Folding proteins by first-passage-times-optimized replica exchange

Walter Nadler,[1,*] Jan H. Meinke,[1,†] and Ulrich H. E. Hansmann[2,1,‡]

[1] *John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany*

[2] *Department of Physics, Michigan Technological University, Houghton, Michigan, 49931 USA*

(Received 6 June 2008; published 3 December 2008)

Replica exchange simulations have become the method of choice in computational protein science, but they still often do not allow an efficient sampling of low-energy protein configurations. Here, we reconstruct replica flow in the temperature ladder from first passage times and use it for temperature optimization, thereby maximizing sampling. The method is applied in simulations of folding thermodynamics for a number of proteins starting from the pentapeptide Met-enkephalin, through the 36-residue HP-36, up to the 67-residue protein GS-α_3W.

DOI: 10.1103/PhysRevE.78.061905 PACS number(s): 87.10.−e, 05.10.Ln, 02.70.Rr, 02.70.Tt

Progress in computational protein studies is still hampered by inefficient sampling at low temperatures. The reason is the inherent roughness of protein energy landscapes leading to barriers and bottlenecks. The resulting slow relaxation can be alleviated by the replica exchange method [1–3]. Monte Carlo or molecular dynamics simulations [4–8] are performed in parallel at different values of a control parameter, most often the temperature. At certain times the current conformations of replicas at neighboring control parameter values are exchanged according to a generalized Metropolis rule [9], and each replica performs a random walk in control parameter space. As convergence is faster than by spending all computer time at one low temperature, replica exchange has become the method of choice for protein simulations.

Even with replica exchange sampling, however, the computational costs of protein simulations can still be prohibitive. This is, in part, because the efficiency of this technique depends strongly on the discretization of control parameter space. Trebst and co-workers [10–12] have shown how sampling can be optimized by maximizing the flow across parameter space. While their work led to a much deeper understanding of the dynamics of the replica exchange technique, its application in protein simulations is still limited: Their schemes rely on a direct analysis of global replica flow, but an accurate measurement of this nonlocal quantity requires computational efforts that are excessively high in most protein simulations. We propose to overcome this difficulty by an alternative approach that relies on measurements of first passage times *within* the temperature ladder. The resulting increased efficiency in optimizing the temperature discretization enables thermal all-atom simulations of larger proteins than previously possible. As an example, we present first results from an ongoing folding study of the 67-residue GS-α_3W.

In this study, we assume replica exchange simulations relying on $N+1$ control parameter values, numbered $n=0,\ldots,N$, which we will call *nodes*. The time evolution of the probability $P(n,t)$ that an individual replica is on node n at time t can be approximated by a master equation in discrete time [13,14],

$$P(n,t+1) = W(\beta_{n-1},\beta_n)[P(n-1,t)-P(n,t)] + W(\beta_n,\beta_{n+1}) \times[P(n+1,t)-P(n,t)] + P(n,t). \quad (1)$$

Here, β_n is the control parameter value used for simulation at node n. The symmetric transition probabilities between neighboring nodes, $W(\beta,\beta')=W(\beta',\beta)$ lead to a constant stationary distribution: $P_0(n)=1/(N+1)$. The flow of replicas across temperature space can be determined from Eq. (1): the probability distribution for the flow from $n=0$ to $n=N$ is the stationary solution of Eq. (1) with the boundary conditions $P_{up}(0)=1$ and $P_{up}(N)=0$, resulting in

$$P_{up}(n) = \left[1 - J\sum_{i=0}^{n-1} \frac{1}{W(\beta_i,\beta_{i+1})}\right], \quad (2)$$

and a similar form for P_{down}. In an actual simulation P_{up} and P_{down} are estimated by measuring the fraction of replicas moving up,

$$f_{up}(n) = \frac{z_{up}(n)}{z_{up}(n)+z_{down}(n)}, \quad (3)$$

and a corresponding quantity for those moving down, $f_{down}(n)=1-f_{up}(n)$. Here, $z_{up}(n)$ [$z_{down}(n)$] is the number of visits at node n by replicas that came from node 0 (N). In the following we will measure flow using f_{up} and denote it by f_{mea}.

In order to optimize sampling one has to maximize the total current J, which is given by

$$J = \left[\sum_{i=0}^{N-1} \frac{1}{W(\beta_i,\beta_{i+1})}\right]^{-1} \quad (4)$$

and also serves as normalization constant in Eq. (2). The control parameter (i.e., temperature) set $\{\beta_0^{(opt)},\ldots,\beta_N^{(opt)}\}$ that maximizes J leads to linear flow distributions [13],

$$P_{up}^{(opt)}(n) = 1 - n/N \quad \text{and } P_{down}^{(opt)}(n) = n/N. \quad (5)$$

The corresponding optimized transition probabilities are constant, $W^{(opt)}(\beta_i,\beta_{i+1})=\text{const.}$

[*] w.nadler@fz-juelich.de

[†] j.meinke@fz-juelich.de

[‡] hansmann@mtu.edu

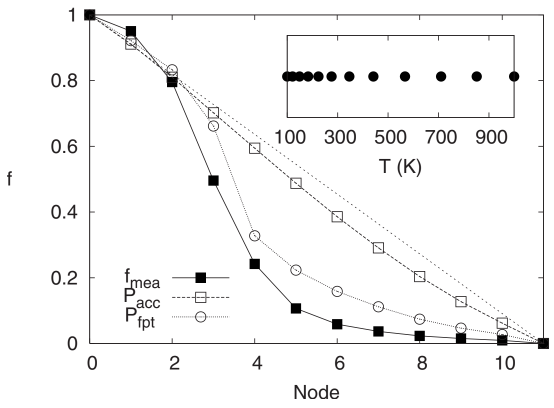

FIG. 1. Flow distribution for a replica-exchange Monte Carlo simulation of Met-Enkephalin: (f_{mea}) from direct observations, Eq. (3), (P_{acc}) reconstructed using Eq. (2) with effective transition probabilities approximated by observed acceptance rates, and (P_{fpt}) using measured first passage times, Eq. (11). The underlying temperature set is given in the inset. The linear curve indicates the ideal distribution, Eq. (5).

We emphasize that usually a temperature set with equal *acceptance rates* will not lead to an optimized flow. Take as an example Fig. 1. Here we show the flow for an implicit-solvent replica-exchange Monte Carlo simulation of the pentapeptide Met-Enkephalin, using 12 replicas with 10^6 sweeps over all degrees of freedom. An exchange attempt is made every ten sweeps. The temperature set (drawn in the inset) was chosen to yield approximately equal transition rates. The *hypothetical flow* that would be observed if the transition probabilities in Eq. (1) were given by the observed acceptance rates is denoted by P_{acc}, and is close to the ideal linear form, Eq. (5). However, even for this simple molecule we observe a strong deviation of the *measured flow*, denoted by f_{mea}, from P_{acc} as well as from the ideal linear form, indicating that the temperatures are far from optimal. This deviation is due to the difference between the observed acceptance rates and the effective transition probabilities entering the master equation (1). Due to broken ergodicity the random walk of replicas can have a hierarchical, treelike structure, and observed acceptance rates as well as flow distributions are projections onto a one-dimensional walk [13]. Equivalence between both would require fast relaxation at all nodes [13,15], which is never the case in protein simulations. Instead, the *effective transition probabilities* describing long-time properties like replica flow are usually different from observed acceptance rates, which describe only short-time properties. Such discrepancies between short-time and long-time properties of stochastic processes are well known [16,17].

A direct measurement of the flow distribution is computationally costly: Individual replicas have to cross the full ladder of nodes many times in order to ensure sufficient statistics; see Eq. (3). The scarcity of such "tunneling" events is a problem particularly at the beginning of the control parameter optimization when round trip times are largest. In order to alleviate this problem we propose to estimate the flow distribution from measurements of mean first passage times [14,18] *within* the temperature ladder. First passage times effectively describe the long-time properties of stochastic processes [16,17] and they incorporate correlations that are not covered by short time properties like observed acceptance probabilities. This approach does not require tunneling of replicas over the whole control parameter range, as global flow can be approximated from observing mean first passage times of replicas crossing only part of it.

A first passage time between nodes n and n' is the time between a particular replica's first encounter with node n and its consecutive first encounter with node n'. The mean first passage time is the average over all such events. For the master equation (1) it is given by [13]

$$\tau(n \to n') = \sum_{i=n}^{n'-1} \frac{1}{P_0(i)W(\beta_i,\beta_{i+1})} \sum_{j=0}^{i} P_0(j), \tag{6}$$

with $n' > n$, and an equivalent form for $n' < n$. We will concentrate on mean first passage times between the lower boundary node 0 and inner nodes n:

$$\tau(0 \to n) = \tau(0 \to n-1) + \frac{n}{W(\beta_{n-1},\beta_n)}, \tag{7}$$

as well as on those between the upper boundary node N and the inner nodes n:

$$\tau(N \to n) = \tau(N \to n+1) + \frac{N-n}{W(\beta_{n+1},\beta_n)}. \tag{8}$$

with $\tau(0 \to 0) \equiv 0 \equiv \tau(N \to N)$. These two relations can be employed to determine the effective transition probabilities entering Eq. (1).

For $\tau(0 \to n)$, the number of first passage events in a simulation decreases with n, while the error increases. Similarly, the error for $\tau(N \to n)$ increases with decreasing n. There exists a node n^* so that the error of $\tau(0 \to n)$ is still smaller than that of $\tau(N \to n)$, while this relation changes for node n^*+1. Therefore the mean first passage times for $\tau(0 \to n)$, $n=1,\ldots,n^*$, and those for $\tau(N \to n)$, $n=n^*+1,\ldots,N-1$, will be the most reliable ones. Hence while in the optimization schemes of Refs. [12,13] the limiting factors are tunneling events across the full ladder of temperatures, here the statistics is only limited by the number of first passage events from either boundary to n^*. Such events occur even in the absence of tunneling events, and their number can be orders of magnitude larger.

In order to simplify our formalism, we introduce sums over adjacent inverse transition probabilities:

$$h(0 \to n) = \sum_{j=1}^{n} \frac{1}{W(\beta_{j-1},\beta_j)} = \sum_{j=1}^{n-1} \frac{\tau(0 \to j)}{j(j+1)} + \frac{\tau(0 \to n)}{n}, \tag{9}$$

and

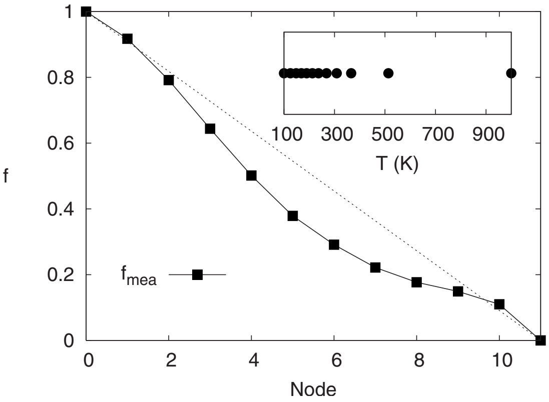

FIG. 2. Flow distribution for an implicit-solvent replica exchange Monte Carlo simulation of Met-Enkephalin. The mean-first-passage-times optimized temperatures are given in the inset and were obtained after *one* iteration. As in Fig. 1, the linear curve indicates the ideal distribution, Eq. (5).

$$h(N \to n) = \sum_{j=n+1}^{N} \frac{1}{W(\beta_j, \beta_{j-1})}$$
$$= \sum_{j=1}^{N-n-1} \frac{\tau(N \to N-j)}{j(j+1)} + \frac{\tau(N \to n)}{N-n}. \tag{10}$$

Using these auxiliary functions, we obtain the following expressions for the flow probabilities:

$$P_{up}^{(MFPT)}(n) = \begin{cases} 1 - \dfrac{h(0 \to n)}{h(0 \to n^*) + h(N \to n^*)}: & n \leqslant n^* \\ \dfrac{h(N \to n)}{h(0 \to n^*) + h(N \to n^*)}: & n > n^* \end{cases} \tag{11}$$

with a similar relation for $P_{down}^{(MFPT)}$. We will abbreviate $P_{up}^{(MFPT)}$ by P_{fpt} in the following. Fig. 1 displays also this quantity for the above described Met-enkephalin simulation, with the temperatures chosen for equal acceptance rates. P_{fpt} follows closely the measured flow distribution f_{mea} of Eq. (3). Differences between the curves are due to sampling variations.

Starting from a flow distribution P_{fpt} reconstructed from mean first passage time analysis, one can now use existing iteration schemes that exhibits fast convergence to the optimal temperature values [11–13]. We found that flow distributions P_{fpt} derived from mean first passage times lead to temperature sets that are more stable upon iteration than those from flows measured directly by way of Eq. (3). For the example of Met-Enkephalin, we show in Fig. 2 the measured flow distribution f_{mea} for a temperature discretization that results from a single iteration based on P_{fpt} of Fig. 1. The deviations from the ideal case (the linear line) are already minimal.

Progress in computational protein science has gone far beyond tiny peptides such as Met-enkephalin. We chose that molecule solely to demonstrate that already for such simple systems equal acceptance rates do not lead to an optimal flow. To test our approach for a more realistic example, we have selected the 36-residue protein HP-36 in the same implicit solvent as in Ref. [12], where this protein was used to introduce flow analysis to replica exchange simulations of biomolecules. HP-36 has a sufficiently complicated structure to serve as a generic instance of a globular protein but is small enough to be numerically accessible. For this reason it has become an often used toy model to study simulation techniques.

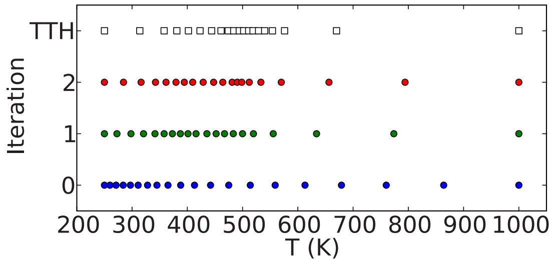

FIG. 3. (Color online) Optimizing the temperature set of a 20-replica implicit-solvent simulation of HP36. TTH is the final temperature discretization of Ref. [12].

The series of two iterations with a total statistics of 2×10^5 sweeps, and replica-swap attempts every ten sweeps, is displayed in Fig. 3. For comparison, we show also the optimal temperature set of Ref. [12] that relied on a total of 7×10^5 sweeps; note the differences. The estimated flow for a replica walking from node 0 to node N, and back, and therefore the sampling efficiency, is already larger than for the temperature set of Ref. [12]. This is not surprising as the latter was determined from direct measurements of the flow distribution f_{mea}, a nonlocal quantity that is difficult to determine in a simulation.

The improved sampling of the 36-residue HP-36 has given us confidence in our technique. For this reason, we decided to apply it to GS-α_3W (PDB-code 1LQ7 [19,20]), a three-helix bundle with a single tryptophan buried in the interior of the protein. It was designed to study the creation

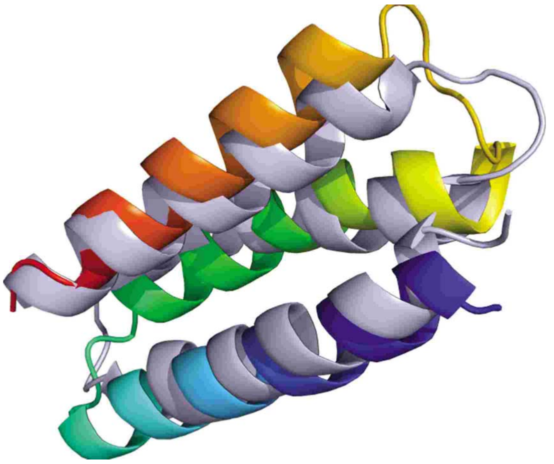

FIG. 4. (Color online) Overlap of the lowest energy configuration (gray) of the 67-residue GS-α_3W with the experimentally determined structure (PDB–code:1LQ7). The root-mean square deviation over heavy atoms between both configurations is 3.3 Å.

and maintenance of a tryptophanyl radical, and therefore serves as a simple model for the function of redox proteins [20]. With 67 residues the protein is of a size that far exceeded what we could study previously in physics-based, thermal all-atom simulations, and previous computational investigations of the protein relied on coarse-grained models only [21]. In Fig. 4 we show the lowest energy structure found so far in an ongoing implicit-solvent replica exchange Monte Carlo simulation relying on mean first-passage-times-optimized temperatures [22]. We used 32 replicas, all starting from a stretched initial configuration, and followed, so far, over 500 000 sweeps. The figure shows the overlap of the minimal structure with 1LQ7. The root mean square deviation is 3.3 Å and the tm score [23] equals 0.5674. A detailed analysis of our data will be published later, but our results do already indicate that replica exchange with flow-optimized temperatures allows thermal folding simulations of such large proteins. To our best knowledge, this is the first time that a protein of this size and complexity has been successfully and reproducibly folded from first principles in unbiased thermal all-atom simulations.

In summary, we have presented a technique employing observed mean first passage times within the control parameter ladder to speed up the flow across the control parameter space, leading to faster sampling of low-energy protein configurations. We have demonstrated the working of this method for two toy models, the pentapeptide Met-Enkephalin and the 36-residue HP-36. Results from preliminary folding simulations of the 67-residue GS-α_3W, a protein of a size previously not accessible to thermal all-atom folding simulations, demonstrate the full power of our approach. We mention in passing that this method can be applied straightforwardly to generalized-ensemble simulations [4], where a weight function has to be determined that optimizes the flow across order parameter space.

This work was supported, in part, by Research Grant No. CHE-0809002 of the National Science Foundation.

[1] C. J. Geyer and A. Thompson, J. Am. Stat. Assoc. **90**, 909 (1995).

[2] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).

[3] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[4] U. H. E. Hansmann and Y. Okamoto, Phys. Rev. E **56**, 2228 (1997).

[5] T. Herges and W. Wenzel, Phys. Rev. Lett. **94**, 018101 (2005).

[6] H. Li, G. Li, B. A. Berg, and W. Yang, J. Chem. Phys. **125**, 144902 (2006).

[7] M. Nanias, C. Czaplewski, and H. A. Scheraga, J. Chem. Theory Comput. **2**, 513 (2006).

[8] D. Gront and A. Kolinski, J. Phys.: Condens. Matter **19**, 036225 (2007).

[9] N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953).

[10] S. Trebst, D. A. Huse, and M. Troyer, Phys. Rev. E **70**, 046701 (2004).

[11] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, J. Stat. Mech.: Theory Exp. (2006), P03018.

[12] S. Trebst, M. Troyer, and U. H. E. Hansmann, J. Chem. Phys. **124**, 174903 (2006).

[13] W. Nadler and U. H. E. Hansmann, Phys. Rev. E **75**, 026109 (2007).

[14] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, (Springer, Berlin, 1985).

[15] C. Predescu, M. Predescu, and C. Ciobanu, J. Phys. Chem. B **109**, 4189 (2005).

[16] K. Schulten, Z. Schulten, and A. Szabo, J. Chem. Phys. **74**, 4426 (1981).

[17] W. Nadler and K. Schulten, J. Chem. Phys. **82**, 151 (1985); Z. Phys. B: Condens. Matter **59**, 53 (1985).

[18] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, Cambridge, 2001).

[19] Protein Data Bank: www.rcsb.org; H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig,I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).

[20] Q-H Dai, C. Thommos, E. J. Fuentes, M. R. A. Blomberg, P. L. Dutton, and A. J. Wand, J. Am. Chem. Soc. **124**, 10952 (2002).

[21] A. Liwo, M. Khalili, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **102**, 2362 (2005).

[22] Since no tunneling events occurred in the initial step, direct flow measurements according to Eq. (3) could not be performed at all in this case.

[23] Y. Zhang and J. Skolnick, Proteins: Struct., Funct., Bioinf. **57**, 702 (2004).