THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

ETNA: Equilibrium Transitions Network and Arrhenius Equation for Extracting Folding Kinetics from REMD Simulations

S. Muff, and A. Caflisch

J. Phys. Chem. B, **2009**, 113 (10), 3218-3226• DOI: 10.1021/jp807261h • Publication Date (Web): 20 February 2009

ETNA: Equilibrium Transitions Network and Arrhenius Equation for Extracting Folding Kinetics from REMD Simulations

S. Muff* and A. Caflisch*

Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Received: August 14, 2008; Revised Manuscript Received: December 1, 2008

It is difficult to investigate folding kinetics by conventional atomistic simulations of proteins. The replica exchange molecular dynamics (REMD) simulation technique enhances conformational sampling at the expenses of reduced kinetic information, which in REMD is directly available only for very short time scales. Here, we propose a procedure for obtaining kinetic data from REMD by making use of the equilibrium transitions network (ETN) sampled at the temperature of interest. This information is supplemented by mean folding times extracted from ETNs at higher REMD temperatures and scaled according to the Arrhenius equation. The procedure is applied to a three-stranded antiparallel β-sheet peptide which has a very heterogeneous denatured state with a broad entropic basin and several enthalpic traps. Despite the complexity of the system and the REMD exchange time of only 0.1 ns, the procedure is able to estimate folding times (ranging from about 0.1 μs at the melting temperature of 330 K to about 8 μs at 286 K) as well as transition times from individual non-native basins to the native state.

I. Introduction

Molecular dynamics (MD) and Metropolis Monte Carlo are simulation techniques widely used for Boltzmann-weighted (i.e., equilibrium) sampling. In principle, the main advantage of MD simulations is the correct description of the dynamics because the time-behavior of the system is not characterized in detail by Monte Carlo sampling.[1] In practice, because of the many degrees of freedom in the (poly)peptide chain and the related complexity of the free-energy landscape it is very challenging to sample the conformational space of peptides and proteins by standard MD techniques, which have an inherently "slow" time step of about 1−2 fs. At low temperatures, MD simulations can get trapped and sample mainly the starting basin. On the other hand, at high temperatures the accessible phase space increases dramatically and not all possible conformations are visited. A number of simulation techniques have been introduced to enhance the sampling of the conformational space.[2−4] At the same time, the availability of hundreds to thousands of processors has been exploited by intrinsically parallel jobs like distributed computing[5,6] and loosely coupled MD simulations.[7] Because of the significant time-scale gap between the actual folding process (microseconds to seconds) and simulation length (nanoseconds), it is not possible to extract folding kinetics directly from distributed computing simulations.[6,8] In this context, Markov chain models have been applied to determine transition probabilities between a small number (usually less than 100) of coarse-grained states from multiple short MD runs[9−11] but the development of an automatic procedure to cluster the MD snapshots into kinetically distinct states is a major obstacle and an active area of research.[12−15]

One simulation technique widely used to enhance sampling is replica exchange MD (REMD). In REMD, several noninteracting copies of the system are evolved in parallel over a range of temperatures.[16] The values of temperature are exchanged periodically using a Metropolis-like criterion that ensures sampling of the canonical ensemble at each of these values. REMD is more efficient than constant temperature MD (CTMD) for equilibrium sampling in particular at low temperatures as shown for peptide folding[17] and aggregation.[18] However, the REMD sampling consists of many discontinuous segments of trajectory, which cannot be used straightforwardly to analyze the kinetics on relevant time scales.

Four approaches to the extraction of kinetics from REMD have been published. Andrec et al. have proposed a network model in which links represent allowed conformational changes between states according to a geometrical similarity criterion, and each snapshot is a node of the network. Sampling at different temperatures is combined according to the kinetic energy of states.[19] Van der Spoel and Seibert assumed a two-state model and fitted the four parameters of a rate equation by employing the fraction of native folded species along a heterogeneous set of 16 REMD and 4 CTMD trajectories of a β-hairpin decapeptide.[20] Yang et. al approximate the folding process by Langevin dynamics along a one-dimensional reaction coordinate R with effective random forces and diffusion coefficient (as a function of R) extracted from REMD.[21] Their approach requires the a priori definition of a one-dimensional reaction coordinate for folding which almost always masks the hidden complexity of the folding process.[22−24] This complexity is also masked in the two-state assumption of van der Spoel and Seibert. Recently, Buchete and Hummer have proposed a procedure to extract rates from the number of transitions on the time scale of replica exchanges by calculating the rate coefficients of a master equation using the maximum likelihood technique.[14,25] They applied this procedure to the blocked alanine pentapeptide in explicit water (which was coarse-grained into 32 states according to a 5-bit string of residue helicity) but concluded their letter by explicitly mentioning that the application to protein folding might "pose a major challenge" because of the large number of states.[14]

Here, we present a procedure for extracting kinetics from REMD which can be applied to systems more complex than those mentioned above, for example, peptides and proteins

* To whom correspondence should be addressed. E-mail: (A.C.) caflisch@bioc.uzh.ch; (S.M.) smuff@bioc.uzh.ch. Tel.: +41 44 635 55 21.

10.1021/jp807261h CCC: $40.75
Published on Web 02/20/2009

simulated at atomistic resolution. First, the equilibrium transitions network (ETN) is constructed for each value of the temperatures used in REMD. More precisely, the ETN is the capacitated graph whose nodes and links represent coarse-grained microstates and transitions, respectively, sampled in the short segments at constant temperature. The ETN often consists of several disconnected components because of the short trajectory segments between replica exchanges and due to free-energy barriers separating states. Within each component an equilibrium phase-space distribution at the respective temperature is sampled because of the canonical-ensemble sampling within the REMD segments. An important aspect of the procedure for extracting kinetics from REMD is that the ETN can be treated as a Markov state model, implying that Monte Carlo simulations on the network reproduce the correct dynamics within each component. To estimate folding rates at each temperature, mean folding times (mft's) are computed as in refs 26 and 27 on the ETN component that is connected to the native state. Finally, the Arrhenius equation and the sampling at high temperatures are used to extract kinetics for the low temperature nodes that are disconnected from the NC of the ETN (Figure 1). The procedure is termed ETNA because of the combination of the ETN and the Arrhenius equation. Moreover, thanks to the thermodynamically correct sampling from the REMD trajectories and the integration of short REMD segments into ETN components, it is possible to extract correct populations of enthalpic free-energy basins from the analysis with cut-based free-energy profiles (cFEPs), which is a method for grouping conformations according to (local) transitions at equilibrium.[28]

ETNA is applied to the miniprotein called Beta3s[29] whose native structure corresponds to a three-stranded antiparallel β-sheet consisting of two β-hairpins.[30] Beta3s has been shown to fold to the native structure determined by NMR[30] in molecular dynamics simulations with the CHARMM polar hydrogen molecular mechanics potential energy function supplemented by a simple implicit solvent model.[29] Since folding simulations of Beta3s are very fast close to its melting temperature of 330 K (folding time of about 0.1 μs, which requires roughly 18 h on a single core of a XEON 2.33 GHz), many studies have been made to elucidate its folding mechanism.[15,17,23,27,29,31,32] ETNA is able to extract from REMD overall folding times of Beta3s, as well as folding times from individual basins in the unfolded state, that are in good agreement with the corresponding values obtained by multiple CTMD folding runs started from the denatured state ensemble at 286 K. Therefore, kinetics on time scales 5 orders of magnitude longer than the REMD segments are accessible, as the REMD exchange time was only 0.1 ns and the folding time of Beta3s is about 8 μs at 286 K.

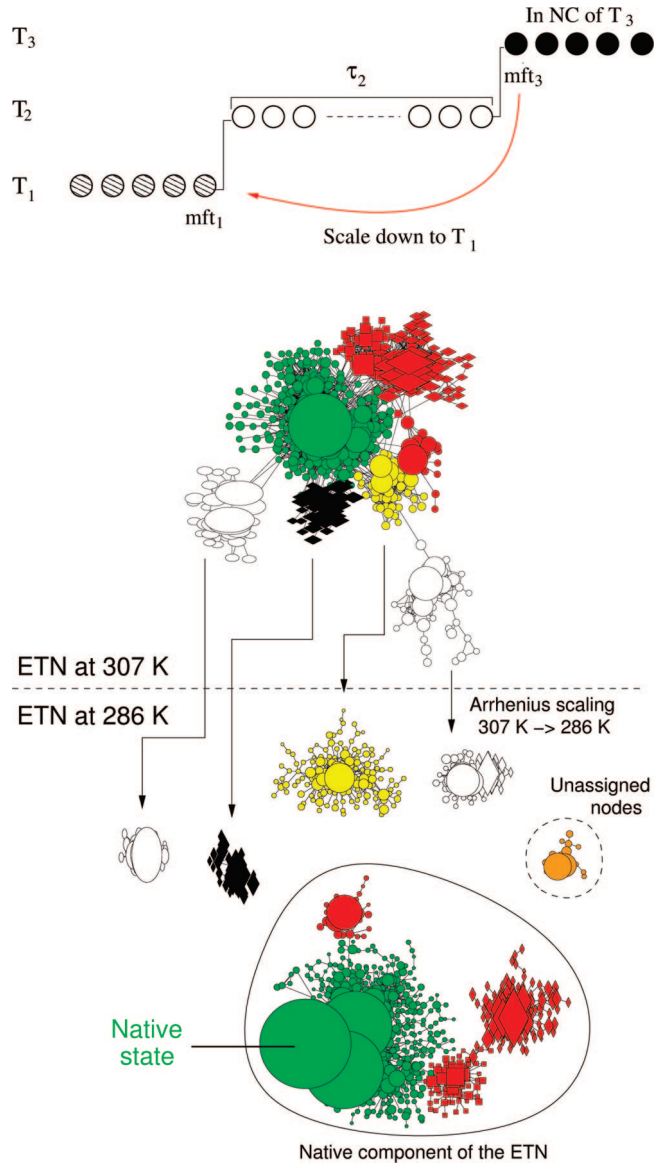

Figure 1. Illustration of the ETNA procedure for folding time evaluation of snapshots outside of the native component (NC) at the temperature of interest. (Top) Whenever a disconnected REMD segment at temperature T_1 is visited (dashed nodes), the next segment in the time series of that replica is considered (white nodes). In the example shown, only the second temperature increase to T_3 is successful in visiting snapshots belonging to nodes of the NC (black nodes). The details of the procedure are given in the text. (Bottom) Schematic view of the main idea behind the Arrhenius scaling approach. Folding time information extracted at high temperature is used to estimate folding kinetics at low temperature. Each color and symbol shape represents a free-energy basin.

II. Theory

A. Equilibrium Transition Network (ETN) from REMD Segments at Constant Temperature. The trajectory segments collected in REMD simulations at a given temperature are much shorter (picoseconds) than the time scales of large conformational transitions or folding (microseconds to seconds). The length of the segments depends on the frequency of the swapping attempts and their acceptance ratio, which is usually 25−30%. These segments of trajectory are between 3 and 6 orders of magnitude shorter, depending on the temperature, than the folding time of a structured peptide or a small protein. The essential idea of ETN is to extract kinetics from the integration of all REMD segments at the same temperature. For complex systems, the ETN at each temperature consists of several disconnected parts, one of which contains the native state and is termed native component (NC) hereafter. The NC is usually the largest component, but its size can be reduced at very low temperature due to large free-energy barriers between states, as well as at high temperature (above 307 K for Beta3s; see section IIIF) because sampling is not sufficient to fully connect the large accessible space, especially in the presence of entropy-dominated regions.

There are two important conditions on the ETNs. First, the individual ETN components must fulfill the property of Markov state models. Markovianity depends on the way how snapshots are grouped into nodes and on the lagtime of the transitions. The second condition is that all components represent locally the correct connectivity and population of states, that is, that the ETN assembled from REMD sampling is indistinguish-

able from the corresponding portion of the ETN from converged CTMD simulations. This requires that the REMD exchange time is long enough for establishing local connectivity.

Note that for an ETN generated by the combination of thousands of short trajectories, like in REMD, it is important to symmetrize the transition matrix (i.e., impose detailed balance) by replacing the absolute number of transitions n_{ji} from node i to node j by $c_{ji} = (n_{ji} + n_{ij})/2$. Such an enforced detailed balance is allowed only if the REMD simulations are long enough to reach equilibrium at all temperatures. While this step is helpful (but not essential) for long equilibrium trajectories,[28] it is necessary for ETNs extracted from REMD to avoid dead-ends. A dead-end may arise when the trajectory is interrupted because of a temperature swap, leaving the last visited node without a next neighbor. Such nodes are problematic if transition times are calculated by solving the respective master equation on the ETN (see next subsection).

B. Mean Folding Time Calculation on the NC at Constant Temperature. The mean folding time (mft) is the mean first passage time to the native node. Given the transition probability p_{ij} between nodes j and i (with $p_{ij} = c_{ij}/\sum_k c_{kj}$), the mft for a node i in the NC at a given temperature is the solution of the equation system $\text{mft}_i = \Delta t + \sum p_{ij} \cdot \text{mft}_j$, which can be determined by iterative multiplication.[26,27] Δt is the lagtime time of the Markov state model. Solving the equation system allows the calculation of the mft from nodes in the NC that actually never fold within any of the short REMD segments. On the other hand, it is not possible to calculate the mft of nodes not belonging to the NC. For snapshots in non-NC nodes, the Arrhenius-scaling approach (ETNA) is introduced as follows in the next subsection.

C. Scaling Folding Times Using the Arrhenius Equation. An essential aspect of the ETNA procedure is the use of the Arrhenius equation and high-temperature sampling to extract kinetics at low temperature for microstates that do not belong to the NC. Assuming both the pre-exponential factor A and the activation energy to exit from a minimum of interest E_a to be temperature independent, the ratio of folding rates from the respective basin k_i at different temperatures T_1 and T_2 is

$$\frac{k_2}{k_1} = \frac{Ae^{-E_a/RT_2}}{Ae^{-E_a/RT_1}} = e^{(E_a/R)(1/T_1 - 1/T_2)}$$
$$\Rightarrow \quad \tau_1 = \tau_2 \cdot e^{(E_a/R)(1/T_1 - 1/T_2)} \tag{1}$$

In a first approximation, E_a/R can be taken as a universal constant of the system and extracted by a linear fit of the $1/T$ versus $\ln(k)$ plot of unfolding rates at several temperatures. Note that this assumption of universality for E_a/R is invalid if folding barriers from different non-native regions are very heterogeneous or very different from the unfolding barrier, but a more general theory with multiple scaling factors can be derived in a straightforward way.

The Arrhenius equation is an approximation that ignores entropic contributions. Using the Eyring equation from transition state theory, the ratio of reaction rates can be written as

$$\frac{k_2}{k_1} = \frac{T_2}{T_1} \cdot e^{[(-\Delta G/RT_2) - (-\Delta G/RT_1)]}$$
$$= \frac{T_2}{T_1} \cdot e^{[(-\Delta H/R)((1/T_2) - (1/T_1))]}$$

Thus, under the simplifying assumption that ΔH and ΔS are temperature independent, the entropic contribution $T\Delta S$ cancels in the ratio of rates even when the Eyring approach is used. Moreover, the "pre-factor" T_2/T_1 is close to 1.0 for similar temperatures, as they are usually employed in REMD simulations. Therefore, the use of the (simpler) Arrhenius equation is justified.

As mentioned above, the scaling of folding times according to the Arrhenius equation comes into play because at low temperatures the ETN from REMD is usually split into disconnected pieces due to high free-energy barriers that separate basins (Figure 1). Therefore, folding times from outside the NC

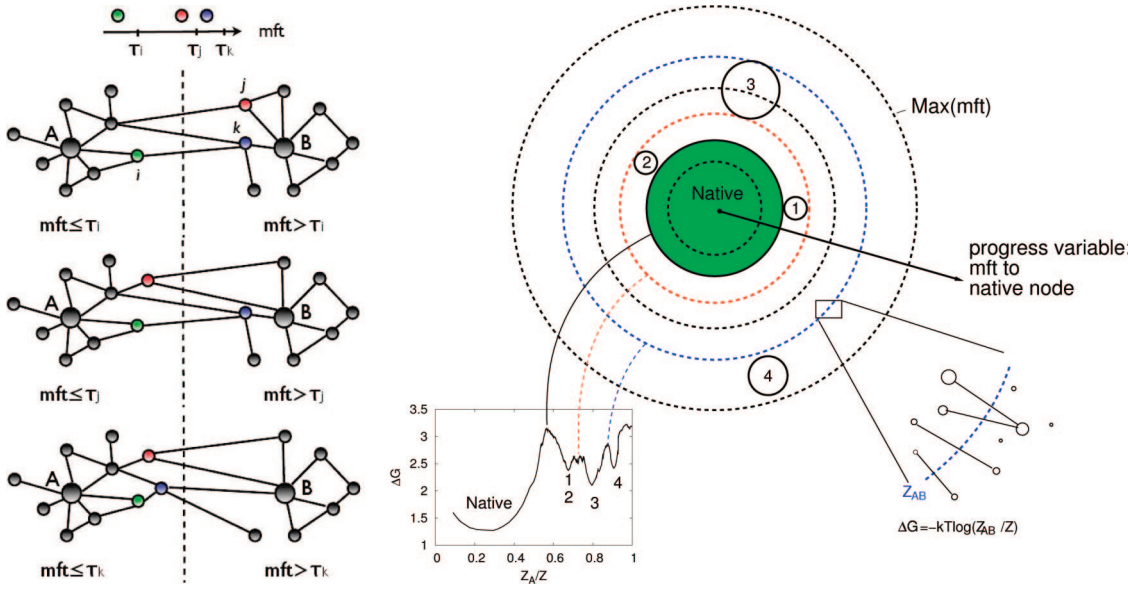

Figure 2. Schematic illustration of the cFEP procedure.[27,28] (Left) Nodes of the ETN are first sorted according to increasing mft. For each value mft_c between 0 (node A) and Max(mft), a value of the cut Z_{AB} between nodes A and B is calculated. The set of nodes on the left of the cut contains node A and all nodes with $\text{mft} \le \text{mft}_c$, where Z_A/Z is its relative partition function. The green, red, and blue nodes have consecutive values of mft in this simplified illustration of the ETN. (Right) Relation between free-energy basins and the cFEP. Each solid circle borders a basin, while concentric dashed circles represent values of mft. To illustrate the cFEP, $\Delta G = -kT \log(Z_{AB}/Z)$ is plotted as a function of Z_A/Z. Basins 1 and 2 overlap because they have the same mft distance from the native state and are therefore not separated in the unfolded part of the profile.

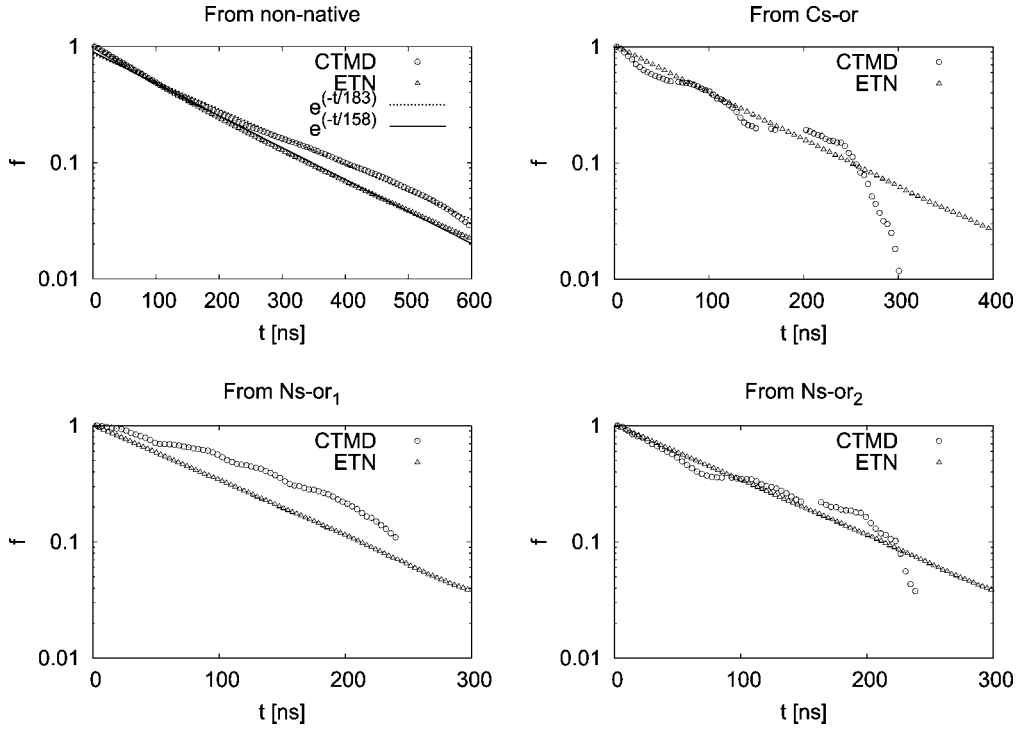

Figure 3. Cumulative folding time distribution as extracted directly from the 330 K CTMD simulation (circles) and from the corresponding ETN, which is treated as a Markov state model (triangles). The folding dynamics from the non-native ensemble (top left) and from specific metastable states (top right and bottom) can be reproduced by the model, which is a very strong indication that the Markov assumption is justified for the lagtime of 20 ps used here.

at a low temperature of interest (T_1) cannot be calculated directly on the ETN. When the temperature of a replica is swapped to the next higher temperature (T_2) in the REMD simulation, the trajectory moves to the ETN at T_2. If nodes of the ETN at T_2 are visited, the mft of the closest (in time) node is scaled to T_1 according to the Arrhenius eq 1 and the snapshots in the previous T_1 segment are assigned an mft. If the procedure is not successful for T_2, the next temperature T_3 is considered and so on. If between two T_1 segments the system does not visit the NC at any other temperature, it is not possible to assign mft's to the previous T_1 segment. Those snapshots remain unassigned and are therefore ignored. Note, however, that the scaling of folding kinetics with ETNA is valid only in temperature ranges where folding times follow the Arrhenius law. If a temperature T_A is known, where the system starts to show an anti-Arrhenius behavior, the data with $T > T_A$ must be discarded from the analysis. Hence, the snapshots in a T_1 segment are ignored if no NC at a temperature between T_1 and T_A is visited before the system continues to $T > T_A$.

Figure 1 top illustrates the ETNA algorithm for the case where nodes at T_3 are used to scale a fragment of the trajectory at T_1. The mft of the first of these T_3 nodes is used to calculate the theoretical mft of the last T_1 snapshot (mft_1), taking into account also the effective time τ_2 spent in the segment at T_2, $\mathrm{mft}_1 = \mathrm{mft}_3 \cdot e^{(E_a/R)(1/T_1-1/T_3)} + \tau_2 \cdot e^{(E_a/R)(1/T_1-1/T_2)}$, where mft_3 was previously calculated by solving the system of equations at T_3. Since a snapshot cannot have a mean folding time, but only one value originating from one folding event along the trajectory, the folding time τ assigned to the last T_1 snapshot is chosen randomly according to the exponential distribution around mft_1 as $P(\tau) = ke^{-k\tau}$ with $k = 1/\mathrm{mft}_1$. This last step is essential in order to obtain, in addition to the average value, a cumulative folding time distribution, which is used later for analysis. All remaining T_1 snapshots in the considered segment are assigned a folding time exponentially distributed around $\mathrm{mft}_1 + i\Delta t$, with i being the number of timesteps backward from the last snapshot in the segment, and Δt the lagtime of the model. Note that with this procedure the mft scaling from higher temperature to the reference temperature is done separately for every snapshot in nodes not connected to the NC of T_1. Therefore the ETNA approach is different from pure Arrhenius-based methods,[20] because each snapshot is assigned an individual folding time value, which depends on the route the system takes for folding.

D. Cut-Based Free-Energy Profiles (cFEPs). The cFEP approach was first introduced in ref 28 and further developed in ref 27. For a node i in the ETN the partition function is $Z_i = \sum_j c_{ij}$ where, as mentioned above, c_{ij} is the symmetrized number of transitions between nodes j and i. If the nodes of the network are partitioned into two groups A and B, then $Z_A = \sum_{i\in A} Z_i$, $Z_B = \sum_{i\in B} Z_i$, $Z_{AB} = \sum_{i\in A,\, j\in B} c_{ij}$ and the free energy of the barrier between the two groups is $-kT \log(Z_{AB}/Z)$ with Z being the partition function of the full network (Figure 2). The cFEP has the advantage with respect to the projections onto geometric coordinates that barriers are preserved.[28] In particular, the relative partition function Z_A/Z includes all pathways to and from the state of interest (e.g., the folded state). The cFEP method groups conformations according to equilibrium kinetics. Their application to components of the ETN is possible, because transitions from constant temperature segments establish locally the correct connectivity. Therefore, the profiles of ETN components are expected to be identical to those that would be extracted from equilibrium sampling. The cFEP analysis was performed with the program WORDOM,[33] which is particularly efficient in handling large sets of trajectories. Here, only cFEPs with mft as progress variable are used. Values of mft for individual nodes are calculated as explained above.

E. Isolation of Free-Energy Basins. Since the ETN constructed from REMD segments at constant temperature yields multiple disconnected components, it is not possible to obtain the complete cFEPs, that is, the profile up to $Z_A/Z = 1$. However,

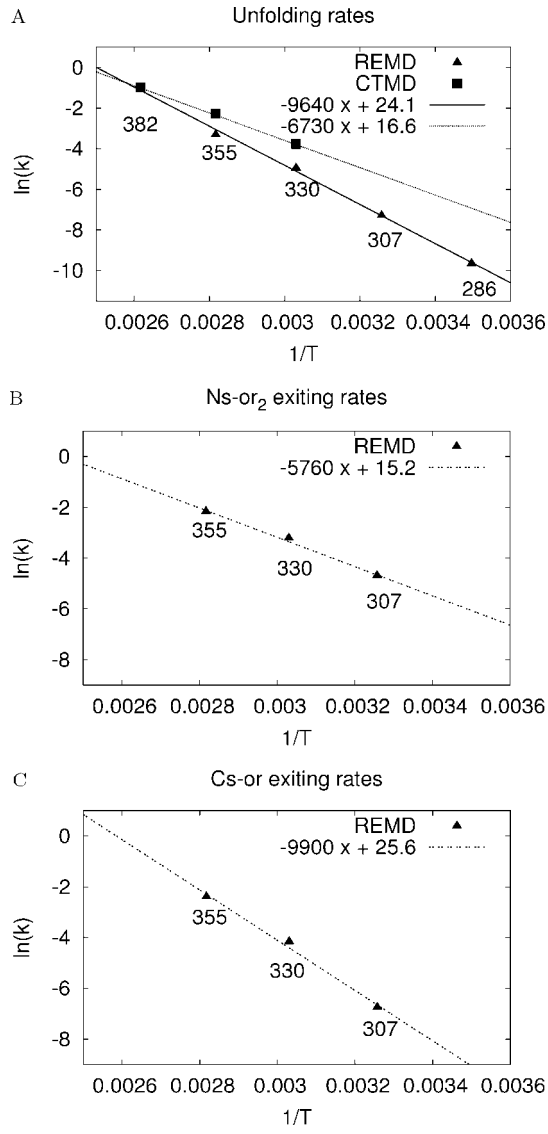

Figure 4. Temperature dependence of rates to exit enthalpic basins. (A) Unfolding rates estimated by Monte Carlo runs on the ETN of the CTMD simulations (squares) and the ETN of the NC from individual REMD temperatures (triangles). The estimates for the Arrhenius constant E_a/R (eq 1), which is used to scale the kinetics at different temperatures, can be extracted from the linear fit of unfolding rates. (B) Exiting rates from the Ns-or$_2$ basin. (C) Exiting rates from the Cs-or basin. Activation enthalpy values to exit individual basins are similar, justifying the use of only one Arrhenius constant in the ETNA approach.

the majority of nodes within a given free-energy basin belong to the same component of the ETN, at least if relaxation in the basin is as fast as the minimal length of the segments, which ensures that different REMD segments are connected through their visits to some of the highly populated nodes. Therefore, the procedure to extract basins from the cFEP remains the same as for the NC, where unfolding cFEPs from a node in the basin of interest (usually its most visited node) are plotted. The nodes lying on the left of the cut at the first barrier make up the basin.

III. Application of ETNA to Beta3s

A. Molecular Dynamics Simulations. All simulations and part of the analysis of the trajectories were performed with the program CHARMM.[34] The designed 20-residue peptide Beta3s[30] was modeled by explicitly considering all heavy atoms, and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[35] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent on the solute.[36] It was shown previously using exactly the same SAS-based implicit solvent model that at 330 K Beta3s folds reversibly to its NMR conformation irrespective of the starting structure, and importantly, 23 of the 26 NOE restraints are satisfied.[29] Despite the absence of collisions with water molecules, in the simulations with implicit solvent relative rates of folding of structured peptides are comparable with the values observed experimentally.[37−39] Importantly, the small variations in total SAS and radius of gyration during folding of Beta3s at 330 K[31] suggest that the lack of solute/solvent friction does not have a significant effect on pathways and kinetics.

B. REMD Setup. In the present simulations, eight replicas were run with temperatures (in K) of 286, 307, 330, 355, 382, 411, 442, and 476 for a simulation time of 11 μs each. Swapping attempts between replicas were performed every $\tau_{swap} = 0.1$ ns with an acceptance ration of about 25% and thus most REMD segments are 0.1−1 ns long. The Berendsen thermostat was used with a much shorter coupling constant of 5 ps to allow the temperature of the system to relax between two swapping attempts. Frames were saved with a frequency of 20 ps and therefore a REMD segment contains at least five consecutive snapshots before a new temperature is accepted. The low swapping versus saving frequency was chosen in order to let the system sample local transitions, which are the essential ingredient in the method presented here.

C. CTMD Folding Runs at 286 and 307 K. It is computationally prohibitive to obtain reversible folding−unfolding of Beta3s by CTMD at low temperature. Therefore, 750 CTMD folding runs at 286 K and 250 at 307 K were performed for comparison with REMD. Starting conformations were chosen uniformly distributed over the denatured state ensemble in the REMD segments at 286 and 307 K (see section IIE for definition of the native basin). Folding is defined by all-atom rmsd ≤2.5 Å from the snapshot in the center of the folded node in the REMD sampling at the respective temperature, as identified by the leader algorithm.[40] Therefore, a folding event is defined through the same structural constraints in both the CTMD folding runs and REMD. The CTMD simulations were stopped upon folding or after 10 μs, even if the folded state was not reached because of the large computational cost (about 300 days on a 200-CPU cluster). Note that 164 out of the 750 and 28 out of the 250 CTMD runs at 286 and 307 K, respectively, did not fold within 10 μs. Nevertheless, the 10 μs could be included in the cumulative folding time distribution $f(t) = \int_t^\infty p(\tau)d\tau$, because $f(t)$ is the probability that a folding event requires at least time t.

D. The Markov State Model of Beta3s. It is necessary to coarse-grain the snapshots because each conformation is visited only once; in other words, any trajectory per se is nothing but a long string of coordinate sets. There are several meaningful ways for clustering individual coordinate sets in the trajectory to obtain coarse-grained microstates (nodes is used synonymously in this paper), and different ones are likely to be most useful for different types of analysis. For a structured peptide like Beta3s or a β-hairpin, rmsd and secondary structural coarse-graining are obvious possibilities.[22,23,41] The coarse-graining used in this work is the leader algorithm based on the all-atom rmsd[40] with a cutoff of 2.5 Å. Note that nodes in the ETN with only one or two neighbors (i.e., one incoming and/or one outgoing neighbor) were grouped to their outgoing neighbor. This

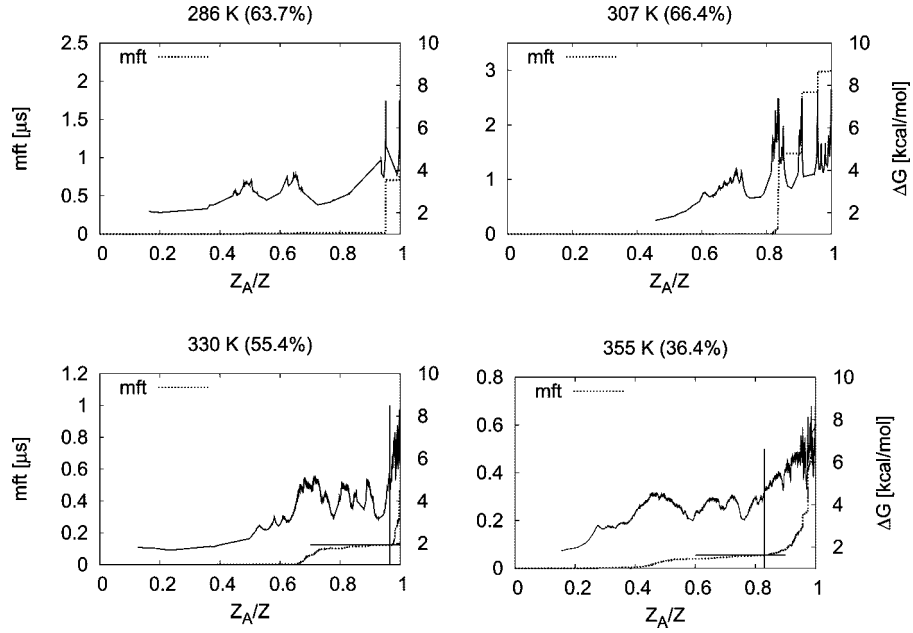

Figure 5. Identification of enthalpic basins for Arrhenius scaling. At each value of the temperature, the cFEP of the NC is shown by solid lines with ΔG values on the right y-axis. The percentage values in parentheses represent the statistical weight of the respective NC. The plot of mft as a function of the relative partition function Z_A/Z is shown with dotted lines, and the selection of the enthalpically stabilized part of the ETN is indicated by perpendicular lines. The criterion was to include as many enthalpic minima as possible by cutting at the point (crossing of perpendicular lines) where the roughness of the cFEP indicates insufficient sampling, which is often the case in entropically stabilized regions. At 286 and 307 K, all nodes of the NC were included, whereas only a subset was used at 330 and 355 K to remove entropic noise (see text for explanation). The mft cutoffs were chosen at 125 ns (330 K) and 55 ns (355 K). Interestingly, these cutoff values correspond roughly to the folding times observed in the CTMD simulations.

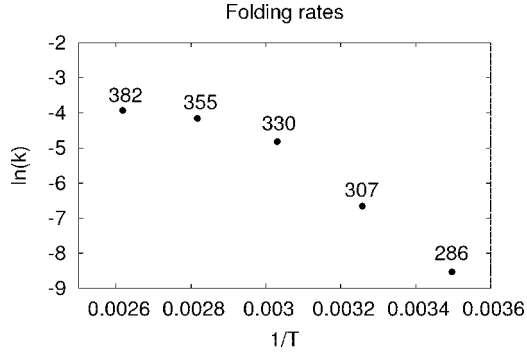

Figure 6. Folding rates from CTMD equilibrium simulations (330, 355, 382 K) and folding runs (286, 307 K) calculated by exponential fitting of the cumulative folding time distribution. There is a clear anti-Arrhenius transition above 355 K, and therefore sampling at higher temperatures was not included in the ETNA analysis.

regrouping is justified because the future of such nodes within a trajectory is completely determined, that is, no information is erased through their regrouping. Upon rmsd coarse-graining and regrouping, the following numbers of nodes were found: 4183 (at 286 K), 11 611 (307 K), 26 719 (330 K), and 38 445 (355 K). These nodes are the states of the Markov state model (i.e., the ETN), and the lagtime was set to $\Delta t = 20$ ps. Figure 3 contains a comparison of folding dynamics from the CTMD simulations and from the corresponding Markov state model at 330 K. There is a very good agreement for the overall dynamics, as well as for folding distributions from various metastable states, indicating that the Markov assumption holds.

E. The Arrhenius Fit. As explained in section IIC, the parameter E_a/R for the fit of the Arrhenius plot can be calculated from unfolding rates at different temperatures (Figure 4). The simplest way to obtain rates is by estimating them from the ETN of a CTMD simulation. Unfolding rates were extracted as the average time the system spends in the folded state before exiting from it, where the folded state is defined from the cFEP of the considered temperature by cutting at the first significant barrier in the profile (Figure 5).[27,28] Unfolding rates extracted with this procedure are shown as squares and fitted by the solid line in Figure 4a, from which the slope $E_a/R = 6730$ K was extracted.

It is also possible to approximate unfolding rates directly from the REMD data. This approach can be especially useful if equilibrium or unfolding simulations are too expensive, for example, for large systems or if the unfolding barrier is very high. The procedure to estimate the rates is the same as for CTMD with the only difference that the ETN is constructed from REMD (and not from CTMD) data and only the NC can be used. Figure 4a contains the rates estimated on the CTMD and REMD ETNs, where the slope $E_a/R = 9640$ K is obtained by fitting the latter. Figure S2 in Supporting Information shows that the main results obtained by ETNA are robust with respect to the type of simulations used to extract the value of E_a/R, implying that the REMD data is sufficient and no costly CTMD simulations to estimate unfolding rates are needed.

Interestingly, rates to exit other enthalpic basins can be fitted with similar E_a/R (Figure 4b,c), which means that the differences in activation energy to leave enthalpic basins of Beta3s are relatively small. Therefore, the approximation of using the activation energy for unfolding as a representative barrier to leave any enthalpic basin of the system is valid in the application of the ETNA procedure to Beta3s.

F. cFEPs from REMD Data at Individual Temperatures. The profiles from the ETN at each of the four lowest REMD temperatures are shown in Figure 5. These cFEPs represent only the NCs of each ETN, and include the indicated portions of the sampled conformational space. A comparison of the same profiles with the results from CTMD simulations at 330, 355, and 382 K shows a remarkable similarity up to the first

TABLE 1: Comparison of Populations and mft Values from Individual Basins Extracted from REMD Simulations by ETN(A) and the Corresponding Values Obtained by CTMD[a]

				weight (%)		mft (ns)	
	sec. str. string	name	NC[b]	REMD	CTMD	REMD	CTMD
330 K							
	-EEEESSEEEEEESSEEEE-	Native	yes	37.8	37.1		
	-EEEESTTEEEEESSEEEE-	Ns-or$_1$	yes	1.9	2.2	115	106
	--EEESSSEEEEESSEEEE-	Ns-or$_2$	yes	3.6	2.7	126	109
	--EESSSEEEEEESSEEEE-	Ns-or$_3$	yes	1.8	1.4	113	105
	-EEEESSEEEEESSSEEEE-	Cs-or	yes	5.3	5.3	101	109
	---SSGGG---EESSEETT-	Ch-curl$_1$	no	2.5	0.6	175 (4.2%)[c]	263
	---SSGGG-EESSTTTTEE-	Ch-curl$_2$	no	1.4	1.2	NA[d]	201
286 K							
	-EEEESSEEEEEESSEEEE-	Native	yes	60.6			
	-EEEESTTEEEEESSEEEE-	Ns-or$_1$	yes	3.1		705	2030
	--EEESSSEEEEESSEEEE-	Ns-or$_2$	no	2.8		6330 (98.0%)	3170
	--EESSSEEEEEESSEEEE-	Ns-or$_3$	no	0.5		6370 (100%)	6690
	-EEE-STTEEEESSSEEE--	Cs-or	no	0.7		13100 (96.4%)	970
	---SSGGG---EESSEETT-	Ch-curl$_1$	no	7.5		8820 (4.4%)	5260
	---SSGGG-EESSTTTTEE-	Ch-curl$_2$	no	4.0		NA[d]	7170

[a] The basins were identified with the cFEP approach and the DSSP secondary structure string[43] is the most frequent in the basin. [b] Several non-native basins at 330 K are in the native component (NC) of REMD, whereas the NC at 286 K consists of only the native basin and Ns-or$_1$. Note that the ETN or ETNA procedures were used for basins in the NC or outside of it, respectively. All folding times were extracted from the fit of the respective cumulative folding time distribution. [c] Values in parentheses are the fraction of snapshots to which a folding time could be assigned by ETNA. [d] The Ch-curl$_2$ basin was disconnected from the NC at all temperatures and therefore it is not possible to estimate its mft. Abbreviations: Ns-or, N-terminal strand out of register and folded C-terminal hairpin; Cs-or, C-terminal strand out of register and folded N-terminal hairpin; Nh-curl, curl-like conformation with folded N-terminal hairpin; Ch-curl, curl-like conformation with folded C-terminal hairpin.

significant barrier (Supporting Information, Figure S6), which indicates that the ETN from REMD sampling contains indeed the correct connectivity information.

At 286 and 307 K the main contribution originates from the native basin (with a weight of 60.6 and 54.3%, respectively). At higher temperatures the native state shrinks. Only 37.8, 16.8, and 1.3% remain native at 330, 355, and 382 K, respectively. Thus, even though the absolute size of the NC decreases for temperatures above 307 K, more non-native basins belong to the NC with increasing temperature.

G. Removal of Entropic Effects. Beta3s is known to spend about one-third of its time in an entropic region at 330 K, that is, in a non-native region with heterogeneous structures stabilized mainly by entropy.[27] Even in a 20 μs equilibrium simulation at 330 K, this entropic region suffers from incomplete sampling and the majority of the nodes is visited only once or a few times.[15,23] The entropic fraction increases dramatically at higher temperatures. The insufficient sampling of these regions introduces large errors to the ETN. Therefore, most parts of the entropic regions were ignored for calculations on the ETNs and only the well-sampled portion, corresponding to the enthalpic basins, was used to be consistent with the Arrhenius equation, which is valid for enthalpic barriers. Figure 5 shows how this selection was carried out for different temperatures with the help of the information from the cFEPs. At 286 and 307 K, no entropic contribution is present and all nodes were considered. At 330 and 355 K, the cFEPs (solid lines in Figure 5) show pronounced minima which represent the enthalpic basins. After the last enthalpic basin along the Z_A/Z coordinate, the cFEPs are clearly entropy-dominated, as can be seen from their rough shape which indicates insufficient sampling. All nodes above the threshold indicated in the profiles were discarded. The removal of these regions at high temperature does not bias the scaling of the kinetics by ETNA from high to low temperature, since a very small part of the free-energy surface is entropic at low temperature. In addition, temperatures higher than 355 K were not used in the application of the ETNA procedure to Beta3s, because there the folding rate shows a clear anti-Arrhenius transition according to Figure 6, that is, T_A (see section IIA) was chosen as 355 K.

H. Free-Energy Basins. All significantly populated free-energy basins with enthalpic stabilization could be determined from different ETN components. cFEPs from various basins that belong to different components at 286 and 330 K are given in Supporting Information, Figures S4 and S5, respectively. States at different temperature are considered to correspond to each other if the most populated DSSP secondary structure string[43] (first column in Table 1) is the same. This comparison ensures that the bottom of the corresponding basins contain similar conformations, but it clearly does not imply that the basins are completely identical and such an assumption is not used anywhere in this work. Populations extracted from the cFEPs are presented in Table 1. At 330 K the thermodynamics can be compared to those from a 20-μs equilibrium CTMD simulation. The results are in high agreement, except for the Ch-curl$_1$ enthalpic trap, which was visited only once in the CTMD trajectory and therefore has a large error. The high agreement between CTMD and REMD thermodynamics, both extracted by the cFEP approach, is not trivial because the cFEP method is based on the information of the *equilibrium* transitions between states, whereas REMD samples the correct ensemble of conformations, but only *local* transitions. Therefore, the use of cFEPs is only possible if transitions at constant temperature are sampled, as it is the case here because the REMD swapping frequency was chosen lower than the saving frequency of conformations.

I. Folding Time Estimates from REMD. The cumulative folding time distribution from nodes outside the native basin at 286 K is shown in the top left panel of Figure 7. The red control distribution from the 750 CTMD folding runs can be fitted between 1 and 10 μs with $e^{-t/7.76\mu s}$. Within the same interval, folding kinetics extracted from REMD with the ETNA procedure scale almost identically as $e^{-t/7.78\mu s}$. As a comparison, if only the non-native part of the NC is used, that is, if the Arrhenius-scaling is not applied, the ETN of REMD would suggest a folding time of roughly 0.7

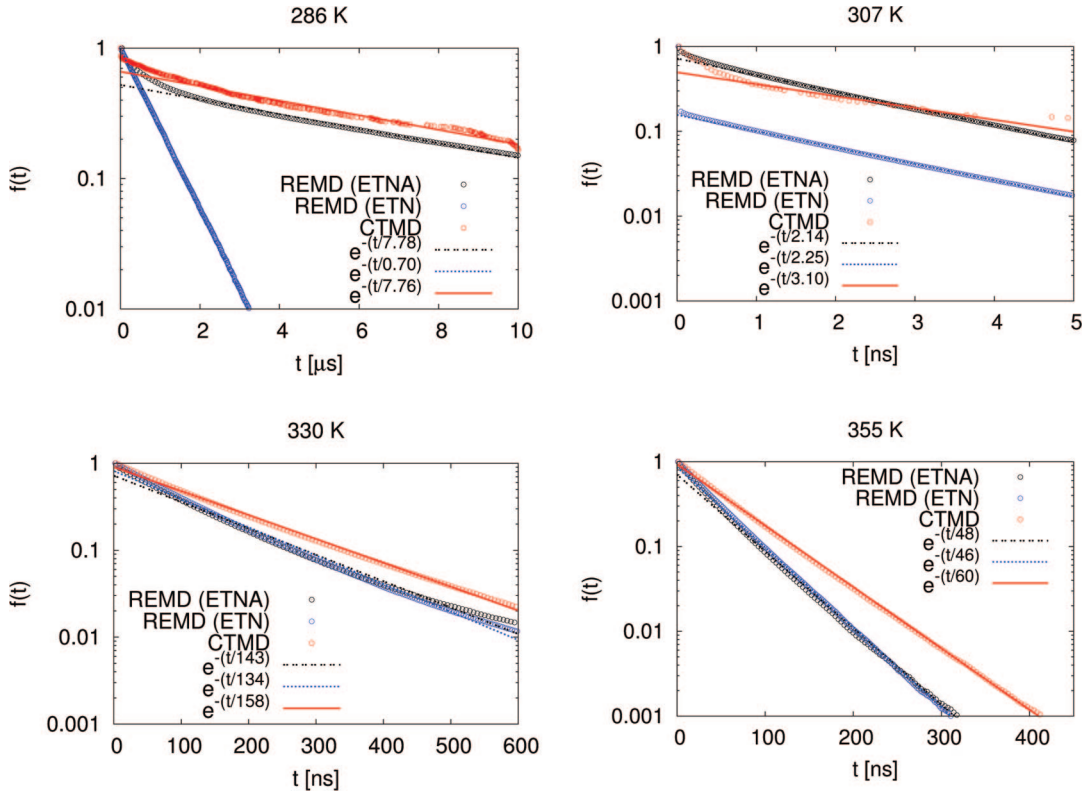

Figure 7. The cumulative folding time distributions $f(t) = \int_t^\infty p(\tau)\mathrm{d}\tau$ extracted from REMD using the ETNA approach (black) or only from the NC of the ETN (blue) are compared to the reference CTMD data (red). $p(\tau)$ is the probability density of the folding time distribution. The CTMD data at 286 and 307 K were extracted from 750 and 250 folding simulations, respectively, and started from the unfolded state ensemble. At 330 and 355 K, equilibrium CTMD simulations of 20 and 10 μs, respectively, were performed to compare the folding time distribution to the REMD approach. The CTMD and ETNA curves at all temperatures are in remarkable agreement. The use of only the NC of the ETN at 286 K yields a folding time that is faster by a factor of 10, whereas for temperatures of 307 K or higher the use of the Arrhenius scaling (ETNA) and only the NC (ETN) are almost identical to the CTMD results.

μs and therefore underestimates the real folding kinetics by 1 order of magnitude. Note that the ETNA procedure is able to scale only folding times from a fraction of all snapshots outside of nodes from the NC, because if no NC-node from the network of a temperature between T_I and T_A is visited before the replica continues to $T > T_\mathrm{A}$ (see section IIE), all snapshots of the previous T_I segment are ignored. At 286 K only 20.5% of the snapshots from nodes outside the NC could be assigned a folding time with ETNA. This result implies that the scaling of even a small fraction of folding kinetics is sufficient to yield correct overall rates.

At 330 K (Figure 7, bottom left), the CTMD kinetics were fitted for values up to 600 ns with $e^{-t/158\mathrm{ns}}$, while the ETNA-scaled times are only moderately faster ($e^{-t/143\mathrm{ns}}$). The folding times for trajectories starting from the ETN (i.e., only considering the NC) are distributed as $e^{-t/134\mathrm{ns}}$, thus unlike at 286 K, the application of the Arrhenius-based scaling of rates from different temperatures has almost no effect at 330 K. Similar cumulative folding time distributions are obtained by ETN and ETNA because at higher T the non-native regions of the NC are significantly populated, which reduces the effect of the Arrhenius scaling. According to Figure 7, the NC-ETN at 355 K and even the one at 307 K are sufficient to reveal approximately correct folding rates. Note that, since folding times from high temperatures are used to scale rates at low temperatures with ETNA, the availability of correct folding times from at least one higher temperature is necessary to obtain correct folding kinetics at low temperature.

In addition to overall folding time distributions, kinetics from individual basins were estimated at 286 and 330 K (Table 1). For basins belonging to the NC at the respective temperature, it is not necessary to use Arrhenius scaling to estimate folding times because values of mft can be calculated directly on the NC. In contrast, the mft of basins not belonging to the NC have to be evaluated with the Arrhenius approach. Because of the assignment of folding times to individual snapshots with ETNA, often even only a portion of all snapshots belonging to nodes of a basin can be scaled. This problem is severe in the case of Ch-curl$_1$, for which less than 5% could be assigned a folding value. In such a case, the folding time estimate is very inaccurate and it can be helpful to plot the cumulative folding time distribution. The latter does not contain all details of the folding kinetics, but is in return less sensitive to noise and statistical errors than plain distributions.[42] Therefore, the exponential fit to the former was used to estimate the folding kinetics from all basins (Supporting Information, Figure S5).

Similarly, the statistics harvested for individual basins from the 286 K CTMD folding runs are relatively low, because starting points of the 750 runs were distributed over all basins. Nevertheless, deviations between folding times from individual basins obtained with REMD or CTMD might originate from low statistics, yet the values lie within the same order of magnitude. The exception is Cs-or at 286 K, which exemplifies the main caveat of the ETNA approach. The Arrhenius equation approximates only the enthalpic contribution of barriers. Therefore, the folding time scaling of entropically stabilized regions of the free-energy landscape is not valid. Because of the considerable entropic stabilization of the Cs-or basin, which was reported earlier[27] and emerges also from a comparison of its statistical weight at 286 K (0.7%) and 330 K (5.3%), the scaling of the folding time at 286 K overestimates by 1 order of magnitude the mft from the CTMD simulations (Table 1).

IV. Conclusions

ETNA (equilibrium transitions network and Arrhenius scaling) is a procedure to extract kinetics from REMD simulations. At each of the REMD temperatures, the procedure makes use of the network whose nodes and links are the clustered snapshots and the transitions observed in the short REMD segments, respectively. These networks consist usually of a component that includes the native state and several disconnected components. An essential element of ETNA is the use of the Arrhenius equation for scaling mean folding times of nodes at temperature values higher than the temperature of interest. In this way, folding times at the latter temperature can be estimated for the nodes that are not connected to the native component. The use of the Arrhenius equation is the main difference between the ETNA procedure and a previously published approach based on the distribution of the kinetic energy.[19]

There are three conditions to apply the ETNA procedure. First, each component must fulfill the properties of a Markov state model. Second, the REMD segments should be long enough (i.e., the temperature-swapping frequency low enough) to allow for local transitions to take place at constant temperature in REMD, so that the ETN components at each REMD temperature are locally indistinguishable from the ETN obtained by a long CTMD simulation. Third, it is assumed that the scaling in terms of the Arrhenius equation is appropriate, that is, the free-energy basins are mainly enthalpic, so that the mean folding rate of a node is essentially identical to the corresponding rate constant for the entire basin. However, folding rates from different basins do not necessarily have to be identical and an adaptive scaling approach might be derived in the future.

ETNA was applied to extract folding kinetics at low temperature from a REMD simulation of Beta3s, a three-stranded antiparallel β-sheet peptide of 20 residues. Beta3s is a challenging test system because of its complex denatured state, which consists of several enthalpic traps, a basin with fluctuating helical conformations, and a heterogeneous entropic region at temperature values close to the melting temperature. Notably, overall folding rates of Beta3s and folding times from mainly enthalpic non-native basins are estimated correctly by ETNA. Moreover, the folding time of about 8 μs at 286 K is in agreement with NMR data (4−14 μs at 283 K).[30]

We plan to apply ETNA to extract folding kinetics of small proteins simulated by REMD with an efficient and accurate implicit solvent model.[44] Moreover, ETNA can be employed to investigate the kinetics of other biologically relevant processes like large conformational transitions involved in enzyme or receptor functions.

Acknowledgment. We thank Sergei V. Krivov, Philipp Schütz, and François Marchand for useful comments to the manuscript. The simulations were performed on the Etna and Matterhorn clusters of the University of Zurich, and we thank Christian Bolliger for hardware support. This work was supported by a Swiss National Science Foundation grant to A.C. Procedures for calculating the cut-based FEPs are available in WORDOM http://www.biochem-caflisch.uzh.ch/wordom.

Supporting Information Available: Cumulative folding time distributions with $E_a/R = 9640$ K and additional cFEPs. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
(2) Berne, B. J.; Straub, J. E. *Curr. Opin. Struct. Biol.* **1997**, *7*, 181.
(3) E Hansmann, U. H.; Okamoto, Y. *Phys. Rev. E* **1997**, *56*, 2228–2233.
(4) Frenkel, D.; Smit, B. *Understanding Molecular Simulations*; Academic Press: San Diego, CA, 2002.
(5) Snow, D. C.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102–106.
(6) Paci, E.; Cavalli, A.; Vendruscolo, M.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8217–8222.
(7) Settanni, G.; Gsponer, J.; Caflisch, A. *Biophys. J.* **2004**, *86*, 1691–1701.
(8) Ferst, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14122–14125.
(9) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415–425.
(10) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581.
(11) Sriraman, S.; Kevrekidis, I. G.; Hummer, G. *J. Phys. Chem. B* **2005**, *109*, 6479–6484.
(12) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(13) Noé, F.; Horenko, I.; Schuette, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(14) Buchete, N.-V.; Hummer, G. *Phys. Rev. E* **2007**, *77*, 030902.
(15) Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1185–1195.
(16) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
(17) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035–4042.
(18) Cecchini, M.; Rao, F.; Seeber, M.; Caflisch, A. *J. Chem. Phys.* **2004**, *121*, 10748–10756.
(19) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6801–6806.
(20) van der Spoel, D.; Marvin Seibert, M. *Phys. Rev. Lett.* **2006**, *96*, 238102.
(21) Yang, S.; Onuchic, J. N.; Garcia, A. E.; Levine, H. *J. Mol. Biol.* **2007**, *372*, 756–763.
(22) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(23) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(24) Caflisch, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.
(25) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B*, in press.
(26) Apaydin, M.; Brutlag, D.; Guesttin, C.; Hsu, D.; Latombe, J. In *International Conference on Computational Molecular Biology (RECOMB)*; ACM: New York, 2002.
(27) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(28) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(29) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780–10785.
(30) De Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. A. *Protein Sci.* **1999**, *8*, 854–865.
(31) Cavalli, A.; Ferrara, P.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 305–314.
(32) Cavalli, A.; Haberthür, U.; Paci, E.; Caflisch, A. *Protein Sci.* **2003**, *12*, 1801–1803.
(33) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(34) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(35) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.
(36) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 24–33.
(37) Ferrara, P.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. B* **2000**, *104*, 5000–5010.
(38) Settanni, G.; Rao, F.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 628–633.
(39) Ihalainen, J. A.; Paoli, B.; Muff, S.; Backus, E.; Bredenbeck, J.; Woolley, G. A.; Caflisch, A.; Hamm, P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 9588–9593.
(40) Hartigan, J. A. *Clustering Algorithms*; Wiley: New York, 1975.
(41) Hubner, I. A.; Deeds, E. J.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17747–17752.
(42) Chekmarev, S. F.; Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2005**, *109*, 5312–5330.
(43) Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 174–184.
(44) Haberthür, U.; Caflisch, A. *J. Comput. Chem.* **2008**, *29*, 701–715.

JP807261H