Using the histogram test to quantify reaction coordinate error

Baron Peters[a)]
CECAM (Centre Européen de Calcul Atomique Moléculaire), Ecole Normale Supérieure, 46 Allée d'Italie, 69364 Lyon Cedex 7, France

(Received 3 October 2006; accepted 21 November 2006; published online 27 December 2006)

Many schemes for calculating reaction rates and free energy barriers require an accurate reaction coordinate, but it is difficult to quantify reaction coordinate accuracy for complex processes like protein folding and nucleation. The histogram test, based on *estimated* committor probabilities, is often used as a qualitative indicator for good reaction coordinates. This paper derives the mean and variance of the intrinsic committor distribution in terms of the mean and variance of the histogram of committor estimates. These convenient formulas enable the first quantitative calculations of reaction coordinate error for complex systems. An example shows that the approximate transition state surface from Peters' and Trout's reaction coordinate for nucleation in the Ising model gives a mean committor probability of 0.495 and a standard deviation of 0.042. *© 2006 American Institute of Physics*. [DOI: 10.1063/1.2409924]

INTRODUCTION

A reaction coordinate is a single degree of freedom that quantifies the dynamical progress along a reaction pathway.[1] Reaction coordinates have many applications in chemical kinetics.[2] For example, accurate reaction coordinates are required to calculate kinetically relevant free energy barriers[3–6] and rates from transition state theory.[2] Accurate reaction coordinates also facilitate transmission coefficient calculations.[3,7] Moreover, knowing the reaction coordinate is tantamount to understanding the reaction mechanism.[1] New methods can approximate reaction coordinates for complex biomolecular isomerizations[8–10] and nucleation processes,[1] but it remains difficult to quantify errors in the approximate reaction coordinates. This paper develops a simple quantitative measure of reaction coordinate error.

The exact reaction coordinate at a configuration $\mathbf{x}$ is the committor probability,[1,9] the fraction of trajectories initiated from $\mathbf{x}$ with Boltzmann distributed initial velocities that commit to the product basin.[11] The committor probability is denoted $p(\mathbf{x})$ and takes the value of 1 for products, 0 for reactants, and 1/2 for transition states.[11] Because the committor probability is an exact reaction coordinate, isosurfaces of a good reaction coordinate, $r(\mathbf{x})=\text{const}$, should closely approximate the isocommittor surfaces, $p(\mathbf{x})=\text{const}$.[1,9,10]

The histogram test[11–13] is a common diagnostic for evaluating an approximate reaction coordinate. The histogram test begins with a Boltzmann distributed sample of configurations from an isosurface of an approximate reaction coordinate.[11] Because transition states are particularly important, an approximate transition state surface is usually tested first. At each sampled configuration $\mathbf{x}$, an estimate of $p(\mathbf{x})$ is calculated by initiating a finite number of trajectories from $\mathbf{x}$ with Boltzmann distributed velocities.[11] The fraction of trajectories that commit to the product basin are counted at each configuration and a histogram of estimated committor probabilities is constructed.[11] For a good reaction coordinate, the histogram will be closely centered on a characteristic committor value.[9,11] For example, an approximate transition state surface should give a histogram centered around the characteristic committor value for transition states, $p=1/2$.[11]

The histogram is often referred to as a committor distribution, but in reality the histogram is a sample from the distribution of noisy committor estimates. This paper quantitatively relates the histogram to the intrinsic committor distribution. The mean of the intrinsic committor distribution identifies the characteristic isocommittor surface that is approximated by the reaction coordinate isosurface. The standard deviation of the intrinsic committor distribution quantifies deviations from the characteristic isocommittor value.

MEAN AND VARIANCE OF THE INTRINSIC COMMITTOR DISTRIBUTION

The distribution of committor probabilities on an approximate reaction coordinate isosurface is an intrinsic property of the isosurface, the equilibrium distribution, and the reaction dynamics. If $\rho(p)$ is the intrinsic committor distribution, $\rho_{\text{EQ}}(\mathbf{x})$ is the equilibrium distribution of configurations $\mathbf{x}$, and the isosurface is $r(\mathbf{x})=r^*$, then[14]

$$\rho(p)=\int_{r(\mathbf{x})=r^*} d\mathbf{x}\rho_{\text{EQ}}(\mathbf{x})\delta[p(\mathbf{x})-p]. \tag{1}$$

The estimator of the committor probability at $\mathbf{x}$ is $\hat{p}=k/N$ where k is the number of trajectories from N trials that commit to the product basin. The estimate is a binomial random variable with parameters $p=p(\mathbf{x})$ and N.[15,16] Let the distribution of committor estimates on the same isosurface be $h(\hat{p})$. The binomial distribution links the distributions $h(\hat{p})$ and $\rho(p)$. Isosurfaces of a perfect reaction coordinate exactly follow the isocommittor surfaces, so $\rho(p)=\delta[p-p(\mathbf{x})]$ and $h(\hat{p})$ must be the binomial distribution,[15,17]

[a)] Electronic mail: bpeters@cecam.org

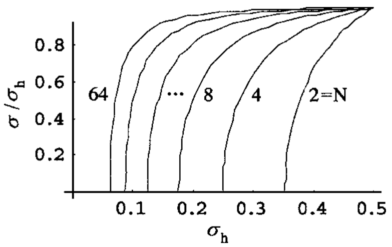

FIG. 1. σ/σ_h as a function of σ_h when μ_h=1/2, i.e., for an approximate transition state ensemble. The curves are labeled according to the number of trajectories per $\hat{p}$ estimate.

$$h(\hat{p}) = B_{N,p}(\hat{p}) = \binom{N}{N\hat{p}} p^{N\hat{p}}(1-p)^{N-N\hat{p}}, \tag{2}$$

where $\hat{p}$ can be $0/N, 1/N, \ldots,$ or N/N.

The isosurfaces of an approximate reaction coordinate will deviate from the true isocommittor surfaces, so the variance of $\rho(p)$ will be nonzero. In this case, ρ(p) and $h(\hat{p})$ are related by the equation

$$h(\hat{p}) = \int dp \rho(p) B_{N,p}(\hat{p}). \tag{3}$$

The mean and variance in $h(\hat{p})$ can be computed from the mean and variance in $\rho(p)$. The mean in $h(\hat{p})$ is

$$\begin{aligned}\mu_h &\equiv \sum_{\hat{p}} \hat{p} h(\hat{p}) = \sum_{\hat{p}} \hat{p} \int dp \rho(p) B_{N,p}(\hat{p}) \\ &= \int dp \rho(p) \sum_{\hat{p}} \hat{p} B_{N,p}(\hat{p}) \\ &= \int dp \rho(p) p \equiv \mu. \end{aligned}\tag{4}$$

So the mean of p in the distribution $\rho(p)$, μ, equals the mean of $\hat{p}$ in the distribution $h(\hat{p})$, μ_h. The variance in $h(\hat{p})$ can be computed similarly. If σ_h^2 and σ^2 are the variances in $h(\hat{p})$ and $\rho(p)$, respectively, then

$$\sigma^2 = \frac{N}{N-1}\left(\sigma_h^2 - \frac{1}{N}\mu_h(1-\mu_h)\right). \tag{5}$$

σ is the root mean square (rms) difference in intrinsic committor probabilities between an approximate reaction coordinate surface and the isocommittor surface $p(\mathbf{x})=\mu$.

Given μ_h, the lower and upper bounds for σ_h^2 are the variances from a perfect reaction coordinate $(\rho(p)=\delta[p-\mu_h])$ and a reactant/product mixture $(\rho(p)=\mu_h\delta[p-1]+(1-\mu_h)\delta[p])$.

$$\mu_h(1-\mu_h)/N \leqslant \sigma_h^2 \leqslant \mu_h(1-\mu_h). \tag{6}$$

Equations (5) and (6) show that σ^2 is proportional to the "excess variance" in $h(\hat{p})$ beyond the minimum possible (binomial) variance given μ_h. Substituting the upper and lower bounds into Eq. (5) shows that σ can take values from zero to σ_h. Figure 1 shows σ/σ_h as a function of σ_h when μ_h =1/2, i.e., for an approximate transition state ensemble.

Figure 1 shows why histogram tests are often performed with $N \geqslant 100$ trajectories per estimate.[1,9,18] For large N and $\sigma_h > 1/N^{1/2}$, $\sigma \approx \sigma_h$. However, for good reaction coordinates where σ_h approaches $1/2N^{1/2}$, the approximation $\sigma \approx \sigma_h$ is inaccurate and Eq. (5) is needed to quantify reaction coordinate error.

UNCERTAINTY IN THE INTRINSIC MEAN AND VARIANCE

The quantities μ_h and σ_h^2 are exact but, in practice, histogram tests are based on a finite sample from $h(\hat{p})$.[11] Uncertainty in estimates of μ_h and σ_h^2 propagate through Eqs. (4) and (5) into the computed values of μ and σ. Let $\hat{\mu}_h$ and $\hat{\sigma}_h^2$ be estimates of μ_h and σ_h^2 from a histogram test based on "n" samples from $h(\hat{p})$. Note that n and N are different numbers. N is the number of trajectories per estimate, and n is the number of estimates in the histogram test. The central limit theorem says that variances in the estimates $\hat{\mu}_h$ and $\hat{\sigma}_h^2$ are

$$\mathrm{var}[\hat{\mu}_h] = \sigma_h^2/n \tag{7}$$

and

$$\mathrm{var}[\hat{\sigma}_h^2] = (\mu_{4,h} - \sigma_h^4)/n, \tag{8}$$

where $\mu_{4,h}$ is the fourth central moment of $h(\hat{p})$. Using Eq. (4), the uncertainty in μ is

$$\delta\mu = \frac{\sigma_h}{\sqrt{n}} = \left(\frac{1}{nN}\mu(1-\mu) + \frac{N-1}{nN}\sigma^2\right)^{1/2}. \tag{9}$$

The uncertainty in σ can be estimated similarly using Eqs. (5) and (8),

$$2\sigma\delta\sigma = \frac{N}{N-1}\left(\frac{\sqrt{\mu_{4h}-\sigma_h^4}}{\sqrt{n}} + \frac{|1-2\mu_h|}{N}\frac{\sigma_h}{\sqrt{n}}\right). \tag{10}$$

A general expression for $\mu_{4,h} - \sigma_h^4$ can be obtained in terms of the first four cumulants of $\rho(p)$, but sufficiently narrow intrinsic committor distributions can be approximated by a Gaussian with mean μ and variance σ^2. Let $t = nN$ be the total number of trajectories in the histogram test. When $\mu \approx 1/2$, the relative uncertainties $\delta\mu/\mu$ and $\delta\sigma/\sigma$ each multiplied by $t^{1/2}$ are

$$\sqrt{t}\frac{\delta\mu}{\mu}\bigg|_{\mu=1/2} = \left(\tfrac{1}{4} + (N-1)\sigma^2\right)^{1/2} \tag{11}$$

and

$$\begin{aligned}\sqrt{t}\frac{\delta\sigma}{\sigma}\bigg|_{\mu=1/2} &= \frac{1}{2\sqrt{N-1}}\left(\frac{1}{8\sigma^4} + \frac{(N-2)}{\sigma^2}\right. \\ &\quad \left. + 2((N-3)^2 - N)\right)^{1/2}. \end{aligned}\tag{12}$$

The calculation of μ is generally more accurate than the calculation of σ. Figure 2 shows that when a histogram test accurately yields σ, it also accurately yields μ.

Figure 3 shows the number of trajectories required to achieve 10% relative uncertainty in σ. The required number of trajectories depends on N, and the optimal value of N depends on the error σ in the reaction coordinate being tested. Like calculating a small transmission coefficient,[3] many trajectories are needed to accurately compute the error

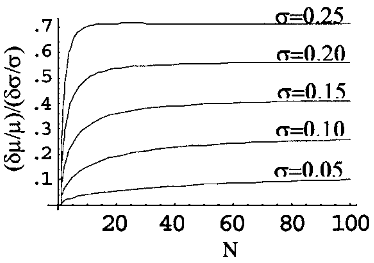

FIG. 2. Ratio of relative uncertainties, $(\delta\mu/\mu)/(\delta\sigma/\sigma)$, as a function of N for an intrinsic committor distribution that is Gaussian with $\mu=1/2$ and variance σ^2.

in a good reaction coordinate. For example, with N=100 it takes 20 000 trajectories to determine that $\sigma=0.05$ with 10% uncertainty. However, when $\sigma\geqslant 0.20$ it takes only 1000 trajectories using $N=8$ to compute σ with 10% relative uncertainty and μ with 5% relative uncertainty.

In applications σ is not known *a priori*, so the optimal value of N and the required number of trajectories are also unknown. The following strategy takes advantage of the finding that histogram tests with a few thousand trajectories are adequate for quantifying error in some reaction coordinates. First, sample $n=200$ configurations for the histogram test before computing any trajectories. Then compute a histogram with a small value of N, say $N=10$. Calculate the mean and variance of the intrinsic committor distribution using Eqs. (4) and (5). Then estimate $\delta\sigma/\sigma$ using Eq. (12). If the uncertainty is unacceptably large, more accurate values of μ and σ can be computed by increasing N with the same set of configurations. For reaction coordinates with $\sigma\geqslant 0.15$ the test will be complete with just 2000 trajectories.

EXAMPLE

The reaction pathway for nucleation links a metastable phase (the reactant) to the globally stable phase (the product).[18] Critical nuclei are transition states in the nucleation process because they have 50% probability of growing to form a globally stable phase and 50% probability of dissolving back into the metastable phase.[18] Equation (5) can quantify errors in the reaction coordinates of Peters and Trout[1] for nucleation in the three-dimensional Ising model.

One of the coordinates of Peters and Trout[1] involves only the surface area of the nucleus, one involves only the nucleus size (volume), and one involves both the nucleus area and size. Peters and Trout[1] published the mean and standard deviation from histogram tests on these approximate reaction coordinates and the number of trajectories per committor estimate, $N=100$. Table I shows the mean and standard deviation of the intrinsic committor distribution from the three approximate transition state (critical nucleus) surfaces. The mean and standard deviation of their original histogram tests[1] are also shown. The coordinate involving nucleus area and size gives the lowest error, with an intrinsic committor distribution characterized by $p=\mu\pm\sigma$ where $\mu=0.495$ and $\sigma=0.042$.

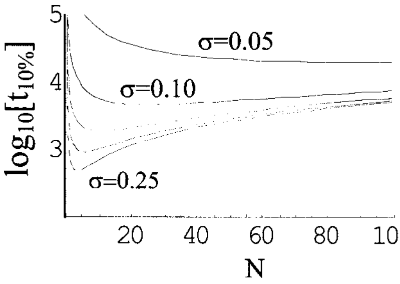

FIG. 3. Log of the total number of trajectories (nN) required to compute the intrinsic standard deviation σ to within 10% of the actual value for an approximate transition state ensemble.

TABLE I. μ and σ for the intrinsic committor distribution in the approximate transition state ensembles from three coordinates of nucleation in the Ising model. The table also shows the mean and standard deviation of committor estimates from the original histogram tests and the number of configurations sampled in each histogram. Relative uncertainties in μ and σ are less than 5%.

	Histogram		Intrinsic		
	$\hat{\mu}_h$	$\hat{\sigma}_h$	μ	σ	n
Nucleus area	0.540	0.211	0.540	0.206	320
Nucleus size	0.494	0.076	0.494	0.058	1000
Area and size	0.495	0.065	0.495	0.042	1000
Exact (binomial)	0.500	0.050	0.500	0.000	

CONCLUSIONS

The intrinsic committor probability distribution on an isosurface of an approximate reaction coordinate is related to the distribution of committor probability estimates on the isosurface, but these distributions are not equal. Equation (1) shows how $\rho(p)$, the intrinsic committor distribution, is related to $h(\hat{p})$, the distribution of committor estimates. Equations (4) and (5) show that the mean and variance in $\rho(p)$ can be computed from the mean and variance in $h(\hat{p})$. In practice, the mean and variance of $h(\hat{p})$ are available only as estimates from a histogram test. The finite number of samples in the histogram test leads to uncertainty in the calculated mean and variance of $\rho(p)$. Expressions for the uncertainty are derived and a practical strategy is recommended to accurately obtain the mean and standard deviation of $\rho(p)$ with a few thousand trajectories when possible. Data from Peters and Trout[1] were used to compute the mean and variance of the intrinsic committor distribution on approximate transition state surfaces for nucleation in the Ising model. The best reaction coordinate from Peters and Trout[1] gives an approximate transition state surface with a mean intrinsic committor probability of $\mu=0.495$ and an intrinsic standard deviation of $\sigma=0.042$.

ACKNOWLEDGMENTS

This work has been partially supported by the EC through the Marie Curie EXT project MEXTCT-2005-023311. The author also would like to thank Gregg Beckham and Bernhardt L. Trout for discussions that inspired this work, and Berend Smit for helpful suggestions on the manuscript.

[1] B. Peters and B. L. Trout, J. Chem. Phys. **125**, 054108 (2006).

[2] K. J. Laidler, in *Chemical Kinetics* (Harper & Row, New York, 1987).

[3] D. Frenkel and B. Smit, in *Understanding Molecular Simulation*, 2nd Edition (Academic Press, San Diego, 2001).

[4] S. C. Yang, J. N. Onuchic, and H. Levine, J. Chem. Phys. **125**, 054910 (2006).

[5] R. Radhakrishnan and B. L. Trout, J. Chem. Phys. **117**, 1786 (2002).

[6] V. K. Shen and P. G. Debenedetti, J. Chem. Phys. **111**, 3581 (1999).

[7] B. J. Berne, M. Borkovec, and J. E. Straub, J. Phys. Chem. **92**, 3711 (1988).

[8] R. B. Best and G. Hummer, PNAS **102**, 6732 (2005).

[9] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).

[10] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).

[11] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).

[12] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).

[13] P. L. Geissler, C. Dellago, and D. Chandler, J. Phys. Chem. B **103**, 3706 (1999).

[14] C. Dellago, P. G. Bolhuis, and P. L. Geissler, Adv. Chem. Phys. **123**, 1 (2002).

[15] mathworld.wolfram.com/BinomialDistribution.html

[16] C. D. Snow, Y. M. Rhee, and V. S. Pande, Biophys. J. **91**, 14 (2006).

[17] A. Waghe, J. C. Rasaiah, and G. Hummer, J. Chem. Phys. **117**, 10789 (2002).

[18] A. C. Pan and D. Chandler, J. Phys. Chem. B **108**, 19681 (2004).