta t)$ is constant for all Δt and thus is equal to N_{AB}, one can determine the diffusion coefficient along the committor as $D(q) = Z_H(q)^{-1}N_{AB}/\Delta t$.

Integrating eq 7 one obtains $\int Z_{C,1}(r, \Delta t)dr = \sum_k [r(k\Delta t + \Delta t) - r(k\Delta t)]^2/2$, hence $Z_{C,1}(r, \Delta t)$ can be considered as the local average density of $\Delta r^2/2$. Difference between two profiles $Z_{C,1}(r, 2\Delta t)$ and $Z_{C,1}(r, \Delta t)$, averaged over some local region, is proportional to $\Delta r^2(2\Delta t) - \Delta r^2(\Delta t) \sim \langle (r(t + \Delta t) - r(t))(r(t) - r(t - \Delta t))\rangle$, the correlation between successive displacements. Which means that the closer $Z_{C,1}(r, \Delta t)$ profiles for different Δt, the closer the correlation is to zero. For the optimal RC or the committor, $Z_{C,1}(q, \Delta t)$ is constant for all Δt, and the correlation is zero. If $Z_{C,1}(r, \Delta t) < Z_{C,1}(r, \Delta t_0)$ for $\Delta t > \Delta t_0$, then the correlation is negative and the dynamics is subdiffusive. The larger the difference, the more negative are correlations and the more suboptimal is the coordinate. If, alternatively, $Z_{C,1}(r, \Delta t) > Z_{C,1}(r, \Delta t_0)$, then the correlation is positive and the dynamics is superdiffusive.

Removing Systematic Bias Due to Finite Value of the Trajectory Sampling Interval Δt_0

Table 1 compares dynamical quantities determined using the diffusive model with that computed directly from the atomistic trajectory. Equations for the equilibrium flux, the mfpt, and the mtpt (eqs 3−5) were derived assuming that the dynamics is observed with infinitely high temporal resolution. In practice, however, the trajectory is saved with a finite sampling interval Δt_0, which means that some of the events, when the system quickly visits a boundary state and comes back, can be missed. This may lead to systematic underestimation of the number of transitions N_{AB} and overestimation of the mfpt and the mtpt. An accurate way, which removes this systematic bias, is to compare the estimates for the same value of Δt_0. One may either compare eqs 3−5 with the results obtained directly from trajectory in the limit of $\Delta t_0 \to 0$, or one may simulate diffusive dynamics on the free energy profile $F(\tilde{q})$ and determine the equilibrium flux, the mfpt, and the mtpt, when observed with the sampling interval of Δt_0. We followed the second option. Diffusive dynamics on the free energy profile $F(\tilde{q})$ was simulated using MC with diffusion coefficient $D(\tilde{q}) = 1$ and a timestep of $0.001\Delta t_0$. The simulation length was chosen to be much longer than the original trajectory, so that statistical errors are negligible. Table 2 compares the dynamical quantities computed with sampling intervals Δt_0 and $0.001\Delta t_0$. One can see that, indeed, the systematic differences due to the finite sampling interval are non-negligible and comparable to the differences shown in Table 1.

Table 3 compares the dynamical quantities computed by simulating diffusion on the free energy profile $F(\tilde{q})$ with that computed directly from the original atomistic trajectory, both with sampling interval Δt_0. The differences are now smaller in comparison to the corresponding differences in Table 1.

Table 3. Comparison of the Dynamical Quantities Computed from the Diffusive Model and Directly from the Trajectory, Both with the Sampling Interval Δt_0[a]

$\tilde{q}(a)$	$\tilde{q}(b)$	N_{ab}	mfpt_{ab}	mfpt_{ba}	mtpt_{ab}
0	84.5	73	3034 (−8%)	1101 (−5%)	234 (−5%)
1.7	83	75	3032 (6%)	1102 (−2%)	208 (3%)
17	68	89	2547 (−3%)	962 (−3%)	66 (4%)
36.5	58.6	115	2072 (−1%)	750 (2%)	10.7 (−13%)
38	55	127	1959 (2%)	712 (5%)	7.8 (−2%)

[a]The numbers show the latter, while percentages in the brackets show the relative difference between the two. Times are given in ns.

AUTHOR INFORMATION

Corresponding Author

*E-mail: s.krivov@leeds.ac.uk.

ORCID

Sergei V. Krivov: 0000-0002-3493-0068

Funding

The work has been partially supported by BBSRC Grant (No. BB/J016055/1).

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

I am grateful to David Shaw and his co-workers for making the folding trajectory available.

REFERENCES

(1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(2) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.

(3) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **2010**, *6*, 751−758.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427

(4) Schwantes, C. R.; Pande, V. S. Modeling Molecular Kinetics with tICA and the Kernel Trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.
(5) Banushkina, P. V.; Krivov, S. V. Optimal reaction coordinates. *WIREs Comput. Mol. Sci.* **2016**, *6*, 748−763.
(6) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141−147.
(7) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 14766−14770.
(8) Best, R. B.; Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6732−6737.
(9) Krivov, S. V. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. *J. Phys. Chem. B* **2011**, *115*, 12315−12324.
(10) Li, W.; Ma, A. Recent developments in methods for identifying reaction coordinates. *Mol. Simul.* **2014**, *40*, 784−793.
(11) Nüske, F.; Keller, B. G.; Pérez-Hernández, G.; Mey, A. S. J. S.; Noé, F. Variational Approach to Molecular Kinetics. *J. Chem. Theory Comput.* **2014**, *10*, 1739−1752.
(12) Banushkina, P. V.; Krivov, S. V. Nonparametric variational optimization of reaction coordinates. *J. Chem. Phys.* **2015**, *143*, 184108.
(13) Mu, Y.; Nguyen, P. H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 45−52.
(14) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.
(15) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2839−2844.
(16) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5−30.
(17) Peters, B.; Trout, B. L. Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys.* **2006**, *125*, 054108.
(18) Peters, B. Reaction Coordinates and Mechanistic Hypothesis Tests. *Annu. Rev. Phys. Chem.* **2016**, *67*, 669−690.
(19) Lu, J.; Vanden-Eijnden, E. Exact dynamical coarse-graining without time-scale separation. *J. Chem. Phys.* **2014**, *141*, 044109.
(20) Berezhkovskii, A. M.; Szabo, A. Diffusion along the Splitting/Commitment Probability Reaction Coordinate. *J. Phys. Chem. B* **2013**, *117*, 13115−13119.
(21) Metzner, P.; Schütte, C.; Vanden-Eijnden, E. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.* **2006**, *125*, 084110.
(22) Krivov, S. V. On Reaction Coordinate Optimality. *J. Chem. Theory Comput.* **2013**, *9*, 135−146.
(23) Hummer, G. From transition paths to transition states and rate coefficients. *J. Chem. Phys.* **2004**, *120*, 516−523.
(24) Krivov, S. V.; Karplus, M. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13841−13846.
(25) Krivov, S. V. Numerical Construction of the pfold (Committor) Reaction Coordinate for a Markov Process. *J. Phys. Chem. B* **2011**, *115*, 11382−11388.
(26) Krivov, S. V.; Fenton, H.; Goldsmith, P. J.; Prasad, R. K.; Fisher, J.; Paci, E. Optimal Reaction Coordinate as a Biomarker for the Dynamics of Recovery from Kidney Transplant. *PLoS Comput. Biol.* **2014**, *10*, e1003685.
(27) Banushkina, P. V.; Krivov, S. V. Fep1d: A script for the analysis of reaction coordinates. *J. Comput. Chem.* **2015**, *36*, 878−882.
(28) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17845−17850.
(29) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-Microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359*, 546−553.
(30) Best, R. B.; Hummer, G.; Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17874−17879.
(31) Krivov, S. V. Method to describe stochastic dynamics using an optimal coordinate. *Phys. Rev. E* **2013**, *88*, 062131.
(32) Tian, P.; Jónsson, S. Æ.; Ferkinghoff-Borg, J.; Krivov, S. V.; Lindorff-Larsen, K.; Irbäck, A.; Boomsma, W. Robust Estimation of Diffusion-Optimized Ensembles for Enhanced Sampling. *J. Chem. Theory Comput.* **2014**, *10*, 543−553.
(33) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **1991**, *254*, 1598−1603.
(34) Krivov, S. V. Is Protein Folding Sub-Diffusive? *PLoS Comput. Biol.* **2010**, *6*, e1000921.
(35) Cote, Y.; Senet, P.; Delarue, P.; Maisuradze, G. G.; Scheraga, H. A. Nonexponential decay of internal rotational correlation functions of native proteins and self-similar structural fluctuations. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 19844−19849.
(36) Hu, X.; Hong, L.; Dean Smith, M.; Neusius, T.; Cheng, X.; Smith, J. C. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nat. Phys.* **2016**, *12*, 171−174.
(37) Chung, H. S.; McHale, K.; Louis, J. M.; Eaton, W. A. Single-Molecule Fluorescence Experiments Determine Protein Folding Transition Path Times. *Science* **2012**, *335*, 981−984.
(38) Chung, H. S.; Piana-Agostinetti, S.; Shaw, D. E.; Eaton, W. A. Structural origin of slow diffusion in protein folding. *Science* **2015**, *349*, 1504−1510.
(39) Neupane, K.; Foster, D. A. N.; Dee, D. R.; Yu, H.; Wang, F.; Woodside, M. T. Direct observation of transition paths during the folding of proteins and nucleic acids. *Science* **2016**, *352*, 239−242.
(40) Wirth, A. J.; Liu, Y.; Prigozhin, M. B.; Schulten, K.; Gruebele, M. Comparing Fast Pressure Jump and Temperature Jump Protein Folding Experiments and Simulations. *J. Am. Chem. Soc.* **2015**, *137*, 7152−7159.
(41) Yang, W. Y.; Gruebele, M. Folding at the speed limit. *Nature* **2003**, *423*, 193−197.
(42) Banushkina, P. V.; Krivov, S. V. High-Resolution Free-Energy Landscape Analysis of α-Helical Protein Folding: HP35 and Its Double Mutant. *J. Chem. Theory Comput.* **2013**, *9*, 5257−5266.
(43) Chung, H. S.; Louis, J. M.; Eaton, W. A. Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 11837−11844.

DOI: 10.1021/acs.jctc.8b00101
J. Chem. Theory Comput. 2018, 14, 3418−3427