Chemical Physics Letters 494 (2010) 100–103

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

$p(\mathrm{TP}|q)$ peak maximization: Necessary but not sufficient for reaction coordinate accuracy

Baron Peters *

Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA
Department of Chemistry and Biochemistry, University of California, Santa Barbara, CA 93106, USA

ARTICLE INFO

Article history:
Received 31 January 2010
In final form 21 May 2010
Available online 27 May 2010

ABSTRACT

Free energy hysteresis errors and other problems can arise from coordinates that are only optimized to accurately parameterize the separatrix. In this sense, an accurate separatrix is necessary but not sufficient to ensure reaction coordinate accuracy. For diffusive dynamics we prove that maximizing the peak in the projected transition path probability $p(\mathrm{TP}|q)$ is equivalent to separatrix optimization. Thus methods based on this criterion [27] may find coordinates that accurately parameterize the separatrix, but not earlier and later stages along the reaction pathway.

1. Introduction

Many reaction rate theories and simulation methods for computing rate constants assume the availability of an accurate reaction coordinate [1]. Reaction coordinate error has been thoroughly investigated in the low-to-intermediate friction regime [2,3]. Theories that resulted from those endeavors include the variational transition state theory (VTST) [2,4] and reactive flux methods for transmission coefficients [5,6]. This Letter focuses on the high friction, or 'diffusive', barrier crossing regime where transmission coefficients are often too small to be calculated efficiently [6,7]. Most theories of reaction dynamics in the high friction limit also assume *a priori* knowledge of accurate reaction coordinates or sufficient components to construct an accurate reaction coordinate. These theories have converged upon the *committor probability*, p_B or p_{fold}, as the ideal reaction coordinate for systems with diffusive dynamics [8–15]. The committor probability at a configuration $\mathbf{x}$ is the probability that a trajectory initiated with Boltzmann distributed momenta from $\mathbf{x}$ will relax to the product state (B) rather than to the reactant state (A). The committor probability identifies reactants as points with $p_B = 0$, products as points with $p_B = 1$, and a spectrum of intermediate values between. The special value $p_B = 1/2$ indicates a transition state. The committor probability itself provides limited physical insight, so models that show how the committor probability depends on intuitively meaningful physical variables are useful [13,15–18]. Several new methods can optimize physical models of the committor probability coordinate from atomistic simulation data [13,16–18]. These new methods provide a rigorous link from molecular simulations to accurate mean first passage times [11,19] and low dimensional Fokker–Planck and Smoluchowski models of the dynamics [1,20–22].

In addition to the mechanistic insights gained from linking simulations to simple dynamical models, accurate reaction coordinates have practical computational advantages. Several authors have noted that inaccurate reaction coordinates can cause hysteresis in computed free energy profiles [3], free energy landscapes that are inconsistent with the dynamics [15], erroneously low activation barriers [2], and paths that do not follow the true mechanism [23]. These problems can arise even when the reaction coordinate seems obvious. For example, Bolhuis et al. [24] showed that the Ramachandran angles which had been used in many studies of the alanine dipeptide, are inaccurate reaction coordinates. Ma and Dinner used a pioneering neural network method to model the committor probability and show that the reaction coordinate includes an essential solvent component [13].

This Letter contrasts methods that optimize the reaction coordinate at all stages of the reaction [13,16–18,25] from methods that only optimize the separatrix [22,26,27]. We discuss problems that can arise from reaction coordinates obtained by separatrix optimization methods. We review expressions for the mean and variance of the true committor distribution on a trial isosurface. Finally, we use these expressions to prove that maximizing the peak in the transition path probability [22,27] is equivalent to optimizing the separatrix.

2. Reaction coordinate and separatrix optimization methods

Committor analysis has been used to identify or validate reaction coordinates in many systems [9,10,15,28–33]. We briefly review committor analysis below, but first note that the expense of the trial-and-error committor analysis procedure has recently motivated more systematic and efficient approaches. The seminal

* Address: Department of Chemical Engineering, University of California, Santa Barbara, CA 93106, USA.
E-mail address: baronp@engineering.ucsb.edu

0009-2614/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2010.05.069

advance from Ma and Dinner [13] generated a training set of p_B-estimates throughout the reaction pathway. They used a genetic algorithm with least-squares to construct a neural network model from thousands of candidate reaction coordinates and the training set data. Peters, Trout, and Beckham noted that the shooting point outcomes from transition path sampling are p_B-realizations that can be used with likelihood maximization to efficiently identify reaction coordinates [16,17]. Borrero and Escobedo [18] extended these ideas to use data from forward flux sampling [34] simulations. The String method of Maragliano et al. [35] requires the subspace of all important variables as inputs, but given an acceptable subspace the string method can provide an accurate coordinate. Bolhuis combined the String method with likelihood maximization to help identify the important variables [36]. Noe et al. [25] use state-to-state transition probabilities from short trajectory data to construct complex maps of the committor probability throughout the reaction pathway. This and related network analysis methods [37–42] are extremely powerful when the overall pathway is a network of parallel and in-series pathways between intermediates. The above methods optimize the reaction coordinate at all stages: in the early pre-organization steps, in the transition state region, and in the late re-organization steps.

Another group of methods are based on properties of the transition state ensemble and transition paths. Antoniou and Schwartz [26] first obtain a transition state ensemble by sampling configurations with committor probabilities near $p_B = 1/2$, and then seek the coordinate with the narrowest distribution in the transition state ensemble. Best and Hummer [22,27] seek the coordinate that gives the maximum peak in $p(\mathrm{TP}|q)$, the projected probability of observing transition paths from a Boltzmann distributed ensemble of points on the q-isosurface. These two methods optimize only the parameterization of the separatrix.

A coordinate having one isosurface that accurately parameterizes the separatrix may still poorly describe progress at earlier and later stages. Events before and after the separatrix are often mechanistically important [43] so reaction coordinates that accurately describe these early and late stages have advantages over coordinates that are only accurate at the separatrix. Consider an example from our work on methane diffusion by hopping from occupied to vacant cages in clathrate hydrates [44]. Fig. 1 shows the free energy landscape with isosurfaces of two coordinates that were studied in our work. The linear progress coordinate and the curvilinear bipolar coordinate [45] both give the same separatrix by symmetry. However, free energy calculations using the linear progress coordinate resulted in hysteresis. Hysteresis occurs because the linear progress coordinate does not follow the direction in which the pathway leaves the stable intermediate. The curvilinear bipolar coordinate does follow the path out of the intermediate at early ($p_B \approx 0.3$) and late ($p_B \approx 0.7$) stages. These two coordinate systems would appear equally acceptable for a separatrix optimization method. This example illustrates the practical difficulties that can arise from coordinate inaccuracy even when the coordinate is accurate at the separatrix.

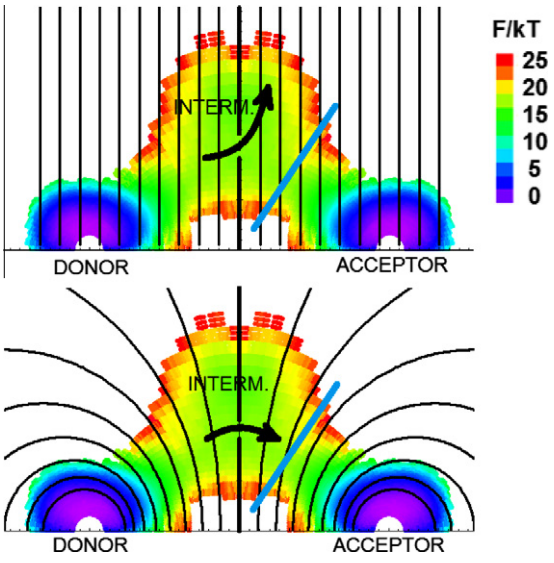

Fig. 1. Free energy surface for cage-to-vacant cage hopping of methane in a gas hydrate [44]. The diagonal blue line indicates a dividing surface between the acceptor cage and a shallow stable intermediate. The separatrix is the plane of symmetry, which passes through a shallow intermediate. The diagonal blue line is a dividing surface between the acceptor and intermediate states. (Above) Free energy calculations using a linear coordinate gave hysteresis because after crossing the separatrix methane can advance along the coordinate by following the black arrow. (Below) Free energy calculations converged without hysteresis for a bipolar coordinate that shares the same separatrix.

Stable intermediates are common along protein folding pathways [46], so reaction coordinates from separatrix optimization may poorly describe progress along the early and late stages of the pathway. It may also be impossible to identify simple and physically meaningful coordinates that describe an entire protein folding pathway, but network analysis methods like that of Noe et al. [25] can compute p_B-values on a complex network of interconnected channels.

3. Improved version of committor analysis

Committor analysis was originally introduced by Du et al. [9] and by Geissler et al. [10] to evaluate reaction coordinate error. The analysis begins with a Boltzmann distributed sample of atomistic configurations on an isosurface $q(\mathbf{x}) = q$ for a trial coordinate $q(\mathbf{x})$. For the remainder of this manuscript, q denotes a specific value of the coordinate dependent function $q(\mathbf{x})$. In our notation, q and $q(\mathbf{x})$ may or may not be equal depending on the argument $\mathbf{x}$ in $q(\mathbf{x})$. At each sampled configuration $\mathbf{x}$, $p_B(\mathbf{x})$ is estimated by initiating N trajectories from $\mathbf{x}$ with Boltzmann distributed velocities. The fraction of trajectories from $\mathbf{x}$ that commit to the product basin provides a p_B-estimate, $\hat{p}_B(\mathbf{x})$ [15]. The p_B-estimates from the Boltzman distribution of configurations on the isosurface are combined into a histogram. For an accurate reaction coordinate according to the p_B-definition, the histogram for each isosurface tested will give a distribution of p_B-estimates that is sharply peaked around a characteristic value for that isosurface [15]. Most studies in the literature only test for accuracy of the $p_B = 1/2$ surface, i.e. for an accurate parameterization of the separatrix. However, multiple isosurfaces of $q(\mathbf{x})$ can be tested to ensure an accurate model of the committor probability at early stages, at the transition state, and at late stages along the reaction pathway [12,13,30].

Committor analysis tests whether a single coordinate $q(\mathbf{x})$ can predict the long-time dynamics along its own q-axis, or whether other variables are needed. *For a good reaction coordinate, the initial position along the q-axis governs the resulting distribution of positions along the q-axis at longer times regardless of other coordinates in the initial positions.* This criterion is not specific to a separatrix, but rather it can be applied to any isosurface of a coordinate $q(\mathbf{x})$.

Peters showed how statistical error in the p_B-estimation process can be deconvoluted from the actual reaction coordinate error [47]. The result is a version of committor analysis that is quantitative, and less expensive than the original p_B-histogram test. Instead of using highly accurate p_B-estimates with $N \approx 100$ trajectories per estimate to directly compute the committor distribution, a histogram of inexpensive ($N \ll 100$) estimates can be related to the distribution of exact p_B-values using the binomial convolution [47]:

$$H(\hat{p}_B|q) = \int_0^1 dp_B B_{N,p_B}(\hat{p}_B) P(p_B|q). \quad (1)$$

In Eq. (1), $\hat{p}_B$ is an estimate of the committor probability computed from N trajectories, p_B is the actual committor probability, $P(p_B|q)$ is the distribution of exact committor probabilities on the trial surface $q(\mathbf{x}) = q$, $B_{N,pB}(\hat{p}_B)$ is the binomial distribution with N trials and parameter p_B, and $H(\hat{p}_B|q)$ is the limiting histogram that would be obtained from p_B-estimates at an infinite set of configurations [47]. Fig. 2 depicts the discrete histogram of p_B-estimates and the corresponding continuous distribution of true committor probabilities.

$P(p_B|q)$ can be defined as

$$P(p_B|q) = \int d\mathbf{x}\rho_{EQ}(\mathbf{x}|q)\delta[p_B - p_B(\mathbf{x})] \tag{2}$$

where

$$\rho_{EQ}(\mathbf{x}|q) = \frac{\rho_{EQ}(\mathbf{x})\delta[q(\mathbf{x}) - q]}{\int d\mathbf{x}\rho_{EQ}(\mathbf{x})\delta[q(\mathbf{x}) - q]} \tag{3}$$

Following our convention for the difference between q and $q(\mathbf{x})$, we have similarly used p_B to denote a specific value of the coordinate dependent committor function $p_B(\mathbf{x})$ in Eq. (2).

Also note that configurations from the distribution $\rho_{EQ}(\mathbf{x}|q)$ should be sampled using a stiff restraining potential or a narrow square-well around the desired value of q in a simulation [7]. As explained in Frenkel and Smit, the use of a restraining potential or a narrow delta function approximant is not equivalent to the formal sharp surface integral over the isosurface $q(\mathbf{x}) = q$ [7]. The difference was not emphasized in our earlier work, but it becomes important in relating $p(\mathrm{TP}|q)$ to the distribution of committor probabilities $P(p_B|q)$.

As we have shown previously [47], Eq. (1) can effectively be inverted by relating moments of $H(\hat{p}_B|q)$ to moments of $P(p_B|q)$:

$$\mu = \mu_H \tag{4}$$

and

$$\sigma^2 = \frac{N\sigma_H^2}{N-1} - \frac{\mu_H(1-\mu_H)}{N-1} \tag{5}$$

Here μ_H and σ_H^2 are the mean and variance of the histogram and μ and σ^2 are the mean and variance of the true distribution. These are properties of the coordinate isosurface being tested. Unlike σ_H and μ_H, σ and μ are protocol-independent in the sense that they no longer depend on the number of trajectories per p_B-estimate used in the shooting procedure [47]. Eqs. (4) and (5) thus quantify the true distribution of p_B-values on the isosurface $q(\mathbf{x}) = q$ as $p_B = \mu \pm \sigma$.

The values of μ and σ can typically be computed accurately with less than 1/10th of the trajectories used in a typical committor analysis [47]. Fig. 3 shows a two-parameter beta distribution model of $P(p_B|q)$ with mean $\mu = 1/2$ (i.e. for a trial separatrix), and with different values of the standard deviation σ. The beta distribution model for $P(p_B|q)$ begins to resemble a peak around $\mu = 1/2$ for $\sigma < 0.15$, thus providing a minimal standard for separatrix accuracy.

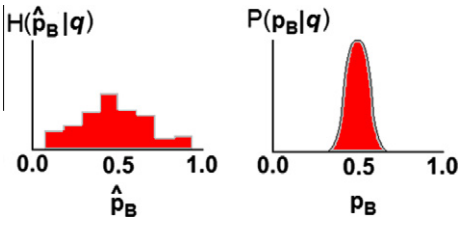

Fig. 2. The histogram of noisy p_B-estimates from an isosurface $q(\mathbf{x}) = q$ is a discrete distribution. In contrast, the true distribution $P(p_B|q)$ is continuous and always narrower than the corresponding histogram because of Eq. (1).

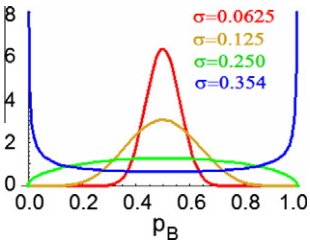

Fig. 3. Beta distribution model of $P(p_B|q)$ for an isosurface $q(\mathbf{x}) = q$ with $\mu = 1/2$ showing different values of the variance σ^2.

4. Proof that $p(\mathrm{TP}|q)$ peak maximization is separatrix optimization

Hummer [48] has given the conditional probability of transition paths after projection to a specific isosurface $q(\mathbf{x}) = q$ as

$$p(\mathrm{TP}|q) = \int d\mathbf{x}\rho_{EQ}(\mathbf{x}|q)p(\mathrm{TP}|\mathbf{x}) \tag{6}$$

where $p(\mathrm{TP}|\mathbf{x})$ is the probability that a trajectory initiated from $\mathbf{x}$ with initial momenta from a Boltzmann distribution will generate a transition path [48]. Shooting procedures can evaluate $p(\mathrm{TP}|q)$ for any type of coupling to the bath modes, but for diffusive dynamics $p(\mathrm{TP}|q)$ can be further simplified [48]. For diffusive dynamics the destinations in forward and backward time for a trajectory initiated at $\mathbf{x}$ are uncorrelated with each other, so again following Hummer [48],

$$p(\mathrm{TP}|\mathbf{x}) = 2p_B(\mathbf{x})(1 - p_B(\mathbf{x})). \tag{7}$$

Using Eqs. (7) and (2) in Eq. (6) gives the central result of this letter

$$\begin{aligned} p(\mathrm{TP}|q) &= 2\int d\mathbf{x}\,\rho_{EQ}(\mathbf{x}|q)p_B(\mathbf{x})(1 - p_B(\mathbf{x})) \\ &= 2\int d\mathbf{x}\,\rho_{EQ}(\mathbf{x}|q)\int_0^1 dp_B\,p_B(1-p_B)\delta[p_B - p_B(\mathbf{x})] \\ &= 2\int_0^1 dp_B p_B(1-p_B)\int d\mathbf{x}\,\rho_{EQ}(\mathbf{x}|q)\delta[p_B - p_B(\mathbf{x})] \\ &= 2\int_0^1 dp_B P(p_B|q)p_B(1-p_B) \\ &= 2\{\mu(1-\mu) - \sigma^2\} \end{aligned} \tag{8}$$

The second line of Eq. (8) again uses the convention that p_B and $p_B(\mathbf{x})$ refer respectively to one value of p_B and to the function $p_B(\mathbf{x})$ which may or may not be equal depending on the argument $\mathbf{x}$. The last two equalities use Eq. (2) and definitions from our earlier work [47] on the committor distribution and its moments. The final equality shows that $p(\mathrm{TP}|q)$ is maximized when $\mu(1-\mu)$ is as large as possible and when σ^2 is as small as possible. Thus maximizing $p(\mathrm{TP}|q)$ over coordinates $q(\mathbf{x})$ and isosurfaces q is equivalent to seeking an isosurface with $\mu = 1/2$ and with the narrowest possible distribution of committor probabilities. Assuming the maximum in $p(\mathrm{TP}|q)$ for the optimized coordinate corresponds to an isosurface with $\mu = 1/2$, we can estimate the committor distribution width on the separatrix. Solving $p(\mathrm{TP}|q)_{\max} = 2(1/4 - \sigma^2)$ with the peak maxima reported by Best and Hummer [22,27], we find $\sigma = 0.23$ from the three-helix bundle study, and an accurate transition state ensemble ($\sigma = 0.12$) from the protein G study. These values illustrate the use of Eqs. (4), (5), and (8) to quantitatively compare accuracy of results obtained by different shooting protocols [47].

Eq. (8) shows that maximizing the peak in $p(\mathrm{TP}|q)$ is equivalent to the original trial-and-error test for reaction coordinate accuracy. However, the $p(\mathrm{TP}|q)$ peak maximization method can *only* test the reaction coordinate at the separatrix, whereas the original committor analysis can (in principle) also test reaction coordinate accuracy at earlier and later stages. We note that rigorous and efficient alternative methods are now available. These include several methods to identify low dimensional collective variable reaction coordi-

nates [13,16–18] and several methods that bypass the need for low dimensional reaction coordinates [25,35–42].

5. Conclusions

Reaction coordinates that remain accurate at all stages of the reaction pathway enable accurate models of dynamics along the reaction pathway and free energy calculations that converge accurately without hysteresis. Methods that optimize the reaction coordinate only at the separatrix may identify reaction coordinates that are inaccurate at earlier and later stages. For the purely diffusive dynamics characteristic of protein folding and nucleation, we used moments of the true committor distribution [47] to prove that maximizing the peak in the transition path probability distribution $p(\mathrm{TP}|q)$ is equivalent to a separatrix optimization. Thus the $p(\mathrm{TP}|q)$ peak maximization method [22,27] does ensure an accurate separatrix, but does not ensure an accurate reaction coordinate for earlier and later stages of the reaction pathway. In systems with long reaction pathways and many stable intermediates, coordinates obtained by $p(\mathrm{TP}|q)$ peak maximization may result in hysteretic free energy landscapes or inaccurate coordinate dependent diffusion models. For these applications, alternative methods that optimize the reaction coordinate over the entire reaction pathway [13,16–18], or methods that circumvent the projection onto specific coordinates are recommended [25,35–42].

Acknowledgements

This work was supported by an NSF CHE CAREER Grant (0955502). Baron thanks Nathan Duff, Gregg Beckham, and the reviewers for helpful comments.

References

[1] P. Hanggi, P. Talkner, M. Borkovec, Rev. Mod. Phys. 62 (1990) 251.
[2] J. Keck, Adv. Chem. Phys. 13 (1967) 85.
[3] T.S. van Erp, J. Chem. Phys. 125 (2006) 20.
[4] B.C. Garrett, D.G. Truhlar, Annu. Rev. Phys. Chem. 35 (1984) 159.
[5] D. Chandler, J. Chem. Phys. 68 (1978) 2959.
[6] E. Vanden-Eijnden, F.A. Tal, J. Chem. Phys. 123 (2005) 184103.
[7] D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications, Academic Press, San Diego, 2002.
[8] L. Onsager, Phys. Rev. 54 (1938) 554.
[9] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, E.S. Shakhnovich, J. Chem. Phys. 108 (1998) 334.
[10] P.G. Geissler, C. Dellago, D. Chandler, J. Phys. Chem. B 103 (1999) 3706.
[11] A. Berezhkovskii, A. Szabo, J. Chem. Phys. 122 (2005) 014503.
[12] W. E, W. Ren, E. Vanden-Eijnden, Chem. Phys. Lett. 413 (2005) 242.
[13] A. Ma, A.R. Dinner, J. Phys. Chem. B 109 (2005) 6769.
[14] C.D. Snow, Y.M. Rhee, V.S. Pande, Biophys. J. 91 (2006) 14.
[15] P.G. Bolhuis, D. Chandler, C. Dellago, P.G. Geissler, Annu. Rev. Phys. Chem. 53 (2002) 291.
[16] B. Peters, B.L. Trout, J. Chem. Phys. 125 (2006) 054108.
[17] B. Peters, G.T. Beckham, B.L. Trout, J. Chem. Phys. 127 (2007) 034109.
[18] E.E. Borrero, F.A. Escobedo, J. Chem. Phys. 127 (2007) 164101.
[19] L. Maragliano, E. Vanden-Eijnden, B. Roux, J. Chem. Theory Comput. 5 (2009) 2589.
[20] A. Ma, A. Nag, A.R. Dinner, J. Chem. Phys. 124 (2006) 144911.
[21] B. Knott, N.C. Duff, M.F. Doherty, B. Peters, J. Chem. Phys. 131 (2009) 224112.
[22] R. Best, G. Hummer, Proc. Natl. Acad. Sci. USA 107 (2010) 1088.
[23] L. Rekvig, D. Frenkel, J. Chem. Phys. 127 (2007) 2007.
[24] P.G. Bolhuis, C. Dellago, D. Chandler, Proc. Natl. Acad. Sci. USA 97 (2000) 5877.
[25] F. Noe, C. Schutte, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Proc. Natl. Acad. Sci. USA 106 (2009) 19011.
[26] D. Antoniou, S.D. Schwartz, J. Chem. Phys. 130 (2009) 151103.
[27] R.B. Best, G. Hummer, Proc. Natl. Acad. Sci. USA 102 (2005) 6732.
[28] R. Radhakrishnan, B.L. Trout, J. Am. Chem. Soc. 125 (2003) 7743.
[29] M.F. Hagan, A.R. Dinner, D. Chandler, A.K. Chakraborty, Proc. Natl. Acad. Sci. USA 100 (2003) 13922.
[30] D. Moroni, P.R. ten Wolde, P.G. Bolhuis, Phys. Rev. Lett. 94 (2005) 4.
[31] Y.A. Mantz, D. Branduardi, G. Bussi, M. Parrinello, J. Phys. Chem. B 113 (2009) 12521.
[32] G.T. Beckham, B. Peters, B.L. Trout, J. Phys. Chem. B 112 (2008) 7460.
[33] J. Gsponer, A. Caflisch, Proc. Natl. Acad. Sci. USA 99 (2002) 6719.
[34] R.J. Allen, P.B. Warren, P.R. ten Wolde, Phys. Rev. Lett. 94 (2005) 4.
[35] L. Maragliano, A. Fischer, E. Vanden-Eijnden, G. Ciccotti, J. Chem. Phys. 125 (2006) 15.
[36] P.G. Bolhuis, personal communication 2010.
[37] D.J. Wales, J. Chem. Phys. 130 (2009) 7.
[38] A. Berezhkovskii, G. Hummer, A. Szabo, J. Chem. Phys. 130 (2009) 205102.
[39] S. Muff, A. Caflisch, J. Chem. Phys. 130 (2009) 125104.
[40] S.V. Krivov, M. Karplus, J. Phys. Chem. B 110 (2006) 12689.
[41] J.D. Chodera, N. Singhal, V.S. Pande, K.A. Dill, W.C. Swope, J. Chem. Phys. 126 (2007) 155101.
[42] F. Noe, I. Horenko, C. Schutte, J.C. Smith, J. Chem. Phys. 126 (2007) 155102.
[43] S.S. Cho, Y. Levy, P.G. Wolynes, Proc. Natl. Acad. Sci. USA 103 (2006) 586.
[44] B. Peters, N.E.R. Zimmerman, G.T. Beckham, J.W. Tester, B.L. Trout, J. Am. Chem. Soc. 130 (2008) 17342.
[45] G.A. Korn, T.M. Korn, Mathematical Handbook for Scientists and Engineers, Dover Publications, Mineola, NY, 2000.
[46] K.A. Dill, B.S. Ozkan, M.S. Shell, T.R. Weikl, Annu. Rev. Biophys. 37 (2008) 289.
[47] B. Peters, J. Chem. Phys. 125 (2006) 241101.
[48] G. Hummer, J. Chem. Phys. 120 (2004) 516.