Obtaining reaction coordinates by likelihood maximization

Baron Peters and Bernhardt L. Trout[a)]
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

(Received 3 May 2006; accepted 27 June 2006; published online 4 August 2006)

We present a new approach for calculating reaction coordinates in complex systems. The new method is based on transition path sampling and likelihood maximization. It requires fewer trajectories than a single iteration of existing procedures, and it applies to both low and high friction dynamics. The new method screens a set of candidate collective variables for a good reaction coordinate that depends on a few relevant variables. The Bayesian information criterion determines whether additional variables significantly improve the reaction coordinate. Additionally, we present an advantageous transition path sampling algorithm and an algorithm to generate the most likely transition path in the space of collective variables. The method is demonstrated on two systems: a bistable model potential energy surface and nucleation in the Ising model. For the Ising model of nucleation, we quantify for the first time the role of nuclei surface area in the nucleation reaction coordinate. Surprisingly, increased surface area increases the stability of nuclei in two dimensions but decreases nuclei stability in three dimensions. © *2006 American Institute of Physics*.
[DOI: 10.1063/1.2234477]

INTRODUCTION

The reaction coordinate is a single variable that quantifies progress along a reaction pathway. Knowing the reaction coordinate is essential for understanding how a reaction proceeds, but reaction coordinates are often extremely difficult to find. For all but the simplest systems, the reaction coordinate involves many degrees of freedom that are not easily identified. The major challenges are to determine which degrees of freedom are important and how they participate in the reaction coordinate. This paper addresses both of these challenges.

For any reaction, the exact reaction coordinate is the *committor probability*,[1–3] the fraction of trajectories initiated with Boltzmann distributed momenta from an atomic configuration $\mathbf{x}$ that commit to the product basin (B).[4–6] If the committor probability is denoted $p_B(\mathbf{x})$, then *transition states* are configurations for which $p_B(\mathbf{x})=1/2$, reactant configurations have $p_B(\mathbf{x})<1/2$, and product configurations have $p_B(\mathbf{x})>1/2$.[4–6] Unfortunately, $p_B(\mathbf{x})$ is costly to compute, and it provides no insight into the physical characteristics that distinguish reactants, products, and transition states.

For complex systems, a simple approximation to $p_B(\mathbf{x})$ in terms of collective variables is more useful and more feasible than the actual function $p_B(\mathbf{x})$. *Collective variables* are functions of the configuration that compress many atomistic details into physically important variables. Examples include the fraction of native protein contacts,[2] coordination numbers,[7] coordination geometries,[8] and nucleus size.[9] The key challenge is to learn which collective variables are important and how they are involved in the reaction coordinate.

Committor probabilities, reaction coordinates, transition states, and collective variables are important concepts used throughout this paper. Collective variables are denoted $q(\mathbf{x})$, where $\mathbf{x}$ is the full configuration of the system. $\mathbf{q}(\mathbf{x})$ denotes a vector of several collective variables. Reaction coordinates are denoted $r(\mathbf{q})$ or $r(\mathbf{x})$, both of which are abbreviations for $r(\mathbf{q}(\mathbf{x}))$. Where the reaction coordinate is written as r without an explicit dependence on $\mathbf{q}$ or $\mathbf{x}$, the meaning is "a particular value of $r(\mathbf{x})$." The symbol $r^{\dagger}$ indicates the value of $r(\mathbf{x})$ corresponding to the transition state surface. Note that $r(\mathbf{x})$ is also used to indicate a trial reaction coordinate.

Because the committor probability is the exact reaction coordinate, isosurfaces of a good reaction coordinate, $r(\mathbf{x})=r$, must closely approximate isocommittor surfaces, $p_B(\mathbf{x})=\text{const}$. The usual test for matching isosurfaces is to sample configurations from the Boltzmann distribution on an isosurface of the trial reaction coordinate. These configurations are then used to construct a histogram of estimated $p_B(\mathbf{x})$ values.[5,6] For brevity, a histogram of estimated p_B values will be called a p_B histogram for the remainder of this paper. If $r(\mathbf{x})$ is a good reaction coordinate, the p_B histogram will be sharply peaked around a p_B value corresponding to the value of r.[5,6] Having a sharply peaked p_B histogram around $p_B=1/2$ for the putative transition state surface is particularly important.

In committor analysis, trial reaction coordinates are iteratively tested and improved based on p_B histograms. Each estimated p_B value in a histogram requires on the order of 100 trajectories, and good statistics require hundreds of estimates per p_B histogram. Figure 1 shows that a single p_B histogram can require tens of thousands of trajectories, each half as long as a reactive trajectory.

The difficulties and computational cost of committor analyses have motivated recent attempts to systematize the search for reaction coordinates.[1–3,10] Recent methods improve upon the trial and error aspects but continue to use

[a)]Author to whom correspondence should be addressed. Electronic mail: trout@mit.edu

 © 2006 American Institute of Physics

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

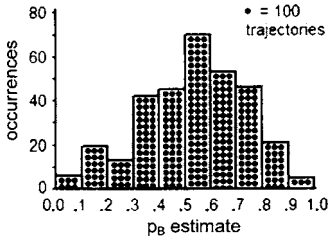

FIG. 1. This p_B histogram for a putative definition of critical nuclei in the Ising model required 32 000 trajectories, each half as long as a reactive trajectory.

expensive histogram calculations. Ma and Dinner developed a neural network scheme that iteratively proposes and improves trial reaction coordinates based on committor analysis.[1] Their procedure requires millions of trajectories to train the neural network and to evaluate the network's guesses.[1]

Hummer reformulated committor analysis in terms of the probability that trajectories initiated from $\mathbf{x}$ are transition paths, $p(\mathrm{TP}|\mathbf{x})$.[11] For diffusive barrier crossings, $p(\mathrm{TP}|\mathbf{x}) = 2p_B(\mathbf{x})(1-p_B(\mathbf{x}))$,[11] so $p(\mathrm{TP}|\mathbf{x})$ attains a maximal value of $1/2$ for $p_B(\mathbf{x})=1/2$, the transition state surface.[11] Projection of $p(\mathrm{TP}|\mathbf{x})$ onto a good reaction coordinate $r(\mathbf{x})$ gives a function $p(\mathrm{TP}|r)$ with the peak at $r^{\ddagger}$ corresponding to the transition state surface, $r(\mathbf{x})=r^{\ddagger}$.[11] Best and Hummer variationally maximized the peak in $p(\mathrm{TP}|r)$ to obtain a good reaction coordinate at the transition state surface.[2] Their method requires a histogram of estimated $p(\mathrm{TP}|r)$ values for each iterative improvement of $r(\mathbf{x})$.[2]

Maragliano *et al.*[10] developed a string method for collective variables to obtain a family of approximate isocommittor surfaces. To our knowledge, theirs is the only other approach that does not rely on costly histograms. However, their method presumes *a priori* knowledge of the relevant variables, and it requires many iterations of mean force and variable entanglement calculations. It is unclear how the method of Maragliano *et al.*[10] will compare to existing strategies.

Transition path sampling[5,6,12,13] (TPS) is a powerful importance sampling scheme for simulating reactive trajectories in complex systems. TPS efficiently calculates rates and reactive trajectories by focusing on rare reactive trajectories. In contrast, trajectories from a straightforward simulation spend the vast majority of time in the reactant or product basins. TPS and related path sampling algorithms[14–16] have been used to study ice nucleation,[17,18] DNA transcription,[19] DNA hybridization,[20] Grotthus proton transfer,[21] protein folding,[22–24] methionine oxidation,[25] transfer of molecules across lipid bilayers,[26] the phage-λ switch,[16] and heterogeneous catalysis.[27]

Each shooting point from TPS provides information about the reaction coordinate: where the trajectory was initiated and whether each end committed to the product state. However, points along the ensemble of transition paths are distributed as $p(\mathbf{x}|\mathrm{TP})$.[11] Points from the equilibrium distribution, $\rho_{\mathrm{eq}}(\mathbf{x})$, are distributed differently.[3,11] If $r(\mathbf{x})$ is an arbitrary trial reaction coordinate, then $p(\mathbf{x}|r,\mathrm{TP}) \neq \rho_{\mathrm{eq}}(\mathbf{x}|r)$.[3,11] Thus, methods that rely on histograms of $p_B(\mathbf{x})$ or $p(\mathrm{TP}|\mathbf{x})$ for a trial reaction coordinate isosurface cannot use the information from TPS.[11] However, E *et al.*[3] showed that good reaction coordinates satisfy $p(\mathbf{x}|r,\mathrm{TP}) = \rho_{\mathrm{eq}}(\mathbf{x}|r)$.

The observation of E *et al.* provides the central justification for the method proposed here. If the candidate variables permit a complete description of the true reaction coordinate and if the reaction coordinate is optimized using TPS shooting points from the full range of p_B values, then the reaction coordinate based on the transition path ensemble is also a good reaction coordinate in the equilibrium ensemble. The method is described below, including a new TPS algorithm that is specially designed for calculating reaction coordinates but simultaneously applicable to calculating rate constants. The final sections present applications of the new method to a model potential energy surface and to an Ising model of nucleation.

OVERVIEW OF METHOD

The approach presented here is very different from previous approaches. The typical approach is to propose a trial reaction coordinate, sample points on the corresponding constraint surfaces, and then test each ensemble of samples. Our approach samples points independent of the reaction coordinates to be tested, and then tests all reaction coordinate candidates for the best coordinate given the sample.

The new method begins by harvesting an ensemble of shooting points from a modified version of TPS. Each shooting point is saved with information on whether the trajectory was accepted or rejected and whether its end points committed to the reactant or product basin. Then specify a set of coordinates to be tested. For example, let $\mathbf{q}(\mathbf{x}) = (q_1(\mathbf{x}),\ldots,q_m(\mathbf{x}))$ be a set of m collective variables that are potentially important in the reaction coordinate. The collective variables need not have the same units, nor be differentiable. The collective variables are evaluated at each shooting point. This information, a sample of collective variables evaluated at shooting points and the corresponding trajectory fates, allows likelihood maximization to find the best reaction coordinate from each combination of collective variables. The combinations of variables are screened starting from single variables as model reaction coordinates and going to combinations of two, then three variables, etc. The search stops when the improvement is no longer significant according to the Bayesian information criterion.

"AIMLESS SHOOTING" ALGORITHM

Aimless shooting is version of transition path sampling where momenta are drawn fresh from the Boltzmann distribution for each trial trajectory. Each trajectory shot from $\mathbf{x}$ with freshly sampled momenta is an independent realization of $p(\mathrm{TP}|\mathbf{x})$. In contrast, successive trajectory outcomes are correlated when momenta are slightly perturbed from previous momenta.[6,12] Aimless shooting generates new reactive trajectories only when the shooting point is near the $p_B = 1/2$ surface. Thus, aimless shooting must also be designed to generate most shooting points near the unknown $p_B=1/2$ surface.

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

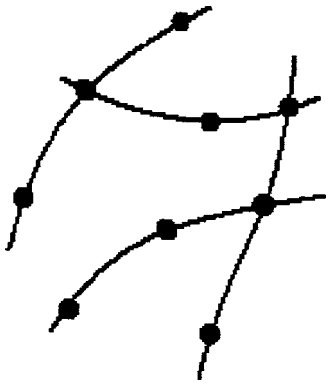

FIG. 2. Aimless shooting creates a sequence of interconnected trajectories. The three points on each trajectory are separated by a short time Δt.

Let Δt be much shorter than the duration of a reactive trajectory, t. The value of Δt has some effect on efficiency, but a wide range of values is acceptable. Define $\mathbf{x}_{-t/2}^{(o)}$ as the point at which the previously accepted trajectory began and $\mathbf{x}_{t/2}^{(o)}$ as the point at which the previously accepted trajectory ended. The superscript (o) denotes "old" and the superscript (n) denotes "new." Select a shooting point on the old trajectory from $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$ with equal probability for each position. Also shift the time t_0 at the shooting point along the new trajectory to $-\Delta t$, 0, or Δt. Denote the time-shifted shooting point on the new trajectory as $\mathbf{x}_{t_0}^{(n)}$, and draw new momenta from the Boltzmann distribution. Dynamically propagate the system to $\pm t/2$. Finally, accept the new trajectory if it joins the reactant and product states A and B. Figure 2 shows how aimless shooting produces new trajectories from old trajectories. Figure 3 shows why the shooting points from aimless shooting stay in regions with large $p(\mathrm{TP}|\mathbf{x})$.

ACCEPTANCE RULE FOR AIMLESS SHOOTING

If the initial velocities are $\mathbf{v}_{t_0}$, then the probability of generating a new trajectory $\{\mathbf{x}_{\pm t/2}^{(n)}\}$ from the old trajectory is

$$p_{\mathrm{gen}}^{o\to n}=\frac{1}{9}\rho_{\mathrm{eq}}(\mathbf{v}_{t_0})p(\{\mathbf{x}_{\pm t/2}^{(n)}\}|\mathbf{v}_{t_0},\mathbf{x}_{t_0}^{(n)},t_0). \tag{1}$$

The probability $p(\{\mathbf{x}_{\pm T/2}^{(n)}\}|\mathbf{v}_0,\mathbf{x}_{t_0}^{(n)},t_0)$ depends on the equations of motion, and ρ_{eq} is the Boltzmann distribution. The factor of 1/9 originates from choosing the shooting point from $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$ and choosing the temporal position of the shooting point on the new trajectory $\mathbf{x}_{t_0}^{(n)}$ with t_0

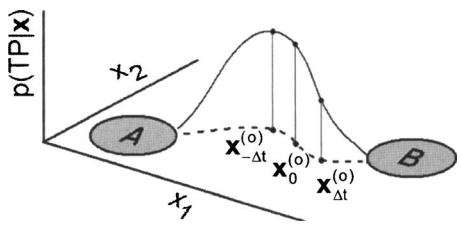

FIG. 3. Of the points $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$ on the old trajectory, the one with the largest $p(\mathrm{TP}|\mathbf{x})$ is most likely to yield a new reactive trajectory. Thus, the aimless shooting algorithm has a statistical "restoring force" that keeps shooting points near the maximum of $p(\mathrm{TP}|\mathbf{x})$.

$=-\Delta t$, 0, or Δt. If the equations of motion conserve the equilibrium distribution, then[5,28]

$$\rho_{\mathrm{eq}}(\mathbf{x}_{-t/2},\mathbf{v}_{-t/2})p(\{\mathbf{x}_{\pm t/2}\}|\mathbf{v}_{-t/2},\mathbf{x}_{-t/2},-t/2)=\rho_{\mathrm{eq}}(\mathbf{x}_{t_0},\mathbf{v}_{t_0})p(\{\mathbf{x}_{\pm t/2}\}|\mathbf{v}_{t_0},\mathbf{x}_{t_0},t_0). \tag{2}$$

The ratio of generation probabilities is

$$\frac{p_{\mathrm{gen}}^{o\to n}}{p_{\mathrm{gen}}^{n\to o}}=\frac{\rho_{\mathrm{eq}}(\mathbf{v}_{t_0}^{(n)})}{\rho_{\mathrm{eq}}(\mathbf{v}_{t_0}^{(o)})}\frac{p(\{\mathbf{x}_{\pm t/2}^{(n)}\}|\mathbf{x},\mathbf{v}_{t_0}^{(n)},t_0^{(n)})}{p(\{\mathbf{x}_{\pm t/2}^{(o)}\}|\mathbf{x},\mathbf{v}_{t_0}^{(o)},t_0^{(o)})}, \tag{3}$$

and the statistical weight[6,12] of a trajectory $\{\mathbf{x}_{\pm t/2}\}$ is

$$S[\{\mathbf{x}_{\pm t/2}\}]=h_A[\mathbf{x}_{-t/2}]h_B[\mathbf{x}_{t/2}]\rho_{\mathrm{eq}}(\mathbf{x}_{-t/2},\mathbf{y}_{-t/2})\times p(\{\mathbf{x}_{\pm t/2}\}|\mathbf{x}_{-t/2},\mathbf{v}_{-t/2},-t/2). \tag{4}$$

The function $h_A[\mathbf{x}]$ returns 1 if $\mathbf{x}$ is in A and 0 otherwise. Similarly, $h_B[\mathbf{x}]$ returns 1 if $\mathbf{x}$ is in B and 0 otherwise. Using Eq. (2), the ratio of statistical weights for the new and old paths is

$$\frac{S[\{\mathbf{x}_{\pm t/2}^{(n)}\}]}{S[\{\mathbf{x}_{\pm t/2}^{(o)}\}]}=h_A[\mathbf{x}_{-t/2}^{(n)}]h_B[\mathbf{x}_{t/2}^{(n)}]\frac{\rho_{\mathrm{eq}}(\mathbf{v}_{t_0}^{(n)})}{\rho_{\mathrm{eq}}(\mathbf{v}_{t_0}^{(o)})}\frac{p(\{\mathbf{x}_{\pm t/2}^{(n)}\}|\mathbf{x},\mathbf{v}_{t_0}^{(n)},t_0^{(n)})}{p(\{\mathbf{x}_{\pm t/2}^{(o)}\}|\mathbf{x},\mathbf{v}_{t_0}^{(o)},t_0^{(o)})}. \tag{5}$$

h_A and h_B at the ends of the old path are unity, or else that path would not have been accepted. The equilibrium probability of the common shooting point $\mathbf{x}$ cancels leaving the probabilities of the old and new shooting velocities. Equations (5) and (3) show that detailed balance[29] in the transition path ensemble is obtained by accepting each new trajectory that goes from state A to state B.

$$p_{\mathrm{acc,TPS}}^{o\to n}=h_A[\mathbf{x}_{-t/2}^{(n)}]h_B[\mathbf{x}_{t/2}^{(n)}]. \tag{6}$$

Forward and backward reactions can be included in the transition path ensemble by accepting trajectories with probability $h_A[\mathbf{x}_{-t/2}^{(n)}]h_B[\mathbf{x}_{t/2}^{(n)}]+h_A[\mathbf{x}_{t/2}^{(n)}]h_B[\mathbf{x}_{-t/2}^{(n)}]$.

SHOOTING POINT DENSITY FROM AIMLESS SHOOTING

As the trajectories generated by aimless shooting evolve in trajectory space, the shooting points from aimless shooting evolve in configuration space. Each time a trajectory is accepted, the previous shooting points, $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$, are replaced by three new points, $\mathbf{x}_{-\Delta t}^{(n)}$, $\mathbf{x}_0^{(n)}$, and $\mathbf{x}_{\Delta t}^{(n)}$. One of the three new points is the old point that generated the new trajectory. Suppose $\mathbf{x}=\mathbf{x}_0^{(n)}$ is the point that generated the new trajectory. The probability of observing the two new points is

$$p(\mathbf{x}_{-\Delta t}^{(n)},\mathbf{x}_{\Delta t}^{(n)}|\mathbf{x}_0^{(n)},\mathrm{TP})=\frac{p(\mathbf{x}_{-\Delta t}^{(n)},\mathbf{x}_0^{(n)},\mathbf{x}_{\Delta t}^{(n)}|\mathrm{TP})}{p(\mathbf{x}_0^{(n)}|\mathrm{TP})}. \tag{7}$$

The probability that the next new trajectory will be obtained by shooting at $\mathbf{x}$ is $p(\mathrm{TP}|\mathbf{x})/3p_{\mathrm{acc}}^{(o)}$, where $\mathbf{x}$ is one of the points $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$ and

$$3p_{\mathrm{acc}}^{(o)}=p(\mathrm{TP}|\mathbf{x}_{-\Delta t}^{(o)})+p(\mathrm{TP}|\mathbf{x}_0^{(o)})+p(\mathrm{TP}|\mathbf{x}_{\Delta t}^{(o)}). \tag{8}$$

The transition probability from $\mathbf{x}_{-\Delta t}^{(o)}$, $\mathbf{x}_0^{(o)}$, and $\mathbf{x}_{\Delta t}^{(o)}$ to $\mathbf{x}_{-\Delta t}^{(n)}$, $\mathbf{x}_0^{(n)}$, and $\mathbf{x}_{\Delta t}^{(n)}$ is

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

$$p^{o\to n}=\frac{p(\mathrm{TP}|\mathbf{x})}{3p_{\mathrm{acc}}^{(n)}}\frac{p(\mathbf{x}_{-\Delta t}^{(n)},\mathbf{x}_0^{(n)},\mathbf{x}_{\Delta t}^{(n)}|\mathrm{TP})}{p(\mathbf{x}|\mathrm{TP})}. \tag{9}$$

Similarly, the probability of generating the old points from the new points is

$$p^{n\to o}=\frac{p(\mathrm{TP}|\mathbf{x})}{3p_{\mathrm{acc}}^{(n)}}\frac{p(\mathbf{x}_{-\Delta t}^{(o)},\mathbf{x}_0^{(o)},\mathbf{x}_{\Delta t}^{(o)}|\mathrm{TP})}{p(\mathbf{x}|\mathrm{TP})}. \tag{10}$$

The shooting point $\mathbf{x}$ is the same in both cases, so the ratio of transition probabilities is

$$\frac{p^{o\to n}}{p^{n\to o}}=\frac{p_{\mathrm{acc}}^{(n)}}{p_{\mathrm{acc}}^{(o)}}\frac{p(\mathbf{x}_{-\Delta t}^{(n)},\mathbf{x}_0^{(n)},\mathbf{x}_{\Delta t}^{(n)}|\mathrm{TP})}{p(\mathbf{x}_{-\Delta t}^{(o)},\mathbf{x}_0^{(o)},\mathbf{x}_{\Delta t}^{(o)}|\mathrm{TP})}. \tag{11}$$

Equation (11) shows that the distribution of shooting points is a "fuzzy" approximation to $p(\mathrm{TP}|\mathbf{x})p(\mathbf{x}|\mathrm{TP})$. For two states connected by a single pathway without any stable intermediates, $p(\mathrm{TP}|\mathbf{x})p(\mathbf{x}|\mathrm{TP})$ is peaked near transition states along the reaction coordinate. In directions orthogonal to the reaction coordinate, $p(\mathbf{x}|\mathrm{TP})$ is peaked at the center of the transition pathway.[10] Thus, aimless shooting distributes shooting points near probable transition states without *a priori* knowledge of their locations. An important exception occurs when a stable intermediate exists along the pathway. Suggestions for detecting and correcting this problem are given in a later section. Another advantage of aimless shooting is that Δt is the only adjustable parameter. Other TPS algorithms have shooting-shifting ratios and momentum perturbation parameters that must be chosen based on the system of interest.

REACTION COORDINATE FROM SHOOTING HISTORY

For a good reaction coordinate $r(\mathbf{x})$, $p(\mathrm{TP}|\mathbf{x})$ depends only on the reaction coordinate. Thus, we seek the function $p(\mathrm{TP}|r(\mathbf{x}))$ that is most likely to explain the realizations of $p(\mathrm{TP}|\mathbf{x})$ that were obtained from aimless shooting. Likelihood maximization is a powerful framework for learning models to explain probabilistic data.[30] Here, the data are the rejected and accepted shooting points from TPS, and the models are trial functions $r(\mathbf{x})$ and $p(\mathrm{TP}|r)$.

The model function for $p(\mathrm{TP}|r)$ must have a peak corresponding to the transition state value of r and decay to zero on both sides of the peak. One possible model is

$$p(\mathrm{TP}|r)=p_0(1-\tanh[r]^2), \tag{12}$$

where p_0 is an adjustable parameter. Equation (12) is symmetric with a peak at $r=0$, so transition states should be on the isosurface $r(\mathbf{x})=0$. For two limiting cases, p_0 can be identified *a priori*. In transition state theory, a transition state $\mathbf{x}$ satisfies $p(\mathrm{TP}|\mathbf{x})=1$, so for systems that obey transition state theory, $p_0=1$. For systems with diffusive barrier crossing dynamics, transition states satisfy $p(\mathrm{TP}|\mathbf{x})=1/2$.[11] Since transition states are on the surface $r(\mathbf{x})=0$ and $p(\mathrm{TP}|\mathbf{x})=2p_B(\mathbf{x})(1-p_B(\mathbf{x}))$,[11] $p_0=1/2$ and $p_B(r)=(1+\tanh[r])/2$. Figure 4 shows $p_B(r)$ and $p(\mathrm{TP}|r)$ from Eq. (12) applied to a system with diffusive barrier crossings.

The model function for $p(\mathrm{TP}|r)$ has only one adjustable parameter, so the model reaction coordinate should be very flexible. Suppose that the model reaction coordinate depends

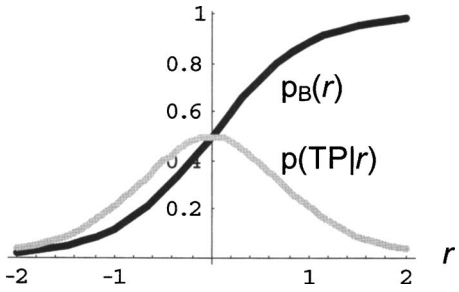

FIG. 4. $p(\mathrm{TP}|r)$ (gray) and $p_B(r)$ (black) as functions of r for a system with diffusive barrier crossings.

on a few collective variables, $\mathbf{q}=q_1,q_2,\ldots,q_m$. A later section extends this treatment to large sets of candidate variables. The simplest model of the reaction coordinate is the linear combination

$$r(\mathbf{q})=\sum_{k=1}^{m}\alpha_k q_k-\alpha_0, \tag{13}$$

where $\alpha_1,\ldots,\alpha_m$ and α_0 are free parameters. The coefficients $\alpha_1,\ldots,\alpha_m$ in the reaction coordinate absorb the units from the collective variables, so the reaction coordinate is dimensionless. The parameter α_0 allows the reaction coordinate to shift so that transition states are at $r(\mathbf{q})=0$.

The symmetry in Eq. (12) and interactions between collective variables may require additional flexibility in Eq. (13). The symmetry in $p(\mathrm{TP}|r)$ can be accommodated by including power law parameters (fixed or adjustable) in the reaction coordinate model, i.e., $r=\Sigma_k\alpha_k q_k^{mk}-\alpha_0$. Interactions can be included by adding a quadratic form to Eq. (13), i.e., $r=\Sigma_k\alpha_k q_k+\mathbf{q}^{\mathbf{T}}\mathbf{A}\mathbf{q}-\alpha_0$, where $\mathbf{A}$ is a matrix of adjustable parameters. Another way to include interactions is to include them explicitly among the q variables, for example, $q_3=q_1q_2$.

The models for $p(\mathrm{TP}|r)$ and $r(\mathbf{x})$ are used to calculate the likelihood of the model given the shooting data. The likelihood depends on the model and the free parameters,[30]

$$L(\boldsymbol{\alpha},p_0)=\prod_{k}^{N_{\mathrm{acc}}}p(\mathrm{TP}|r(\mathbf{q}_{\mathrm{acc}}^{(k)}))\prod_{k}^{N_{\mathrm{rej}}}(1-p(\mathrm{TP}|r(\mathbf{q}_{\mathrm{rej}}^{(k)}))), \tag{14}$$

where $\mathbf{q}_{\mathrm{rej}}^{(k)}$ and $\mathbf{q}_{\mathrm{acc}}^{(k)}$ are the collective variables at the kth rejected and accepted shooting points. N_{acc} and N_{rej} are the numbers of accepted and rejected shooting moves, respectively.

For diffusive dynamics each shooting point constitutes two independent realizations of $p_B(\mathbf{x})$, so the likelihood can be written in terms of $r(\mathbf{q}(\mathbf{x}))$ and $p_B(r)=(1+\tanh[r])/2$,

$$L(\boldsymbol{\alpha})=\prod_{k=1}^{B}p_B(r(\mathbf{q}^{(k)}))\prod_{k=1}^{\neq B}(1-p_B(r(\mathbf{q}^{(k)}))), \tag{15}$$

where B is the number of trajectory end points in B, $\neq B$ is the number of trajectory end points in A, and the $\mathbf{q}^{(k)}$ are the collective variables at the shooting points. Equation (15) distinguishes rejected trajectories with both end points in B from rejected trajectories with both end points in A. The likelihood in Eq. (14) does not distinguish between these types of rejected trajectories. Thus, Eq. (15) uses more of the

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

available information but applies only for diffusive dynamics where $p(\mathrm{TP}|\mathbf{x})=2p_B(\mathbf{x})(1-p_B(\mathbf{x}))$.[11]

The log likelihood[30,31] $\ell(\boldsymbol{\alpha})=\ln L(\boldsymbol{\alpha})$ is easily maximized to obtain the optimal parameters $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}=\arg\max_{\boldsymbol{\alpha}'}\ell(\boldsymbol{\alpha}'). \tag{16}$$

Equation (16) does not include p_0 in the optimization. This is appropriate for the two limits where p_0 is known. Otherwise, Eq. (14) should be used and p_0 should be optimized along with $\boldsymbol{\alpha}$.

Because the likelihood uses information from the full range of p_B values, the reaction coordinate applies at every p_B along the transition pathway. Likelihood maximization constructs the best possible reaction coordinate from a few collective variables. A few variables are often sufficient, but a major challenge is to learn *which variables* are important. The algorithm below compares the best reaction coordinates from combinations of many collective variables.

ALGORITHM TO SCREEN MANY CANDIDATE VARIABLES

Consider a large set of M collective variables as possible components of the reaction coordinate. The likelihood will increase each time a new parameter is added to the model,[31] but simple reaction coordinates are better for gaining insight.

Our strategy is to search the set of M collective variables starting from models of the reaction coordinate with just one variable. The best single variable reaction coordinate is then compared to the best reaction coordinate from all pairs of collective variables and that to the best coordinate from all combinations of three variables, etc. The Bayesian information criterion[32] determines when the benefit of additional model complexity is no longer significant. If N_R is the number of realizations in the likelihood function, an additional parameter (or an additional variable) gives a significant improvement if the likelihood increases by $(1/2)\ln N_R$.[30,32]

(1) Perform aimless shooting to harvest N_R independent realizations of $p(\mathrm{TP}|\mathbf{x})$ or $p_B(\mathbf{x})$ as appropriate for the dynamics of the system. This step *never has to be repeated*.
(2) Propose M candidate collective variables, $q_1,\ldots,q_M$. Choose forms for $r(\mathbf{q})$ and $p(\mathrm{TP}|r)$ such as Eqs. (12) and (13). Set $m=1$.
(3) For each of the $C(m,M)=M!/((M-m)!m!)$ combinations of m variables, maximize the log likelihood ℓ as in Eq. (16).
(4) Let ℓ_m be the maximum log likelihood among combinations of m variables from step (3).
(5) If $\ell_m-\ell_{m-1}<\frac{1}{2}\ln N_R$ or if $m=M$, stop. If $\ell_m-\ell_{m-1}>\frac{1}{2}\ln N_R$ or if $m=1$, set $m=m+1$ and repeat (3)–(5).

This algorithm finds the best reaction coordinate as a function of the candidate variables that are significantly involved in the reaction coordinate. If no combination from the candidate variables gives a satisfactory reaction coordinate, additional variables should be tested. Suggestions for detecting and correcting such problems are given below.

DIAGNOSING AND CORRECTING PROBLEMS

The optimal reaction coordinate should be tested with a p_B histogram to ensure that the set of candidate variables is adequate. Two types of problems can occur and each has a distinct signature in the histogram. A p_B histogram that is too broad but centered at the correct value of p_B indicates that additional candidate variables are needed. The data from the initial TPS simulation can be reused to screen the additional candidates. The additional variables are needed only at the saved shooting points. Trajectories and collective variables at their end points *need not be recomputed*. Optimal reaction coordinates involving the new variables should be compared to the best models without the new variable.

The second problem arises because the shooting points from TPS are distributed only in the transition pathway, where $p(\mathbf{x}|\mathrm{TP})\neq 0$. Estimated isocommittor surfaces may cut the transition pathway correctly but also pass through nonreactive low energy regions that were not sampled in TPS. This problem is unlikely because the reaction coordinate is based on information from the full range of p_B values. However, if it occurs, the p_B histogram for the predicted $p_B=1/2$ surface will be peaked at 0 or 1 or will have double peaks at $\frac{1}{2}$ and 0 or $\frac{1}{2}$ and 1. To correct this problem, add model complexity with power laws or variable interactions to "bend" isosurfaces away from the stable basins while retaining the characteristics of the original simple model in the transition pathway. Before reoptimization, the original shooting point data should be augmented with the low energy points sampled in computing the diagnostic p_B histogram. The points from off-pathway regions should be entered as points that generated nontransition paths in the likelihood function. It may be possible to prevent reaction coordinates from cutting through stable off-pathway regions by including points from the stable basins as points that generate nontransition paths in the likelihood function.

Problems can also occur in the sampling of shooting points. "Inconclusive" trajectories are paths for which one end fails to reach a basin. Inconclusive trajectories are a problem because they cannot be classified in the likelihood function. During the TPS simulation, inconclusive trajectories can be counted by monitoring $h_A[\mathbf{x}_{-t/2}]+h_B[\mathbf{x}_{-t/2}]$ and $h_A[\mathbf{x}_{t/2}]+h_B[\mathbf{x}_{t/2}]$. If either of these is zero, the trajectory is inconclusive. If both sums are 1, the trajectory is a transition path or a nontransition path. If either of these sums is 2, the A and B basins overlap. Note that it is sufficient to monitor the product $(h_A[\mathbf{x}^{(n)}_{-t/2}]+h_B[\mathbf{x}^{(n)}_{-t/2}])(h_A[\mathbf{x}^{(n)}_{t/2}]+h_B[\mathbf{x}^{(n)}_{t/2}])$. One reason for inconclusive trajectories is that the transition paths are not of long enough duration. This can be fixed by increasing the transition path time t.[6]

Stable intermediates can also hamper the harvesting of shooting points by aimless shooting. Stable intermediates could result in many inconclusive trajectories. If the fraction of inconclusive trajectories does not decrease upon increasing t, then a stable intermediate may be trapping the trajectories. On a diagram such as that shown in Fig. 11, the stable intermediate will appear as a highly concentrated region in the swarm of inconclusive trajectories. Figures such as Fig.

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

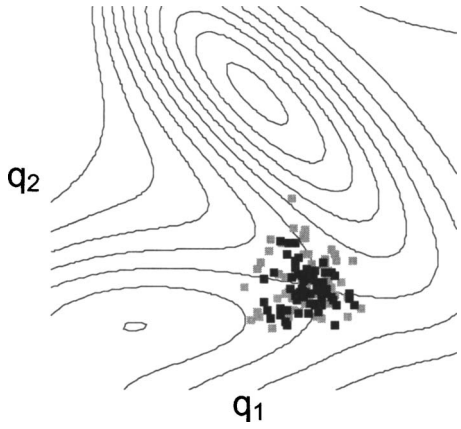

FIG. 5. Accepted (black) and rejected (gray) shooting points from the shooting algorithm. The shooting points are clustered near low energy transition states.

11 can thus help identify stable intermediates. Stable intermediates can then be treated as separate states or included in definitions of the reactant or product states.

EXAMPLE 1: LANGEVIN DYNAMICS ON A MODEL POTENTIAL ENERGY SURFACE

Here we apply the new method to a system evolving via Langevin dynamics on a model two-dimensional potential energy surface. The model surface[33] provides a convenient illustration because reactants, products, and transition states are visually identifiable. The Langevin equation[34] adds friction and random forces to Newtonian dynamics to simulate the effects of bath degrees of freedom,

$$-\nabla V-\gamma\dot{\mathbf{q}}+\boldsymbol{f}(t)=\ddot{\mathbf{q}}. \tag{17}$$

Here γ is the friction coefficient, $V(\mathbf{q})$ is a potential energy, $\mathbf{q}$ is a vector of two collective variables, and $\boldsymbol{f}(t)$ is a Gaussian random force with zero mean.[34] The variance of the random force is related to the friction coefficient by $\langle \boldsymbol{f}(t)\cdot\boldsymbol{f}(t')\rangle = 4k_BT\gamma\delta[t-t']$.[34]

Figure 5 shows rejected and accepted aimless shooting points on the model potential energy surface.[33] The energy contour spacing is $4k_BT$ and the friction coefficient was set to 125. Figure 6 shows $p(\mathrm{TP}|r(\mathbf{q}))$ from log-likelihood optimization based on the aimless shooting points in Fig. 5. We stress that all results in Figs. 5 and 6 are from aimless shooting. The energy contours are shown *only* to illustrate that the method identifies the correct dividing surface.

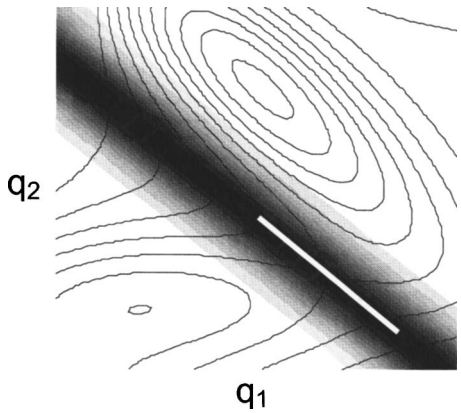

FIG. 6. $p(\mathrm{TP}|r(\mathbf{q}))$ from the shooting points in Fig. 5. The shading at $\mathbf{q}$ is proportional to $p(\mathrm{TP}|r(\mathbf{q}))$. The white line is the estimated $p_B=1/2$ surface.

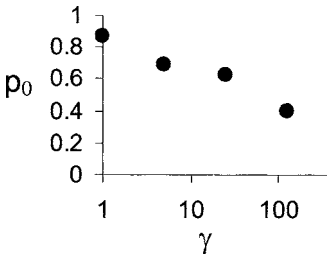

FIG. 7. p_0 as a function of the friction coefficient γ for the model free energy surface.

Figure 7 shows how friction affects the parameter $p_0 = p(\mathrm{TP}|0)$. For the exact reaction coordinate, $p_0 \to 1$ at low friction and $p_0 \to 1/2$ at high friction. The actual value of p_0 is smaller than expected at both high and low friction because the true dividing surface is not exactly linear. The linear approximation to the dividing surface includes as transition states some points with low $p(\mathrm{TP}|\mathbf{x})$ values.

EXAMPLE 2: NUCLEATION IN THE ISING MODEL

Nucleation is an activated process that initiates a transition from a metastable phase to a stable phase.[35] The process begins with a fluctuation in the metastable phase that forms an embryo, or nucleus, of the stable phase. Nucleation is activated because an interface must be created between the nucleus and the surrounding metastable phase.[35] Most nuclei are too small to overcome the interfacial energy barrier, so the nuclei vanish back into the metastable phase. Critical nuclei have equal probability of vanishing into the metastable phase and growing to a macroscopic domain of the stable phase.

Classical nucleation theory (CNT) assumes that the nuclei are small spheres with the free energy density of the macroscopic stable phase.[35] CNT also assumes that the interfacial energy is the nucleus area times the surface tension between the macroscopic phases.[35] These assumptions imply that the free energy to form a nucleus is entirely determined by the nucleus diameter d, $\Delta G = \pi\sigma d^2 - \Delta\mu\pi d^3/6$, where $\Delta\mu$ is the free energy difference between the two phases and σ is the surface tension. The size of the critical nucleus in CNT maximizes the free energy barrier, $d^{\ddagger} = 4\sigma/\Delta\mu$.[35]

In nucleation, the metastable phase is the reactant, critical nuclei are transition states, and the variable that determines the probability of nucleation is the reaction coordinate. Classical nucleation theory assumes that nucleus size is the reaction coordinate. Size is an important variable for nucleation,[35] but other variables such as surface area and internal order of nuclei may also be important.[9,36] The role of these variables in the reaction coordinate for nucleation has not been quantified.

The Ising model[37] has been used in several studies of nucleation.[9,38,39] For the Ising model, the energy is a function of the spins on a lattice.[37] Each spin takes values of ±1. To model nucleation, each spin indicates which phase is present

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

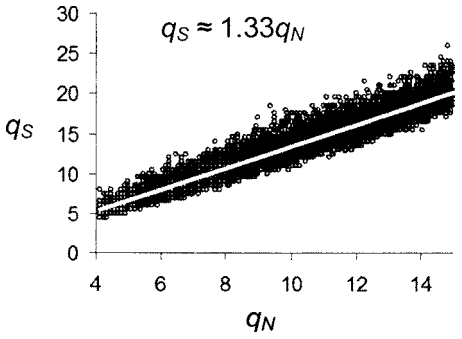

FIG. 8. Points from typical transition paths projected into collective variables q_N and q_s for the 2D Ising model. The white line is a least squares fit, $q_s=1.33q_N$. The slope indicates that nuclei are irregularly shaped throughout the transformation.

at a particular volume element, the coupling between spins is the interfacial energy between the phases, and the field strength is the difference in chemical potential between the perfect phases.[9] Thus, the energy as a function of the spins is

$$E(\mathbf{s}) = \frac{1}{2}\Delta\mu\sum_i s_i + \frac{1}{2}\sigma\sum_{\langle i,j\rangle} s_i s_j, \tag{18}$$

where the s_l are particular spins and $\langle i,j\rangle$ indicates a sum over all nearest neighbors.[9,37] Each attempted change in spin configuration consists of selecting a single random spin. Δt for aimless shooting is the number of spins in the lattice. Nucleation was studied on a square 32×32 periodic lattice and a cubic $32\times32\times32$ periodic lattice. The simulations were repeated on 24×24 and $24\times24\times24$ lattices to ensure that finite size effects are negligible. For the two-dimensional (2D) lattice, the temperature is $k_BT=0.7\sigma$ and the chemical potential is $\Delta\mu=0.2\sigma$. For the three-dimensional (3D) lattice, the temperature is $k_BT=1.35\sigma$ and the chemical potential is $\Delta\mu=0.55\sigma$. The 3D conditions and parameters were selected to enable a comparison with the results of Pan and Chandler.[9]

Equation (15) can be used for the likelihood function because the dynamics clearly satisfy $p(\mathrm{TP}|\mathbf{x})=2p_B(\mathbf{x})(1-p_B(\mathbf{x}))$.[9] In the Ising model we expect three nucleation regimes.[39] At very low temperatures the nuclei are perfect squares (or cubes).[39] Above the roughening temperature the nuclei become rounded.[39] Near the critical point, the nuclei will have nonconvex branched shapes.[35] To examine the role of cluster size N and surface area S, we define two parameters q_N and q_S,

$$\text{2D}\begin{cases} q_N=\sqrt{N}, \\ q_S=S/4, \end{cases} \qquad \text{3D}\begin{cases} q_N=\sqrt[3]{N}, \\ q_S=\sqrt{S/6}. \end{cases} \tag{19}$$

These are the lengths of a nucleus based on N and S assuming minimal surface area as in CNT. (On a lattice, squares and cubes have minimal surface area to volume ratio.) In two dimensions, S is the perimeter and N is the area. If the nuclei are perfect squares, $q_S/q_N=1$ in both 2D and 3D. For rounded nuclei, $q_S/q_N=2/\pi^{1/2}$ in 2D and $q_S/q_N=(6/\pi)^{1/3}$ in 3D. As the nuclei become elongated or nonconvex, the q_S/q_N ratios grow above the rounded values.

Figure 8 shows points from typical transition paths projected into q_N and q_S coordinates for the 2D Ising model. The linear relation between q_N and q_S shows that S scales as $N^{1/2}$ in agreement with classical nucleation theory. However, the slope (1.33) is larger than $2/\pi^{1/2}$ ($\approx$1.13), so the nuclei are irregularly shaped. Similarly, in the 3D system, S scales as $N^{2/3}$, but the slope is 1.31. A slope of $(6/\pi)^{1/3}$ ($\approx$1.25) is expected for nearly spherical nuclei.

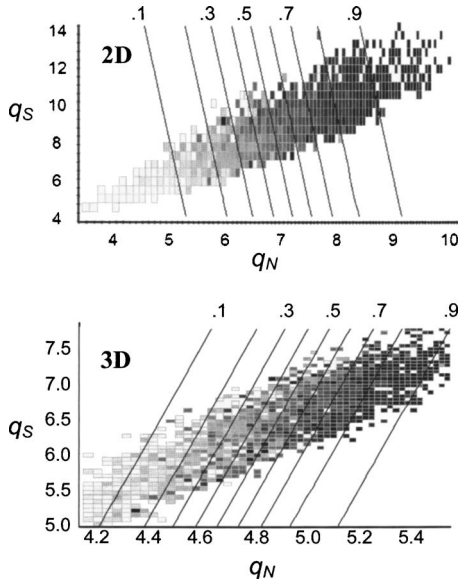

FIG. 9. (2D and 3D) The shading shows the fraction of trajectories that committed to B from each shooting point at (q_N,q_S). White corresponds to 0% and black to 100% with a gray scale in between. Only outlined boxes were sampled. The diagonal lines are the predicted $p_B(\mathbf{q})$ isocommittors from likelihood maximization. The isocommittor values are labeled above each graph.

Figure 9 shows the fraction of trajectories that commit to B from the shooting points projected into q_N and q_S coordinates. Figure 9 also shows p_B contours for optimal reaction coordinates of the form

$$r(q_N,q_S) = \alpha_N q_N + \alpha_S q_S - \alpha_0. \tag{20}$$

Table I shows maximum likelihood parameters and scores for reaction coordinate models that depend on q_S, q_N, or both variables.

If surface area is omitted from the reaction coordinate, likelihood maximization gives the same transition state sur-

TABLE I. Maximum likelihood parameters and scores for three models of the reaction coordinate. The threshold of significant likelihood increase from the Bayesian information criterion is given as BIC. The predicated transition state surfaces are also for each model. For the full models, tangents to the curve of transition states are given. The actual curves can be obtained by solving $r(N, S)=0$.

r	$(\alpha)_{\max}$	$r=0$ surface
2D: BIC$=(1/2)\ln N_R=4.8$		
$1.226q_S-7.260$	-6723.8	$S^{\ddagger}=35.1$
$0.681q_N-4.637$	-6273.5	$N^{\ddagger}=46.3$
$0.623q_N+0.042q_S-4.617$	-6268.5	$S^{\ddagger}=37.1-4.4(N^{\ddagger}-46)$
3D: BIC$=(1/2)\ln N_R=4.9$		
$1.106q_S-7.103$	-7400.2	$S^{\ddagger}=247.3$
$1.953q_N-9.504$	-6656.9	$N^{\ddagger}=115.3$
$2.559q_N-0.409q_S-9.826$	-6596.5	$S^{\ddagger}=246.5+6.8(N^{\ddagger}-115)$

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

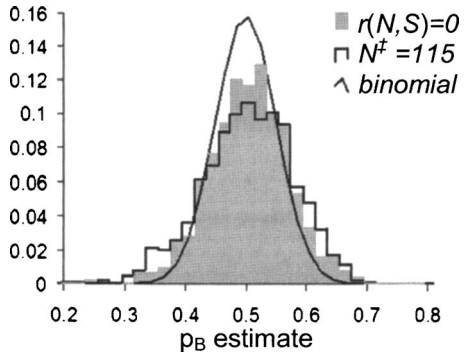

FIG. 10. p_B histograms for $N^{\ddagger}=115$ and $r(N,S)=0$ surfaces in the 3D Ising model. Individual p_B values are estimated from 100 trajectories. The histograms include estimates at 1000 points obtained by umbrella sampling. The bin width is 0.02. The binomial distribution that would result from sampling on the exact $p_B=1/2$ surface is also shown.

face that Pan and Chandler[9] obtained by computing the free energy as a function of N, $N^{\ddagger}=115$. Table I confirms that N is an important variable in the reaction coordinate,[9] but Table I also shows that the reaction coordinate involving both N and S is significantly better. The transition state ensemble includes both small compact nuclei and large distorted nuclei. Increasing the surface area at constant nucleus size decreases the probability that the nucleus will grow. The reaction coordinate from likelihood maximization quantifies the effect of added surface area on the stability of nuclei. Each unit increase in the size of a nucleus must be accompanied by a seven unit increase in surface area to maintain the same probability of nucleus growth or dissolution.

Surface area (perimeter) is also a significant factor in two dimensions. Interestingly, Fig. 9 (2D) shows that in two dimensions, increasing the surface area at constant nucleus size actually increases the probability that the nucleus will grow. Differences between the 2D and 3D results can be explained in terms of the probability that a single protruding spin from a smooth nucleus face will disappear or grow. Protrusions from an otherwise flat face on the two-dimensional nuclei tend to grow a new layer on the face. For the three-dimensional system, such protrusions tend to disappear. This can be verified with a simple calculation involving $\Delta\mu$, σ, k_BT, and the number of spins neighboring a protrusion.[40,41]

Figure 10 shows the p_B histogram for reaction coordinates from the three-dimensional Ising model. The p_B histogram for $S^{\ddagger}=247.3$ is shown in Fig. 1. The transition state surface for the best reaction coordinate is labeled $r(N,S)=0$. (The same window, $|r|<0.0275$, was sampled for each reaction coordinate.) Figure 10 also shows the binomial distribution that would result if every configuration sampled was a true transition state. Likelihood maximization correctly ranks the reaction coordinates, and the p_B histogram for the best reaction coordinate closely reproduces the binomial distribution. Table II summarizes the p_B-histogram results.

From the examples in this paper, the efficiency of aimless shooting appears to be near 25%. Efficiencies of 40% are typical for conventional shooting and shifting.[6,13] However, all successive aimless shooting trajectories are independent realizations of $p(\mathrm{TP}|\mathbf{x})$. Conventional shooting and shifting with slightly altered momenta may generate correlated $p(\mathrm{TP}|\mathbf{x})$ realizations for successive trajectories. The ability to use every trajectory in the likelihood makes aimless shooting preferable to conventional shooting and shifting for calculating reaction coordinates. A more detailed study of sampling efficiency is needed to determine which algorithm is more efficient for the calculation of rate constants by TPS.[13]

TABLE II. Mean and standard deviation (SD) of P_B histograms for predicted transition state surfaces. The exact transition state surface gives a binomial distribution.

Surface	Mean	SD
$S=247.3$	0.540	0.211
$N=115.3$	0.494	0.076
$r(N,S)=0$	0.495	0.065
Exact	0.500	0.050

CONCLUSIONS

This paper presented a powerful new method, based on likelihood maximization, to obtain a reaction coordinate from a list of candidate collective variables. The model reaction coordinates are constructed from candidate variables, and simple single variable models are tested first. The Bayesian information criterion[30] indicates when the benefit of increasing model complexity is no longer significant.

Unlike existing approaches, additional variables can be tested without sampling new trajectories or configurations on a constraint surface. The new variable is simply calculated at each of the saved shooting points. Optimal models involving the new variable are then compared to optimal models without the new variable.

Existing methods require iterative calculations of costly p_B histograms. The new method is approximately an order of magnitude less costly than the calculation of a single histogram. Additionally, the new method does not involve sampling on a constraint surface. In principle, sampling with a constraint is simple, but in practice, the constrained variables must be differentiable or the system must be amenable to Monte Carlo sampling. In truly complex systems, relevant variables may be discontinuous or nondifferentiable, and Monte Carlo moves that efficiently sample configuration space may be extremely complex.

The new method was applied to find reaction coordinates and transition states in two example problems. The example of Langevin dynamics on a bistable potential energy surface shows that the method is applicable for both high and low friction dynamics. The example of nucleation in the Ising model provides a new understanding of nucleation in the Ising model. These simple examples illustrate the power and efficiency of the new method. It should be very useful for understanding reactions in complex systems.

ACKNOWLEDGMENT

The authors thank Giovanni Ciccotti and Eric Vanden-Eijnden for valuable insight, suggestions, and encouragement.

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

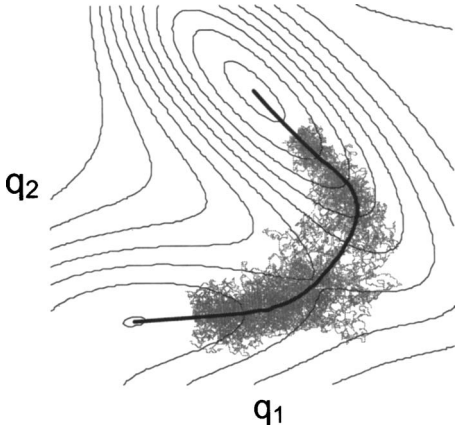

FIG. 11. The MLTP (black) follows the path of maximum density through $p(\mathbf{x}|\mathrm{TP})$, approximated by points from 100 trajectories in the TPE (gray). The MLTP connects the two minima. Contours of the free energy surface are shown only for visual reference.

APPENDIX: MOST LIKELY TRANSITION PATH

Each trajectory in the transition path ensemble is different, but the paths share some common features. If the collective variables in the reaction coordinate have been identified and used to compute a free energy surface, the transition paths should lie in a valley, or reaction pathway, that joins two minima on the free energy surface. The reaction pathway describes which degrees of freedom are changing at each point during the reaction. For reaction between small molecules, Miller *et al.*[42] described the reaction pathway as a harmonic valley centered on the steepest descent pathway downward from the saddle point to the neighboring minima on a potential energy surface.

For complex systems, Maragliano *et al.*[10] developed an equation for the analogous pathway on a free energy landscape as a function of collective variables. We refer to the pathway of Maragliano *et al.* as the most likely transition pathway (MLTP). The MLTP in collective variables $\mathbf{q}$ follows

$$\dot{\mathbf{q}} = -M(\mathbf{q})\nabla_{\mathbf{q}}F(\mathbf{q}) \tag{A1}$$

from the saddle point on the free energy surface $F(\mathbf{q})$ down to the neighboring minima.[10] The matrix $M(\mathbf{q})$ accounts for interdependencies within the set of order parameters $\mathbf{q}(\mathbf{x})$,[10]

$$M_{ij}(\mathbf{q}) = \langle \delta[\mathbf{q}(\mathbf{x}) - \mathbf{q}](\nabla_x q_i \cdot \nabla_x q_j)\rangle. \tag{A2}$$

The MLTP, like the steepest descent pathway of Miller *et al.*,[42] is not a dynamic path.[10] Maragliano *et al.*[10] compute $M(q)$ and $F(q)$ and use these in a collective variables version of the string method.[43]

In addition to the approach of Maragliano *et al.*, we propose that the MLTP can be approximated using trajectories from the transition path ensemble. Let the average of the collective variables in the reactant basin (A) be denoted $\langle \mathbf{q}\rangle_A = (\langle q_1\rangle_A, \ldots, \langle q_m\rangle_A)$ and define $\langle \mathbf{q}\rangle_B$ similarly for the product basin. The MLTP can be approximated by a curve from $\langle \mathbf{q}\rangle_A$ to $\langle \mathbf{q}\rangle_B$ that follows the path of maximum density through $p(\mathbf{x}|\mathrm{TP})$, and $p(\mathbf{x}|\mathrm{TP})$ can be approximated from points on the ensemble of transition paths. If $\mathbf{q}_{\mathrm{MLTP}}(s)$ is a parametrization of the MLTP, then $\mathbf{q}_{\mathrm{MLTP}}(s)$ satisfies

$$(\nabla\langle \delta[\mathbf{q}_{\mathrm{MLTP}}(s) - \mathbf{q}]\rangle_{\mathrm{TPE}})_{\perp} = 0, \tag{A3}$$

where the gradient is with respect to collective variables $\mathbf{q}$, the subscript TPE indicates an average over $p(\mathbf{x}|\mathrm{TP})$, and the subscript $\perp$ indicates components perpendicular to the MLTP tangent. For details on calculating such paths, see existing algorithms such as the string method[43,44] or the nudged elastic band.[45] To find the MLTP, substitute projected gradients of $p(\mathbf{x}|\mathrm{TP})$ for the projected forces that are used in those algorithms.[43–45] Figure 11 shows the MLTP on the model free energy surface of example 1 with 100 trajectories from the TPE.

[1] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[2] R. B. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[3] W. E, W. Ren, and E. Vanden-Eijnden, Chem. Phys. Lett. **413**, 242 (2005).
[4] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. Shaknovich, J. Chem. Phys. **108**, 334 (1998).
[5] P. G. Bolhuis, D. Chandler, C. Dellago, and P. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[6] C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[7] B. Ensing, A. Laio, F. L. Gervasio, M. Parrinello, and M. L. Klein, J. Am. Chem. Soc. **126**, 9492 (2004).
[8] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, Phys. Rev. B **28**, 784 (1983).
[9] A. C. Pan and D. Chandler, J. Phys.: Condens. Matter **1**, 0408331 (2004).
[10] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).
[11] G. Hummer, J. Chem. Phys. **120**, 516 (2004).
[12] C. Dellago, P. G. Bolhuis, F. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[13] C. Dellago, P. G. Bolhuis, and D. Chandler, J. Chem. Phys. **110**, 6617 (1999).
[14] T. van Erp, D. Moroni, and P. G. Bolhuis, J. Chem. Phys. **118**, 7762 (2003).
[15] D. Moroni, T. van Erp, and P. G. Bolhuis, J. Chem. Phys. **120**, 4055 (2004).
[16] R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).
[17] R. Radhakrishnan and B. L. Trout, J. Chem. Phys. **117**, 1786 (2002).
[18] R. Radhakrishnan and B. L. Trout, J. Am. Chem. Soc. **125**, 7743 (2003).
[19] R. Radhakrishnan and T. Schlick, Proc. Natl. Acad. Sci. U.S.A. **101**, 5970 (2004).
[20] M. F. Hagan, A. R. Dinner, D. Chandler, and A. Chakraborty, Proc. Natl. Acad. Sci. U.S.A. **100**, 13922 (2003).
[21] P. L. Geissler, C. Dellago, D. Chandler, J. Hutter, and M. Parrinello, Science **291**, 2121 (2001).
[22] X. Wu and B. R. Brooks, Biophys. J. **86**, 1946 (2004).
[23] P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **100**, 12129 (2003).
[24] S. Brown and T. Head-Gordon, Protein Sci. **13**, 958 (2004).
[25] J. W. Chu and B. L. Trout, J. Am. Chem. Soc. **126**, 16601 (2004).
[26] J. Marti, J. Phys.: Condens. Matter **16**, 5669 (2004).
[27] C. Lo, C. A. Giurumescu, R. Radhakrishnan, and B. L. Trout, Mol. Phys. **102**, 281 (2004).
[28] R. Radhakrishnan and T. Schlick, J. Chem. Phys. **121**, 2436 (2004).
[29] D. Frenkel and B. Smit, *Understanding Molecular Simulation* (Academic, San Diego, 2002).
[30] D. Husmeier, in *Probabilistic Modeling in Bioinformatics and Medical Informatics*, edited by D. Husmeier, R. Dybowski, and S. Roberts (Springer, London, 2005), p. 17.
[31] P. H. Garthwaite, I. T. Jolliffe, and B. Jones, *Statistical Inference* (Oxford, New York, 2002).
[32] G. Schwarz, Ann. Stat. **6**, 461 (1978).

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp

[33] K. Muller and L. D. Brown, Theor. Chim. Acta **53**, 75 (1979).
[34] D. J. Tildesley and M. P. Allen, *Computer Simulation of Liquids* (Oxford, New York, 1987).
[35] P. G. Debenedetti, *Metastable Liquids* (Princeton University Press, Princeton, 1996).
[36] D. Moroni, P. R. ten Wolde, and P. G. Bolhuis, Phys. Rev. Lett. **94**, 235703 (2005).
[37] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, New York, 1987).
[38] V. A. Schneidman, K. A. Jackson, and K. M. Beatty, J. Chem. Phys. **111**, 6932 (1999).
[39] S. Wonczak, R. Strey, and D. J. Stauffer, J. Chem. Phys. **113**, 1976 (2000).
[40] G. Berim and E. Ruckenstein, J. Chem. Phys. **117**, 4542 (2002).
[41] G. Berim and E. Ruckenstein, J. Chem. Phys. **117**, 7732 (2002).
[42] W. H. Miller, N. C. Handy, and J. E. Adams, J. Chem. Phys. **72**, 99 (1980).
[43] W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[44] B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, J. Chem. Phys. **120**, 7877 (2004).
[45] G. Mills and H. Jonsson, Phys. Rev. Lett. **72**, 1124 (1994).

Downloaded 12 Sep 2006 to 193.49.39.4. Redistribution subject to AIP license or copyright, see http://jcp.aip.org/jcp/copyright.jsp