PHYSICAL REVIEW E **77**, 1(R) (2008)

Peptide folding kinetics from replica exchange molecular dynamics

Nicolae-Viorel Buchete* and Gerhard Hummer†

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Building 5, Bethesda, Maryland 20892-0520, USA

(Received 29 October 2007)

We show how accurate kinetic information, such as the rates of protein folding and unfolding, can be extracted from replica-exchange molecular dynamics (REMD) simulations. From the brief and continuous trajectory segments between replica exchanges, we estimate short-time propagators in conformation space and use them to construct a master equation. For a helical peptide in explicit water, we determine the rates of transitions both locally between microscopic conformational states and globally for folding and unfolding. We show that accurate rates in the $\sim 1/(100\ \mathrm{ns})$ to $\sim 1/(1\ \mathrm{ns})$ range can be obtained from REMD with exchange times of 5 ps, in excellent agreement with results from long equilibrium molecular dynamics.

DOI: XXXX PACS number(s): 87.15.A−

Replica exchange molecular dynamics (REMD) [1,2] is a powerful method to enhance the conformational sampling, addressing a serious challenge in molecular simulations [3]. Multiple noninteracting copies (or "replicas") of the system are simulated in parallel, each at a different temperature. To transfer the barrier-crossing efficiency from runs at high temperature to those at low temperature, configuration exchanges are attempted periodically (e.g., at time intervals δt_{REMD}) between replicas at different temperatures (T_i and T_j). Those exchange attempts are accepted with a Metropolis probability $P_{\mathrm{REMD}}(i \leftrightarrow j) = \min\{1, \exp[(\beta_j - \beta_i)(U_j - U_i)]\}$ that enforces detailed balance and maintains canonical distributions at each temperature [with U_i the potential energy of the ith replica, $\beta_i = 1/(k_B T_i)$, and k_B the Boltzmann constant]. After an accepted exchange, the particle velocities are appropriately rescaled to the new temperature, or redrawn from respective Maxwell-Boltzmann distributions. Through a series of exchanges, high-temperature conformations are transferred occasionally to low temperature runs, facilitating the exploration of new configuration-space regions.

While enhancing the exploration of conformation space, REMD apparently does not permit the extraction of useful kinetic information. Conformation exchanges result in discontinuous trajectories, precluding the calculation of equilibrium time correlation functions for times longer than the exchange time δt_{REMD}. To improve the sampling efficiency of REMD, the shortest possible δt_{REMD} should be used [4]. With δt_{REMD} much shorter than the time scales of slow conformational changes, the rates of conformational changes appear inaccessible to REMD simulations. To overcome this problem, at least for the special case of a two-state system, an indirect method has recently been proposed in which the two rate coefficients describing the assumed folding or unfolding dynamics are assumed to obey an Arrhenius temperature dependence [5]. However, the protein-folding rate often exhibits non-Arrhenius temperature dependence [6], and folding intermediates are common. To avoid the resulting problems, master-equation approaches have been described by Levy and co-workers [7] in a qualitative, yet insightful analysis. As a quantitative alternative, REMD has recently been used to estimate the local drift and diffusion coefficients [8] within the framework of coarse diffusion equations [9–11].

Here we show how one can efficiently extract accurate transition rates from REMD simulations, both locally between microscopic conformational states and globally between folded and unfolded conformations (and possible intermediates), without the assumption of a certain temperature dependence of the underlying kinetics. In fact, our method can be used to investigate the Arrhenius or non-Arrhenius character of a particular system. We determine short-time propagators in conformation space to overcome the problems arising from the intrinsically discontinuous character of REMD trajectories [12,13].

We first realize that REMD permits the accurate (and formally exact) calculation of short-time correlation functions. The initial configurations after a replica exchange (with appropriate velocity assignment) constitute valid representatives of the equilibrium phase-space distributions at the respective temperatures. From the subsequent Hamiltonian dynamics until the next exchange, we can obtain exact correlation functions. The maximum time scale will be a few δt_{REMD}, given by the longest time between accepted replica exchanges.

Specifically, we here determine the frequency of transitions between conformational states. From the observed molecular transitions, we construct a master equation describing the dynamics in a conformation space divided into N distinct states. We later verify that the dynamics in the resulting projected space is captured by a master equation, $dP_i(t)/dt = \sum_{j=1}^{N} [k_{ij} P_j(t) - k_{ji} P_i(t)]$, where $P_i(t)$ is the population in state i, and $k_{ij} \geq 0$ is the transition rate from j to $i \neq j$. In vector-matrix notation, we have $d\boldsymbol{P}(t)/dt = \boldsymbol{K}\boldsymbol{P}(t)$, where the $N \times N$ rate matrix $\boldsymbol{K}$ has off-diagonal elements k_{ij} and diagonal elements $k_{ii} = -\sum_{j \neq i} k_{ji} < 0$. The propagators, defined as the probability of being in state j at time t given that the system was in state i at time 0, can be written in terms of the matrix exponential, $p(j,t|i,0) = [\exp(\boldsymbol{K}t)]_{ji}$. To estimate the elements of the rate matrix $\boldsymbol{K}$ from either long equilibrium

*Present address: School of Physics, University College Dublin, Belfield, Dublin 4, Ireland.

†gerhard.hummer@nih.gov

simulations or REMD, we use a maximum-likelihood procedure. We first determine the number N_{ji} of transitions from state i to state j within a time interval Δt, irrespective of intermediate states. The log-likelihood of observing transition numbers N_{ji} is [12,13]

$$\ln \mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{N} N_{ji} \ln p(j, \Delta t | i, 0). \quad (1)$$

To obtain the rate coefficients of the master equation (with upper and lower diagonal elements related by detailed balance), we maximize $\ln \mathcal{L}$ with respect to ~~the~~ k_{ij} [12,13].

Effects of non-Markovian dynamics not captured by the master equation result in a dependence of the rate matrix on the time interval Δt. Ultimately, for long lag times Δt, fast non-Markovian dynamics is effectively suppressed and the propagators are dominated by the slow transitions [11–13]. However, if Δt is short, fast motions lead to improper assignments of conformational states. As a consequence, the extracted rate matrices tend to predict overly fast conformational relaxation.

The problem of fast non-Markovian dynamics can be suppressed by assigning the states with the help of transition paths that connect well-defined regions within two conformational cells [Figs. 1(a) and 1(b)]. A new state is assigned only if the trajectory crosses from one well-defined region to another. Fast equilibrium fluctuations in the projected space thus do not lead to a state change. We showed previously that for peptide folding in long standard molecular dynamics (MD) simulations, this procedure gives accurate rate matrices for observation times Δt as short as 1 ps [13].

Here, we adapt this state-assignment procedure to REMD. In a first step, we follow each replica irrespective of exchanges, and identify transition paths for these continuous trajectories to assign states. In a second step, transition numbers N_{ji} for each of the REMD temperatures are determined from the respective short trajectory segments uninterrupted by replica exchange. From the N_{ji}, we then estimate the coefficients of the master equation through likelihood maximization.

In the following, we demonstrate the general procedure to calculate slow rates from REMD with fast exchange. Master-equation approaches have been used extensively in peptide folding studies [7,12,14–18]. We used the GROMACS 3.3 package [19] to run both standard MD and REMD simulations for the folding of a short helical peptide, blocked Ala_5 (i.e., CH3CO-Ala5-NHCH3), in explicit water [20,21]. We used the AMBER-GSS force field [22] ported to GROMACS [23], with peptide (Φ, Ψ) torsional potentials modified to reproduce experimental helix-coil equilibria [2]. Simulation details can be found in Ref. [13].

Four independent MD and REMD runs were initiated from different configurations (11111—"all helix," 00000—"all coil," 01010, and 10101, where 1 denotes a residue in the helical region of the Ramachandran map, ordered left to right from N to C terminus [13]). The reference MD runs covered 250 ns at two different temperatures (300 and 350 K), for a total combined simulation time of 2 μs. The 150-ns REMD simulations used 12 replicas spanning the

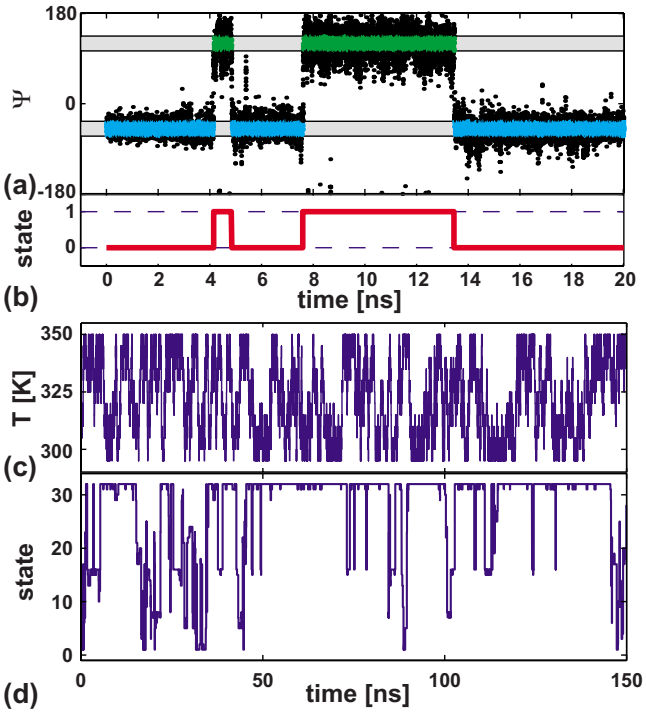

FIG. 1. (Color online) REMD simulations. (a) Schematic of transition-path based assignment of conformational states, shown for illustrative purposes in 1D (with the actual assignment done using both Φ and Ψ [13]). The backbone dihedral angle Ψ of alanine exhibits transitions between helical (blue, $\Psi \approx -50°$) and non-helical states (green, $\Psi \approx 120°$). Conformations within narrow regions around the two free energy minima (gray) can be assigned as helical or coil with high confidence. For other conformations (black dots), the assigned state changes only if the trajectory crosses between the well-defined regions, but not on equilibrium excursions that revert without actual crossing. (b) State assignment corresponding to (a). (c) Temperatures sampled by a typical Ala5 replica during a 150 ns REMD simulation. (d) Conformational states sampled by the Ala5 replica during the same run.

AQ: #1

295–350 K temperature range for a total combined simulation time of 600 ns per replica. Coordinates were saved every 1 ps and REMD exchanges were attempted every δt_{REMD}=5 ps. Figure 1(c) shows that the resulting REMD trajectories pass through the whole range of temperatures multiple times. Each individual trajectory also has a high likelihood to visit most, if not all, of the 32 coarse-grained conformational states [Fig. 1(d); with 00000 and 11111 corresponding to states 1 and 32 in binary notation plus 1]. In the resulting master equation model, the transition rates k_{ij} are different from zero only if states i and j in binary notation differ by at most one bit, producing the connectivity of a five-dimensional hypercube.

Figure 2 shows the equilibrium populations in each of the 32 conformational states at 300 and 350 K from REMD trajectories. The inset illustrates the excellent agreement between equilibrium distributions from MD and REMD at 300 K. At 350 K, the sampling is more efficient and the agreement even better (data not shown).

Figure 3 demonstrates that the master equation accurately captures the kinetics. Shown are the two slowest relaxation

ln k_{ij} (i>j) and ln p_i, where p_i is the equilibrium population of state i.

insert space

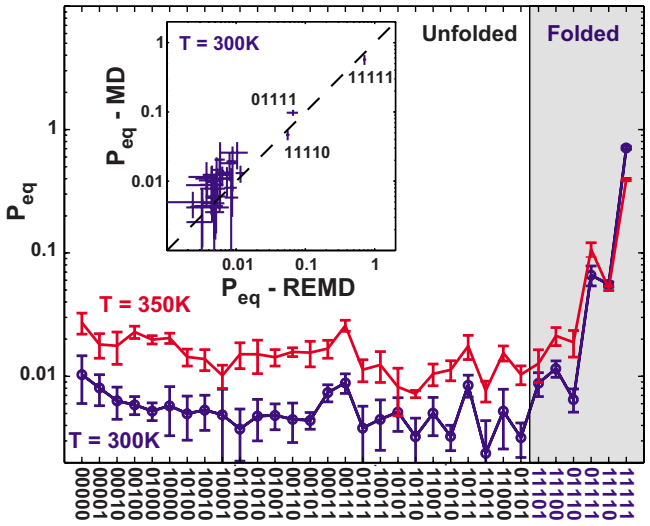

FIG. 2. (Color online) REMD equilibrium populations P_{eq} at 300 and 350 K. Shading indicates the folded basin. (Inset) Scatter plot of P_{eq} from standard MD and REMD at 300 K. Error bars indicate standard deviations of the mean.

times, τ_2 and τ_3, at the 12 temperatures sampled in the REMD runs (where $\tau_i = -1/\lambda_i$, with λ_i the ordered eigenvalues of $\boldsymbol{K}$). The REMD relaxation times agree perfectly with those obtained from standard MD runs at 300 and 350 K [13]. This agreement holds also for all relaxation times τ_i (not shown for $i \geq 4$), and the individual coefficients k_{ij} of the master equation, as shown in Fig. 4(a) (with linear correlation coefficients ≥ 0.94).

From the slowest relaxation time τ_2, and the relative populations in the folded (helical) and unfolded (coil) state of the peptide, we estimate folding and unfolding rates as a function of temperature under the assumption of a two-state relaxation (Fig. 3 inset). The 32 states i were assigned as folded or unfolded based on the left-hand eigenvector of $\boldsymbol{K}$ corresponding to eigenvalue λ_2 [13,24] (see Fig. 2). The resulting folded basin consists of all structures with at least one α-helical $(i, i+4)$ backbone hydrogen bond among the four N-terminal residues. Consistent with the assumptions of Ref. [5], we find that the resulting folding and unfolding rates exhibit Arrhenius-like dependence on temperature. The activation energies for folding and unfolding are $E_a^F \approx 22.1$ kJ/mol and $E_a^U \approx 46.5$ kJ/mol.

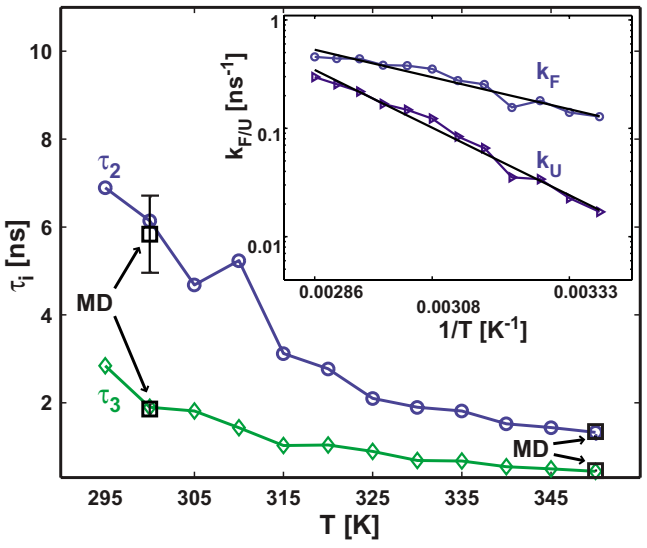

FIG. 3. (Color online) Relaxation times τ_2 (circles, blue) and τ_3 (diamonds, green) as a function of temperature. Open squares show τ_2 and τ_3 from standard MD at 300 and 350 K [13]. (Inset) Folding (k_F) and unfolding (k_U) rate constants as a function of $1/T$, together with Arhenius fits.

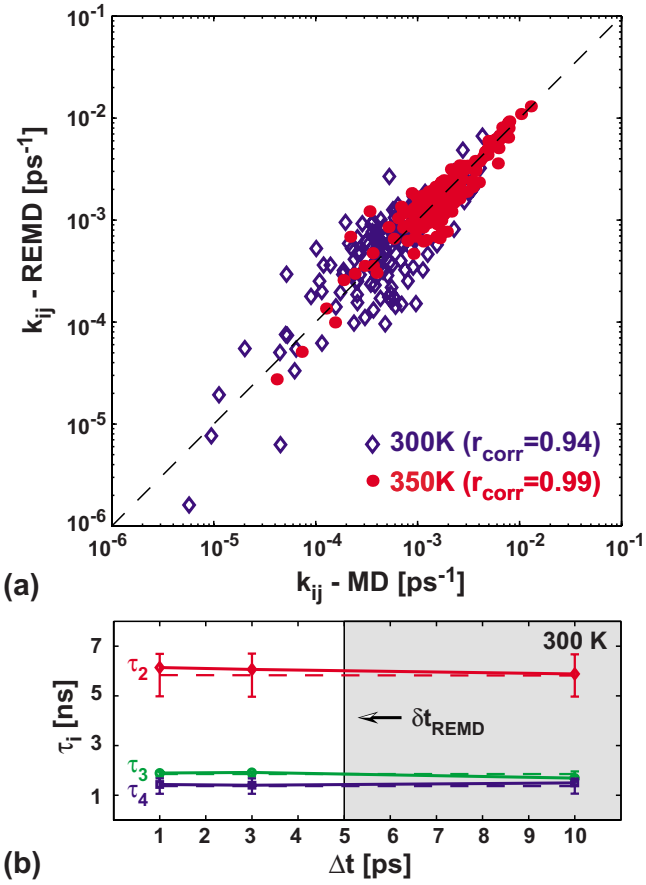

FIG. 4. (Color online) Validation of transition rates estimated from REMD trajectories. (a) Rates k_{ij} from REMD versus those from standard MD [13] at 300 K (diamonds, blue) and 350 K (circles, red). (b) Dependence of the relaxation times τ_2 (top, red), τ_3 (middle, green), and τ_4 (bottom, blue) on the lag time Δt at 300 K. REMD results are shown as symbols connected by solid lines. Reference values from standard MD are shown as dashed lines with error bars. Results for $\Delta t > \delta t_{REMD}$ were obtained from continuous trajectory segments in which replica exchange attempts had been rejected.

A possible concern is the influence of fast non-Markovian dynamics not taken into account by the master equation model. We can explicitly probe for such effects by plotting the calculated relaxation times τ_i as a function of the lag time Δt used to determine the propagators. Figure 4 shows that the relaxation times from REMD are independent of Δt from 1 to 10 ps ($2\delta t_{REMD}$), and agree with the results from standard MD.

We showed how accurate rates for the conformational dynamics of a molecular system can be extracted from REMD simulations. For a short helical peptide in water, the REMD kinetics was in perfect agreement with that from standard MD. The key elements of the procedure are (1) the suppression of non-Markovian noise by using transition paths in the assignment of states, (2) the calculation of transition numbers N_{ij} on the time scale of replica exchanges, and (3) the construction of a master equation from the N_{ij} using a maximum likelihood procedure. The formalism is general, and can be adapted to Hamiltonian REMD [25], resolution exchange [26], non-Boltzmann reservoirs [27], serial replica exchange [28], etc.

In practical applications, such as protein folding, the combinatorial explosion in the number of states poses a major challenge for large systems. To reduce the dimension of the master equation, states could be defined by using conformational clustering [29], subsets of the dihedral-angle coordinates (that produce the most Markovian dynamics), or alternative coordinates such as native or non-native amino-acid contacts or contact fractions, the radius of gyration, or distances between key residues, with our formalism applicable to both discrete and continuous variables [10]. In addition, hierarchical coarse graining [30] can be used to combine fine and coarse-grained master equations [9,12]. As a second challenge, the need to collect sufficient transitions at all temperatures to construct a connected master equation could be overcome by assuming that the individual rates k_{ij}, but not necessarily the slow relaxations τ_i, satisfy an Arrhenius law. In that way, transitions observed at higher temperatures can be used to estimate the relaxation time scales at lower temperatures, augmented by the accurate equilibrium populations of REMD through the requirement of detailed balance. Such a procedure is easily implemented within our likelihood-maximization framework by replacing the individual rates with temperature-independent prefactors and activation energies.

We thank A. Szabo, A. M. Berezhkovskii, E. Rosta, and R. B. Best for many helpful and stimulating discussions. This research used the Biowulf Linux cluster at the NIH and was supported by the Intramural Research Program of the NIDDK, NIH.

[1] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[2] A. E. García and K. Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).
[3] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).
[4] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **104**, 15340 (2007).
[5] D. van der Spoel and M. M. Seibert, Phys. Rev. Lett. **96**, 238102 (2006).
[6] M. Oliveberg, Y. J. Tan, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 8926 (1995).
[7] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **102**, 6801 (2005).
[8] S. C. Yang, J. N. Onuchic, A. E. García, and H. Levine, J. Mol. Biol. **372**, 756 (2007).
[9] G. Hummer and I. G. Kevrekidis, J. Chem. Phys. **118**, 10762 (2003).
[10] G. Hummer, New J. Phys. **7**, 34 (2005).
[11] R. B. Best and G. Hummer, Phys. Rev. Lett. **96**, 228104 (2006).
[12] S. Sriraman, I. G. Kevrekidis, and G. Hummer, J. Phys. Chem. B **109**, 6479 (2005).
[13] N. V. Buchete and G. Hummer, J. Phys. Chem. B (to be published) (http://dx.doi.org/10.1021/jp0761665).
[14] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, J. Comput. Phys. **151**, 146 (1999).
[15] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[16] B. L. de Groot, X. Daura, A. E. Mark, and H. Grubmüller, J. Mol. Biol. **309**, 299 (2001).
[17] O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[18] N. V. Buchete and J. E. Straub, J. Phys. Chem. B **105**, 6684 (2001).
[19] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[20] G. Hummer, A. E. García, and S. Garde, Phys. Rev. Lett. **85**, 2637 (2000).
[21] C. J. Margulis, H. A. Stern, and B. J. Berne, J. Phys. Chem. B **106**, 10748 (2002).
[22] H. Nymeyer and A. E. García, Proc. Natl. Acad. Sci. U.S.A. **100**, 13934 (2003).
[23] E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
[24] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **121**, 9186 (2004).
[25] H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).
[26] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, Phys. Rev. Lett. **96**, 028105 (2006).
[27] A. E. Roitberg, A. Okur, and C. Simmerling, J. Phys. Chem. B **111**, 2415 (2007).
[28] M. Hagen, B. Kim, P. Liu, R. A. Friesner, and B. J. Berne, J. Phys. Chem. B **111**, 1416 (2007).
[29] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, J. Chem. Phys. **126**, 155101 (2007).
[30] F. Noe, I. Horenko, C. Schütte, and J. C. Smith, J. Chem. Phys. **126**, 155102 (2007).

AUTHOR QUERIES —

#1 Please check your figures in both the b/w and the color versions.