Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations

Frank Noé[a,1], Christof Schütte[a], Eric Vanden-Eijnden[b], Lothar Reich[c], and Thomas R. Weikl[c]

[a]Deutsche Forschungsgemeinschaft Research Center Matheon, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany; [b]Courant Institute, New York University, 251 Mercer Street, New York, NY 10012; and [c]Max Planck Institute for Colloids and Interfaces, Science Park Golm, 14424 Potsdam, Germany

Edited by Kenneth A. Dill, University of California, San Francisco, CA, and approved September 16, 2009 (received for review May 20, 2009)

Characterizing the equilibrium ensemble of folding pathways, including their relative probability, is one of the major challenges in protein folding theory today. Although this information is in principle accessible via all-atom molecular dynamics simulations, it is difficult to compute in practice because protein folding is a rare event and the affordable simulation length is typically not sufficient to observe an appreciable number of folding events, unless very simplified protein models are used. Here we present an approach that allows for the reconstruction of the full ensemble of folding pathways from simulations that are much shorter than the folding time. This approach can be applied to all-atom protein simulations in explicit solvent. It does not use a predefined reaction coordinate but is based on partitioning the state space into small conformational states and constructing a Markov model between them. A theory is presented that allows for the extraction of the full ensemble of transition pathways from the unfolded to the folded configurations. The approach is applied to the folding of a PinWW domain in explicit solvent where the folding time is two orders of magnitude larger than the length of individual simulations. The results are in good agreement with kinetic experimental data and give detailed insights about the nature of the folding process which is shown to be surprisingly complex and parallel. The analysis reveals the existence of misfolded trap states outside the network of efficient folding intermediates that significantly reduce the folding speed.

Discovering the mechanism by which proteins fold into their native 3D structure remains an intriguing problem (1, 2). Essential questions are: How does an ensemble of denatured molecules find the same native structure, starting from different conformations? Are there particular sequences in which the structural elements of a protein are formed (3–5)? Are there multiple parallel routes by which protein structure formation can proceed (6, 7)?

Full answers to these questions require one to characterize the ensemble of folding pathways, including their relative probabilities. In principle, this detailed information is accessible via molecular dynamics (MD) simulations which, when used in concert with experimental evidence, are becoming an increasingly accepted tool to understanding structural details that are not easily accessible via the experimental observables (8). MD simulations with atomistic models of proteins have been used to study the dynamics of small proteins with folding times in the microsecond range (9–13). However, even though MD simulations make the full spatiotemporal detail accessible to observation, the characterization of the pathway ensemble is computationally difficult: A brute-force approach would start simulations from an equilibrium of unfolded structures, say A, and simulate until they relax into a set of folded state B. The analysis would then only be comprised of those trajectory segments that leave A and relax to B without returning to A (Fig. 1). Such a procedure is generally unpractical, since protein folding is a rare event and often the affordable simulation length is insufficient in order to observe an appreciable number of folding events, unless very simplified simulation models are used. On the other hand, sampling techniques that enhance the sampling of predefined reaction coordinates may work well for systems with simple transition pathways but are not likely to produce unbiased results in protein folding, which likely involves many statistically relevant but a priori unknown transition states.

It is thus desirable to develop a method that can reconstruct the equilibrium ensemble of folding pathways from simulations that are not driven and yet are much shorter than the folding time. Ideally, these simulations sample different parts of conformation space and thus provide information on different subsegments of the folding process. The reconstruction of folding pathways needs to be done in a statistically correct manner, so as to avoid a bias toward fast folding pathways (14). Master-equation or Markov models of the molecular kinetics are a natural approach toward this goal, as they decompose the macroscopic and possibly slow transition (folding) into a network of faster transitions between individual conformational states (15–17). It has been shown for small model systems that Markov models permit the combination of information from short simulations and extract the correct long-time kinetics (18, 19). Markov models of small proteins have also been reported to reproduce experimental folding time scales (20). In addition, Markov models have the advantage over many other analysis methods that they do not require the definition of a reaction coordinate that may provide a biased or oversimplified view on the kinetics (21).

The present paper proposes an approach for computing the equilibrium ensemble of folding pathways from short simulations started out of equilibrium based on (*i*) combining the information from the short trajectories into a joint Markov model, and (*ii*) a theory of folding pathways that is based on the mathematical framework of transition-path theory (22, 23) for computing the ensemble of folding pathways along with their relative probabilities.

In order to illustrate the approach for computing folding pathways, consider the model potential in Fig. 2*A*, which is a funnel-like energy landscape with the "unfolded" high-energy states $A1$, $A2$, $A3$, "intermediates" I1,I2, and the "native" state B. A discrete state space was defined by a 50×50 grid lattice and the dynamics are given by a Markov jump process between neighboring grid cells (see *SI Appendix* for details). The main ingredient for computing the transition pathways from A to B is the committor probability, also known as probability of folding, p_{fold} (see Fig. 2*B*) (22, 24, 25). For every state i, the committor gives the probability to fold from that state (toward B), instead of unfolding (toward A). Based on the Markov model and the committor, transition-path theory allows the "folding flux" to be computed, which represents the net flux of folding trajectories leaving the unfolded and entering the folded set (see Fig. 2*A*) and provides the probability of any

Author contributions: F.N., C.S., E.V.-E., and T.R.W. designed research; F.N. and L.R. performed research; F.N. and L.R. analyzed data; and F.N., C.S., E.V.-E., and T.R.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

See Commentary on page 18879.

[1]To whom correspondence should be addressed. E-mail: frank.noe@fu-berlin.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905466106/DCSupplemental.

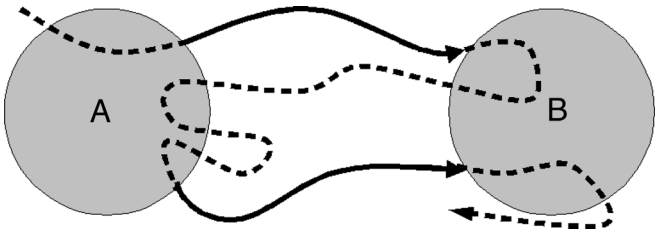

Fig. 1. Transition-path theory provides the probability distribution of productive *A* (unfolded) to *B* (folded) segments (solid lines) of a hypothetical, infinitely long trajectory.

realization of the folding pathway. This folding flux allows many quantitative properties of interest to be to computed, such as the probability that one structural element forms before the other.

Although the full information of all possible folding pathways is contained in the folding flux, this information is too detailed to provide a simple illustration of the folding mechanism. Therefore, a flux lumping procedure is derived, which allows one to coarse-grain the folding flux into a net flux between macroscopic conformations (Fig. 2*C*), showing the essential network of folding pathways. This network can be decomposed into individual pathways, if desired (Fig. 2*D*). The relative probability of each folding pathway is immediately given by the magnitude of the flux along it.

The PinWW domain is required for the regulation of many cellular processes and has been a model system for studying protein folding (26–28). PinWW is comprised of a twisted, triple-stranded, antiparallel β-sheet containing two Trp residues. Extensive thermodynamic and kinetic experiments of the PinWW–domain folding process have been conducted (26). These experiments show that a single time scale on the 10-μs order dominates the slow-relaxation kinetics, thus PinWW is an apparent two-state folder. Recent measurements, however, find that some mutants have more than one time scale, indicating that additional relaxation processes between local energy minima appear when examining the kinetics in detail (29).

The approach to computing folding pathways is applied to a set of 180 explicit solvent simulations of PinWW, whose individual length is two orders of magnitude shorter than the slowest time scales in the system. The Markov model allows the probability of the folded state and the kinetic relaxation to be calculated, and these results are in good agreement with experiment. It is demonstrated that the Markov model is an approximately unbiased model for the dynamics on long time scales. Remarkably, the method allows the folding pathways to be reconstructed despite the fact that no contiguous pathway from the unfolded to the folded state is observed in a single simulation. The results provide detailed insights into the surprisingly complex and parallel nature of the folding process and reveal the existence of misregistered trap states that slow down the folding.

Theory

Markovian Dynamics and Transition Probability Matrix. In order to combine the information contained in many short MD trajectories, a model for the dynamics between the various molecular conformational states is needed. Assume that the state space of the molecule is discretized into a set of $S = \{1, \ldots, m\}$ conformational states (typically a few thousand) and that then a $m \times m$ transition probability matrix $\mathbf{T}(\tau)$ is computed, where each element T_{ij} measures the probability of going from state i to state j within time τ, by $T_{ij} = c_{ij} / \sum_k c_{ik}$. Here, c_{ij} counts the number of times the trajectory was in i at time t and in j at time τ later. Although this expression provides the most likely transition matrix, the full probability distribution of $\mathbf{T}(\tau)$, given c_{ij}, must be considered when statistical uncertainties of $\mathbf{T}(\tau)$ and properties computed from it are desired. Here, 80% confidence intervals

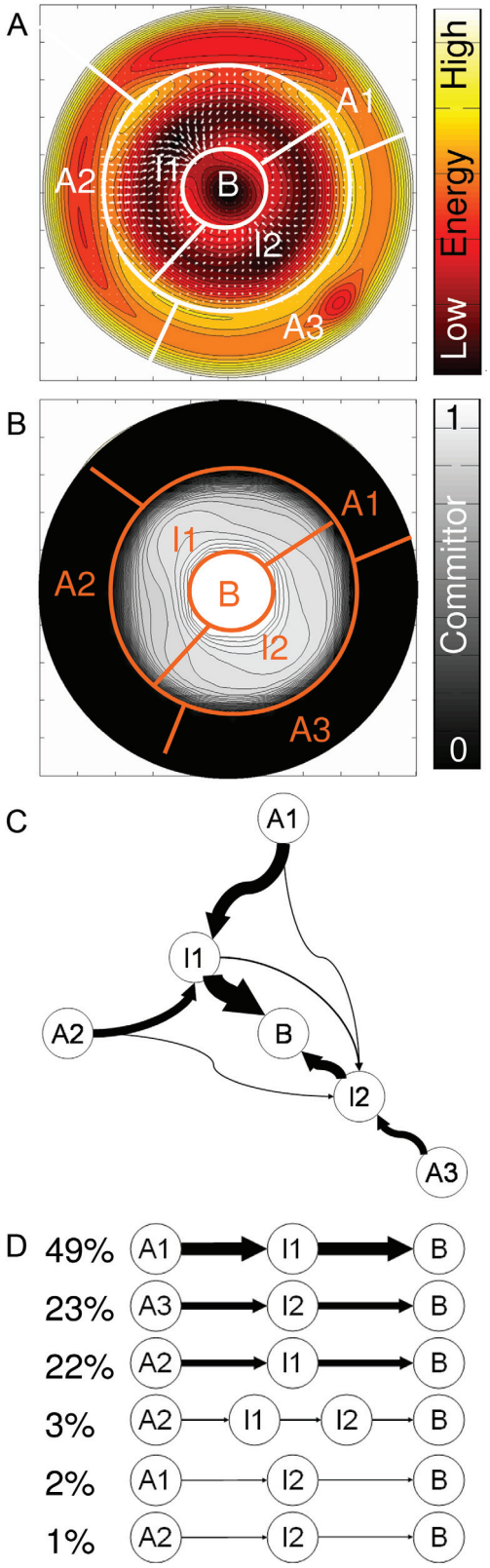

Fig. 2. Illustration of transition-path theory on a model potential. (*A*) The model potential. Three "denatured" source sets (*A*1, *A*2, *A*3), two intermediates (I1,I2), and one "native" target set (*B*) are defined. The folding flux, i.e., the net flux of the $A \rightarrow B$ transition among conformational states, is shown with white arrows. (*B*) The forward committor (probability of folding). (*C*) The coarse-grained flux of the $A \rightarrow B$ transition among macrostates. (*D*) The hierarchy of the $A \rightarrow B$ transition pathways with their contributions to the total flux.

are used to estimate the uncertainty of computed properties (see *SI Appendix* for details).

We require that $\mathbf{T}(\tau)$ is ergodic, i.e., any state can be reached from any other state within a finite time. Then, $\mathbf{T}(\tau)$ has a single eigenvector with eigenvalue 1. When normalizing this eigenvector, it gives the stationary probability π. For equilibrium MD, π is the Boltzmann distribution, and the detailed balance condition holds: $\pi_i T_{ij} = \pi_j T_{ji}$.

When the vector $\mathbf{p}(t)$ denotes the probability of the system to be in any of its m states at time t, the probabilities at time $t+\tau$ are given by $\mathbf{p}(t+\tau) = \mathbf{p}(t)\mathbf{T}(\tau)$. In order for the Markov model to correctly represent the long-time MD, it must be tested whether $\mathbf{T}(n\tau) \approx \mathbf{T}^n(\tau)$ holds, thus allowing $\mathbf{p}(t+n\tau) = \mathbf{p}(t)\mathbf{T}^n(\tau)$ to be computed (see *SI Appendix* for details).

Transition-Path Theory. In order to compute transition pathways, two subsets of the state space, A and B, are defined to specify the transition process one wants to investigate. Here, A and B correspond to the nearly unstructured unfolded and the native set, respectively. All remaining states are unassigned intermediate states I. What is the probability distribution of the trajectories leaving A and continuing on to B? That is, what is the typical sequence of I states used along the transition pathways?

The essential ingredient required to compute the statistics of transition pathways is the committor probability, q_i^+, defined as the probability, when being at state i, that the system will reach the set B next rather than A (22, 24, 25). In the current context, it is the probability of folding, often denoted as p_{fold} (24). By definition, $q_i^+ = 0$ for all i in A and $q_i^+ = 1$ for all i in B. The committor probability for all intermediate states i can be computed by solving the following system of equations:

$$-q_i^+ + \sum_{k\in I} T_{ik} q_k^+ = -\sum_{k\in B} T_{ik}.$$

The committor gradually increases from state A to state B (see Fig. 2*B* for illustration).

Furthermore, the backward-committor probability q_i^- is the probability, when being at state i, that the system was in set A previously, rather than in B. For a molecule in equilibrium, this probability is simply $q^- = 1 - q^+$.

The transition probability T_{ij} contains contributions from all trajectories, including trajectories that leave A and return to A before hitting B, or $B \to A$ trajectories. In order to evaluate the statistics of $A \to B$ trajectories, only a fraction of the transitions which come from A and go on to B is relevant, i.e. $q_i^- T_{ij} q_j^+$. The effective flux f_{ij} is defined as the probability flux along edge i,j, contributing to the transition $A \to B$:

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+.$$

The effective flux still contains unnecessary detours, such as recrossings $A \to \ldots \to i \to j \to i \to j \to \ldots \to B$. Thus, for any pair i,j in the intermediate set of states, both f_{ij} and f_{ji} are positive. In order to only consider the net flux of $A \to B$ trajectories, f_{ij}^+, one computes

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}.$$

f_{ij}^+ defines the folding flux and is a network of fluxes leaving states A and entering states B (see Fig. 2*D* for an illustration). An equivalent expression for the folding flux has recently been proposed in ref. 30. This network is flux-conserving, i.e. the total amount of flux F that leaves states A will enter B whereas for every intermediate state i, input flux equals output flux.

Although we will be mostly interested in the relative weight of different pathways within the folding flux, note that the absolute value of the flux still has a physical meaning. In particular, the expected number of observed $A \to B$ transitions per time unit τ is given by the total folding flux

$$F = \sum_{i\in A} \sum_{j\notin A} \pi_i T_{ij} q_j^+.$$

Another quantity of interest is the rate of the reaction $A \to B$, k_{AB}:

$$k_{AB} = F \Big/ \left(\tau \sum_{i=1}^{m} \pi_i q_i^- \right). \qquad [1]$$

Note that all states that trap the trajectory for some time will reduce k_{AB}. The effect of these traps is properly accounted for in the folding flux, even if they do not contribute to productive pathways. See the *SI Appendix* for more general transition path theory equations and derivations.

Coarse-Graining of the Folding Flux. Because the number of m conformational states used to construct a Markov model is typically large, it is convenient for illustration purposes to compute the net flux of $A \to B$ trajectories among only a few coarse sets of conformations. Let us consider the coarse partition of state space $S = \{C_1, C_2, \ldots, C_n\}$ with $n \ll m$, defined such that the boundaries of A, B, and I are preserved, i.e. A and B are either identical to individual C_i or to a collection of multiple C_i. Then we define the coarse-grained folding flux:

$$F_{ij} = \sum_{k\in C_i, l\in C_j} f_{kl}. \qquad [2]$$

$$F_{ij}^+ = \max\{0, F_{ij} - F_{ji}\}. \qquad [3]$$

The exact way of defining the sets C_i is arbitrary and depends on which features the user is interested in. For example, C_i might be chosen so as to distinguish between different secondary or tertiary structures, or they might be defined so as to separate the free energy minima (metastable states) of the system (15).

Decomposition into Individual Pathways. It may be convenient to decompose the folding flux or the coarse-grained folding flux into individual pathways P_i that connect A and B. For molecules in equilibrium, the flux can always be fully decomposed into a sum of circle-free $A \to B$ pathways. Note that this decomposition is generally nonunique, i.e. there may be many ways a given flux can be decomposed into individual pathways. However, any such decomposition has the property that the statistical questions about the order of events can computed from this set of pathways, which makes them a useful analysis tool. The simplest decomposition is obtained by starting in A and just selecting successor states in a random order until B is found, then determining the minimal net flux f along that chosen pathway and removing that pathway from the network by subtracting f from every f_{ij}^+ along the pathway. A particularly interesting way of decomposing the network is to identify the strongest pathways first (see *SI Appendix*).

The decomposition generates a set of pathways, P_i, along with their fluxes, f_i. The flux f_i provides the relative probability with which pathway i is used when considering the set of pathways P_i as possible options:

$$p_i = f_i / \sum_j f_j.$$

Order of Structural Events. The characterization of protein folding mechanisms is usually made in terms of describing an order of events. For example, in PinWW, is hairpin 1 formed before hairpin 2? For each individual pathway P_i, this question can be answered. Let E be any event whose probability $p(E)$ is to be evaluated, such as E = "hairpin 1 forms before hairpin 2", then this probability can be computed by decomposing the folding flux into pathways P_i with individual fluxes f_i and probabilities p_i, and then

$$p(E) = \sum_i p_i \delta_i(E), \qquad [4]$$

where $\delta_i(E) = 1$ if E occurs in pathway P_i and 0 otherwise. Eq. **4** can only provide the correct probability if the discretization into conformational states used is fine enough, such that the event E can be distinguished from the alternative events. In the example above, the conformational states need to be fine enough such that there is no state containing states where hairpin 1 is both formed and not formed (and likewise for hairpin 2). As $p(E)$ is a well-defined property of a given system, it is independent of the way by which the folding flux is decomposed into pathways.

Application to the Folding of PinWW

In order to illustrate the utility of our approach for studying folding mechanisms, the folding dynamics of the PinWW domain (26) is studied here. A total of 180 MD simulations were started, 100 from near-native conformations and 80 from different denatured conformations and run for 115 ns each at a temperature of 360 K. The simulations were conducted with the GROMACS program (31) by using explicit SPC solvent, the GROMOS96 force field (32), and the reaction field method for computing nonbonded forces. The simulation setup is described in detail in the *SI Appendix*. The simulated structures were aligned onto the native structure and then clustered finely into 1,734 kinetically connected and well-populated clusters. A transition matrix $\mathbf{T}(\tau)$ was constructed by counting transitions between these clusters at a lag time of $\tau = 2$ ns (see *Theory*). It was verified that $\mathbf{T}(\tau)$ is a good model for the long-time kinetics (details on the Markov model construction and validation are given in the *SI Appendix*). All properties computed from the Markov model are associated with statistical uncertainty resulting from the fact that only a finite amount of simulation data has been used to construct the model. These uncertainties are computed by using a Bayesian inference method described in ref. 33; the details are given in the *SI Appendix*. The Markov model can further be validated by comparison with kinetic experimental data recorded at the simulation temperature. The kinetic relaxation curve obtained from tryptophan (Trp) fluorescence temperature-jump experiments (26) can be compared with the relaxation from an off-equilibrium distribution of states (mimicking the situation before the T-jump) into the new equilibrium distribution monitored by a kinetic relaxation curve that is defined via the Trp solvent-accessible surface area (see *SI Appendix* for details). In both the experiment and the model, this kinetic relaxation has a fast, nonexponential decay from 1 to about 0.4, presumably resulting from fast relaxation processes that affect the Trp configurations, followed by a slow, single-exponential decay with a timescale of 26 μs in the model (confidence intervals 8–78 μs) whereas a relaxation time of 13.2 μs was computed from the experimentally determined kinetic parameters given in ref. 26.

In order to study the folding mechanism, a folded set B was defined to be the set of clusters with average backbone root mean square difference to the X-ray structure of less than 0.3 nm. The denatured set A was defined to be the set of all clusters with little β-structure (having a mean of <3 h-bonds in hairpin 1, which has 6 h-bonds in the native state, and <1 h-bonds in hairpin 2, which has 3 h-bonds in the native state). Based on these definitions and the transition matrix $\mathbf{T}(\tau)$ between the 1,734 clusters, the committor probabilities and the folding flux were computed as described in *Theory*.

In order to obtain a view of the sequence of events that is unbiased by defining reaction coordinates, the folding pathways must be considered individually. Therefore, the folding flux was decomposed into individual pathways (see *Theory*) and for each of them the times when hairpin 1 or 2 forms and remains stable were computed. "Formation" was defined as having 80% of the average number of hydrogen bonds that are present in the native state, but variations of this threshold did not change the results

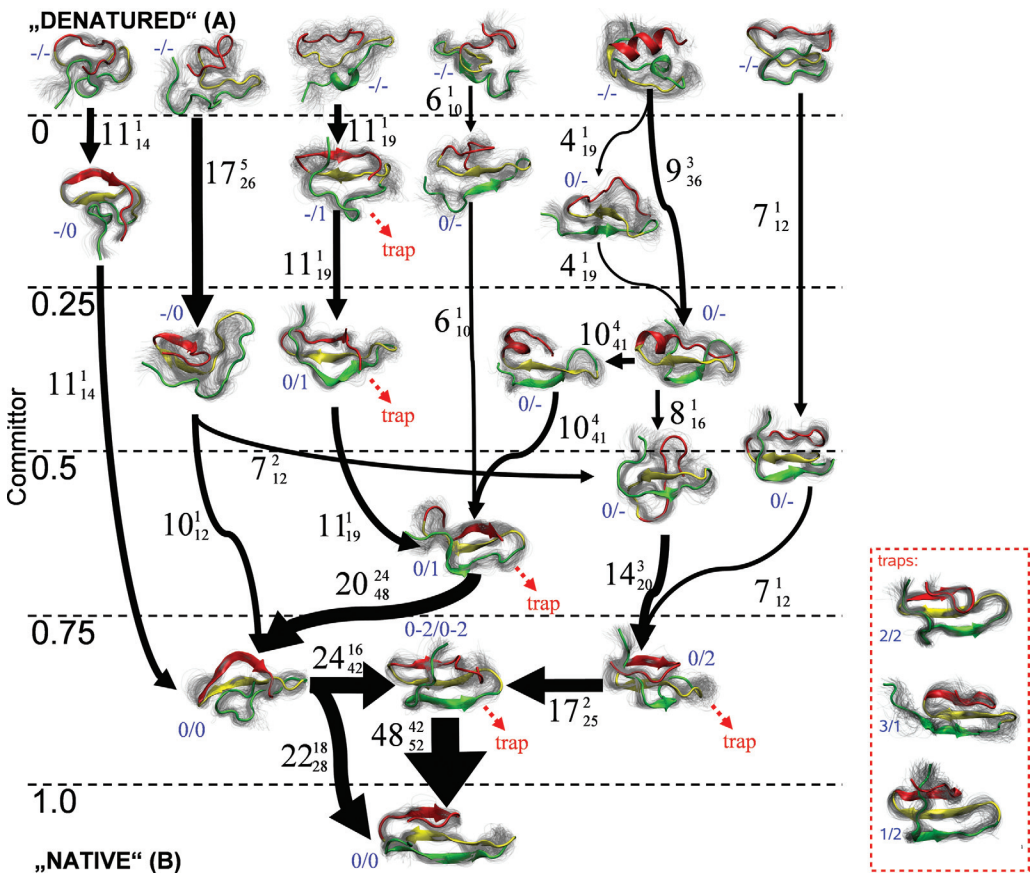

Fig. 3. PinWW folding flux. (*Left*) The network of the 70% most relevant folding pathways for PinWW. The numbers on the left indicate the committor probabilities, the thickness of the arrow indicates the flux of folding trajectories between each pair of conformations. For each conformation, a representative mean structure is shown in color along with an overlay of equilibrium-distributed structures from that conformation indicating the structural flexibility (gray cloud). The numbers next to the arrows give the normalized net flux (large number) and the 80% confidence-interval limits (small numbers) in percentages. The blue numbers next to the structures indicate whether the first/second hairpin has the native register (0), is register-shifted by one or two residues (1,2) or is not formed at all (-). (*Lower Right*) Register-shifted trap states that do not carry significant folding flux but reduce the folding speed by nearly a factor of 2.

qualitatively. The probability that hairpin 1 forms before hairpin 2 was computed from Eq. **4**: In 30% of the folding trajectories, hairpin 1 forms before hairpin 2 (confidence interval 18%–34%) and in 70% it is the other way around. Thus, there is no unique mechanism in terms of the order of secondary structure formation, which is in qualitative agreement with a structural interpretation of mutational Φ values for the PinWW domain (34).

In order to visualize the "essential folding pathways," coarse conformational sets were defined, onto which the folding flux was projected (see *Theory*). We employed a definition of 50 sets that separate the most slowly converting ("metastable") parts of state space. The number of sets can be chosen by fixing a time scale of interest (here 100 ns); then the number of metastable sets are given by the number of implied time scales of the transition matrix slower than that time scale of interest (15). The definition of metastable states can be obtained from eigenvectors of the transition matrix $\mathbf{T}(\tau)$ as suggested in refs. 15, 18, and 35 (see *SI Appendix* for details). Fig. 3 shows the network of the 70% most relevant pathways, which involves only 21 of these 50 conformational sets. The remaining 30% of the flux is mainly in small pathways between the structures shown in Fig. 3 and is omitted here for clarity of the visualization. The 29 of the 50 conformational sets not shown in the figure are only weakly involved in the $A \rightarrow B$ flux.

The denatured set (A) consists of mostly globular structures. No completely stretched structures are observed in the simulation. The coarse-grained folding flux suggests that there is a large number of unfolded states and early intermediates that narrow down when coming closer to the native state. The picture reemphasizes the existence of many structurally different parallel pathways. Pathways where hairpin 1 forms first are shown on the right, pathways where hairpin 2 forms first on the left. It is apparent that the pathways in which hairpin 1 forms first also include some partially helical structures formed by the sequence that will later become the third β strand.

Fig. 3 also indicates whether a set of structures with hairpins formed has the same register pattern as in the native state (0) or is register-shifted by one or two residues (1,2). Most of the productive folding pathways proceed from no hairpins over on-register intermediates to the native state. Some of the folding-efficient structures have the smaller hairpin 2 register-shifted, but none of them have hairpin 1 register-shifted. A special case is a structure which has both chain ends curled in such a way that they are on-register near the termini but register-shifted by 2 residues in between (indicated by "0–2").

For the 50 coarse states defined here, the coarse flux network was decomposed into individual pathways according to decreasing flux as described in *Theory*. The *Upper* frame of Fig. 4 shows the cumulative flux depending on the number of pathways, showing that about 3–5 pathways are needed to carry 50% of the total flux and about 11–20 pathways are needed to carry 90% of the total flux. Although the absolute number of parallel pathways depends on the number of states one defines, i.e., on the amount of coarse-graining, the structural differences between the 50 sets defined here imply a remarkable degree of parallelness of the folding mechanism in the present system.

The six pathways which carry most of the total flux are depicted in the *Lower* frame of Fig. 4, highlighting that there are routes where hairpin 1 forms first (paths 3,4,6), where hairpin 2 forms first (paths 1,2), and where there is a more or less concurrent formation of both (path 5). Note that the percentages of individual pathways given in Fig. 4 should not be misinterpreted as the absolute probability of finding the exact sequence of conformations. For example, these pathways do not consider the possibility of recrossing events or changing between different paths. However, these percentages do provide the relative probabilities of choosing each folding pathway from the ensemble of productive folding pathways. For example, pathway 1 is nearly twice as probable as pathway 6.

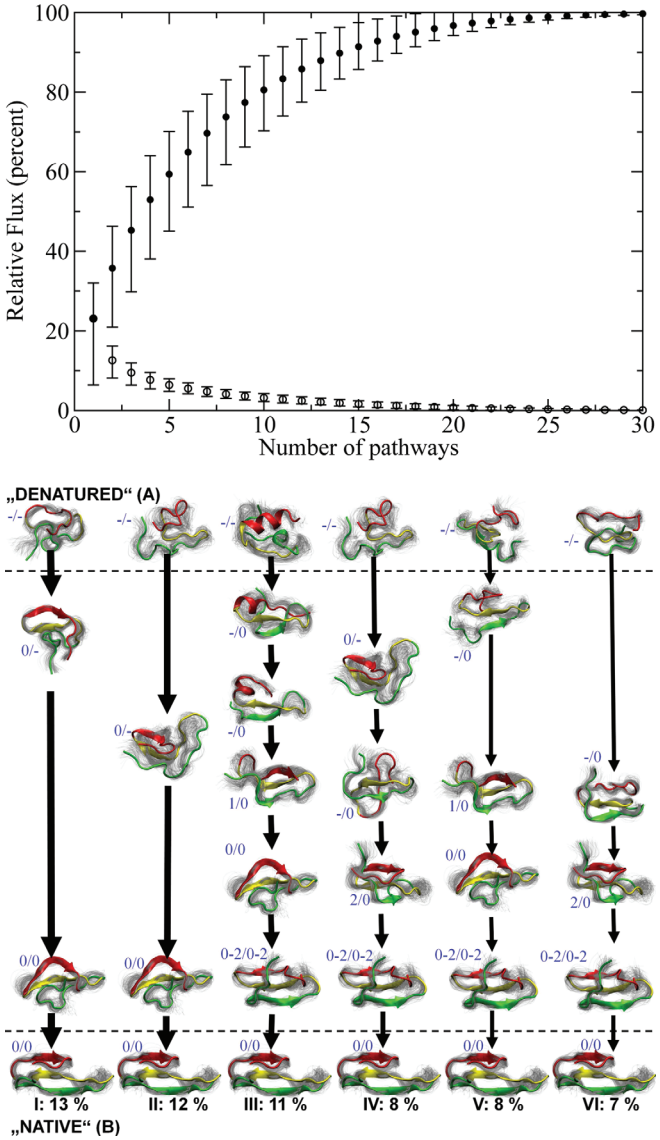

Fig. 4. PinWW folding pathways. (*Upper*) Fluxes of individual pathways and the cumulative flux. The bullets indicate the mean of the distribution, and the error bars mark the 80% confidence interval. (*Lower*) The six individual pathways that carry most of the total flux (nearly 60%).

Interestingly, there are three metastable sets that contribute almost no folding flux (<5%), but the system still spends a significant fraction of the time in them (stationary probability 18% with confidence intervals 3%–45%). These "trap" states, depicted in Fig. 3, have almost full β content, but the hairpins are register-shifted with respect to the native structure, in particular at hairpin 1, which is not fully shifted in any of the intermediates that significantly contribute to the folding flux. The effective flux (Eq. **2**), reveals that these traps are accessible from different metastable states, all of which already have a register shift in hairpin 2 or a partial register shift in hairpin 1 (see Fig. 3).

Removing the trap states from the Markov model increases the absolute folding rate k_{AB} (Eq. **1**), by almost a factor of 2, showing that there is a significant probability that the system gets stuck in one of the trap states for some time.

Conclusions

This study suggests an approach to extract the full equilibrium ensemble of folding pathways from short simulations started out of equilibrium.

The approach was applied to MD simulations of the PinWW miniprotein whose individual length was 2 orders of magnitude below the slowest time scale in the system. Nevertheless, the kinetic model computed here is consistent both with the detailed short-time information contained in the MD simulations and the long-time observables available from experiments.

The results suggest that at the simulation temperature (360 K), there is no unique or statistically dominant folding pathway in terms of the order of secondary structure formation, although hairpin 2 demonstrates a preference for forming before hairpin 1 (70% probability). The ϕ values being larger for hairpin 1 than for hairpin 2 in refs. 26 and 27 indicates a preference for hairpin 1 forming prior to hairpin 2 at temperatures around 320 K, and thus a reverse preference of pathways at these temperatures. This indication is consistent with the present results in that the larger hairpin 1 is expected to have both a larger enthalpic contribution stabilizing its folded state and a larger entropic contribution destabilizing it, which would increase upon increasing temperature. The folding pathway ensemble also includes pathways where both hairpins are partially formed and pathways that include transient α-helical structure.

Outside the efficient folding flux network, trap states have been found comprising register-shifted, i.e. "misfolded" structures. The trap is accessible from several folding-efficient conformations, which have a small degree of register-shift already. The presence of the trap slows down folding by nearly a factor of two. The existence of this trap should be experimentally testable, e.g., with time-dependent IR techniques using site-specific labeling techniques sensitive to register-shifted conformations of hairpin 1. The model also predicts that destabilizing these trap states by mutations that would disfavor misregistered hairpin configurations could speed up the folding significantly.

Although the present results allow no direct conclusions regarding the folding kinetics of PinWW at body temperature, they do suggest that protein folding kinetics may often be more parallel and more complex than intuitively expected from experimental evidence of two-state thermodynamics and kinetics. This finding agrees with the fact that conformational heterogeneity has been observed in a few experiments that have been designed to look for it (36, 37). Moreover, the picture we suggest here is fully compatible with the widely accepted folding-funnel model (38, 39), which suggests a narrowing down of a large conformational heterogeneity to the native conformations via parallel routes. The existence of parallel pathways is also supported by protein folding experiments (6, 7, 36, 37). As a result, our study suggests that a purely mechanistic question, such as "In which order do secondary structure elements fold?" is ill-defined and should rather be replaced by a probabilistic question, such as "What is the probability of a particular order of structure-formation events under a particular set of conditions?"

The combination of simulation, Markov models, and transition-path theory suggested here is a practical route to study the detailed folding kinetics of small to moderately sized macromolecules for which a sufficiently large number of explicit solvent trajectories can be generated on available parallel computing facilities. The approach avoids a biased or oversimplified model that would result from projection onto a single or few predefined reaction coordinates and could thus pave the way toward new insights into protein folding kinetics.

In a broader sense, the present approach will be useful to study the kinetics of many other biophysical processes, including protein misfolding, conformational transitions in the native state, complex enzymatic reactions, protein:ligand binding, and protein:protein aggregation.

ACKNOWLEDGMENTS. The authors are indebted to John D. Chodera, Vijay Pande, Martin Gruebele, and Volker Knecht for enlightening discussions. F.N. and C.S. acknowledge German Science Foundation for funding through the research center Matheon.

1. Kennedy D, Norman C (2005) So much more to know. *Science* 309:78–102.
2. Dill KA, Ozkan BS, Shell SM, Weikl TR (2008) The protein folding problem. *Annu Rev Biophys* 37:289–316.
3. Feng H, Zhou Z, Bai Y (2005) A protein folding pathway with multiple folding intermediates at atomic resolution. *Proc Natl Acad Sci USA* 102:5026–5031.
4. Cellitti J, Bernstein R, Marqusee S (2007) Exploring subdomain cooperativity in t4 lysozyme ii: Uncovering the c-terminal subdomain as a hidden intermediate in the kinetic folding pathway. *Protein Sci* 16:852–862.
5. Friel CT, Beddard GS, Radford SE (2004) Switching two-state to three-state kinetics in the helical protein im9 via the optimisation of stabilising non-native interactions by design. *J Mol Biol* 342:261–273.
6. Goldbeck RA, Thomas YG, Chen E, Esquerra RM, Kliger DS (1999) Multiple pathways on a protein-folding energy landscape: kinetic evidence. *Proc Natl Acad Sci USA* 96:2782–2787.
7. Matagne A, Radford SE, Dobson CM (1997) Fast and slow tracks in lysozyme folding: Insight into the role of domains in the folding process. *J Mol Biol* 267:1068–1074.
8. Schaeffer DR, Fersht AR, Daggett V (2008) Combining experiment and simulation in protein folding: Closing the gap for small model systems. *Curr Opin Struct Biol* 18:4–9.
9. Ensign DL, Kasson PM, Pande VS (2007) Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J Mol Biol* 374:806–816.
10. Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc Natl Acad Sci USA* 97:10780–10785.
11. Snow CD, Nguyen H, Pande VS, Gruebele M (2002) Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* 420:102–106.
12. Lei H, Wu C, Liu H, Duan Y (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc Natl Acad Sci USA* 104:4925–4930.
13. Freddolino PL, Liu F, Gruebele MH, Schulten K (2008) Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys J* 94:L75–L77.
14. Fersht AR (2002) On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc Natl Acad Sci USA* 99:14122–14125.
15. Noé F, Horenko I, Schütte C, Smith JC (2007) Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J Chem Phys* 126:155102.
16. Chodera JD, et al. (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 126:155101.
17. Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B* 112:6057–6069.
18. Schütte C, Fischer A, Huisinga W, Deuflhard P (1999) A direct approach to conformational dynamics based on hybrid monte carlo. *J Comput Phys* 151:146–168.
19. Chodera JD, Swope WC, Pitera JW, Dill KA (2006) Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model Simul* 5:1214–1226.
20. Jayachandran G, Vishal V, Pande VS (2006) Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J Chem Phys* 124:164902.
21. Krivov SV, Karplus M (2004) Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc Natl Acad Sci USA* 101:14766–14770.
22. E W, Vanden-Eijnden E, (2006) Toward a theory of transition paths. *J Stat Phys* 123:503–523.
23. Metzner P, Schütte C, Vanden–Eijnden E (2009) Transition path theory for Markov jump processes. *Multiscale Model Simul* 7:1192–1219.
24. Du R, Pande VS, Alexander, Tanaka T, Shakhnovich ES (1998) On the transition coordinate for protein folding. *J Chem Phys* 108:334–350.
25. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* 53:291–318.
26. Jäger M, Nguyen H, Crane JC, Kelly JW, Gruebele M (2001) The folding mechanism of a beta-sheet: The ww domain1. *J Mol Biol* 311:373–393.
27. Deechongkit S, et al. (2004) Context-dependent contributions of backbone hydrogen bonding to β-sheet folding energetics. *Nature* 430:101–105.
28. Jäger M, et al. (2006) Structure-function-folding relationship in a ww domain. *Proc Natl Acad Sci USA* 103:10648–10653.
29. Liu F, et al. (2008) An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc Natl Acad Sci USA* 105:2369–2374.
30. Berezhkovskii A, Hummer G, Szabo, (2009) A reactive flux and folding pathways in network models of coarse-grained protein dynamics. *J Chem Phys* 130:205102.
31. van der Spoel, et al. (2005) GROMACS: Fast, flexible and free. *J Comput Chem* 26:1701–1718.
32. van Gunsteren WF, Berendsen HJC (1990) Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. *Angew Chem Int Ed* 29:992–1023.
33. Noé F (2008) Probability distributions of molecular observables computed from Markov models. *J Chem Phys* 128:244103.
34. Weikl TR (2008) Transition states in protein folding kinetics: Modeling phi-values of small beta-sheet proteins. *Biophys J* 94:929–937.
35. Weber M (2003) *Improved Perron Cluster Analysis*. (Zuse Institute, Berlin) ZIB Report 352.
36. Lindberg MO, Oliveberg M (2007) Malleability of protein folding pathways: A simple reason for complex behaviour. *Curr Opin Struct Biol* 17:21–29.
37. Mello CC, Barrick D (2004) An experimentally determined protein folding energy landscape. *Proc Natl Acad Sci USA* 101:14102–14107.
38. Dill KA, Chan HS (1997) From levinthal to pathways to funnels. *Nat Struct Mol Biol* 4:10–19.
39. Wolynes PG, Onuchic JN, Thirumalai D (1995) Navigating the folding routes. *Science* 267:1619–6616.