THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Evaluation and Optimization of Discrete State Models of Protein Folding

Elizabeth H. Kellogg,[†] Oliver F. Lange,[‡,§] and David Baker*[,†]

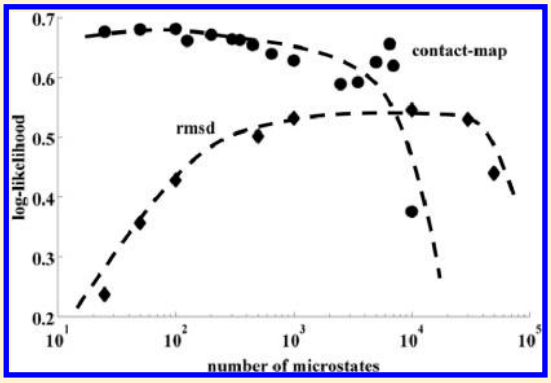
[†]Department of Biochemistry, University of Washington, Seattle, Washington 98105, United States

[‡]Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, 85747 Garching, Germany

[§]Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

Supporting Information

ABSTRACT: The space accessed by a folding macromolecule is vast, and how to best project computer simulations of protein folding trajectories into an interpretable sequence of discrete states is an open research problem. There are numerous alternative ways of associating individual configurations into collective states, and in deciding on the number of such clustered states there is a trade-off between human interpretability (smaller number of states) and accuracy of representation (larger number of states). Here we introduce a trajectory likelihood measure for assessing alternative discrete state models of protein folding. We find that widely used rmsd-based clustering methods require large numbers of initial states and a second agglomeration step based on kinetic connectivity to produce models with high predictive power; this is the approach taken in elegant recent work with Markov State Models of protein folding. In contrast, we find that grouping of states based on secondary structure pairings or contact maps, when refined with K-means clustering, yields higher likelihood models with many fewer states. Using the most predictive contact map representation to study the folding transitions of the WW domain in very long molecular dynamics simulations, we identify new states and transitions. The methods should be generally useful for investigating the structural transitions in protein folding simulations for larger proteins.



INTRODUCTION

Exciting breakthroughs in computer hardware have made possible simulation of proteins for time scales longer than the time required to fold, allowing observation of multiple folding and unfolding events.[1,2] Simulations spanning the ~microsecond folding times of even the smallest proteins require trajectories of $>10^6$ configurations, and methods for associating these configurations into a much smaller number of significant states are of considerable importance for analyzing the key structural transitions.[3,4] Traditional methods have required projection of simulation data onto one- or two-dimensional reaction coordinates,[5−19] but thus can obscure important features of the folding free-energy landscape.[20]

Markov State Models (MSMs) have overcome this limitation by representing dynamics as a network of transitions between discrete states and have been used to analyze folding pathways for many proteins of interest.[21−31] A MSM is constructed by first using geometric similarity to combine very similar configurations into discrete microstates, assuming that high geometric similarity implies high kinetic connectivity. Subsequently, these microstates are assembled into sets of kinetically related microstates, called macrostates. Care must be taken in defining the initial set of states, as incorrect initial grouping of configurations can result in incorrect conclusions about folding dynamics if, for example, a microstate contains configurations separated by high-energy barriers.[32]

There are many open questions in constructing MSMs of protein folding. In order to assign configurations from different trajectories to discrete states, some clustering based on geometric similarity is clearly necessary. What distance measure to use, to what extent configurations should be clustered, and how this impacts the resulting model are not fully understood. To address these and related problems, quantitative metrics for evaluating alternative models are needed.[33,34] Here we describe a likelihood measure for assessing alternative MSMs and investigate the trade-off between geometric and kinetic based lumping as well as alternative structural similarity metrics for grouping configurations.

METHODS

Representations and Distance Measures. *Secondary Structure Pairing.* Hydrogen-bonding patterns were identified and classified for each configuration using the Rosetta software.[35−38] The peptide conformation in each trajectory

Received: May 7, 2012
Revised: July 27, 2012
Published: August 13, 2012

   dx.doi.org/10.1021/jp3044303 | J. Phys. Chem. B 2012, 116, 11405−11413

snapshot (configuration) was represented by a feature vector describing the sets of paired strands and their registers, pleatings, and orientations. Configurations with a common feature vector are assigned to the same microstate. Details of the feature descriptions and the assignment procedure are in the Supporting Information. Depending on the detail of the description the number of distinct feature vectors ranged from 175 to 4857.

Contact Map. Residues separated by more than one residue in sequence were considered in contact if the Cα residues were within 8 Å. Clustering of contact maps was performed according to the greedy K-centers algorithm,[39] using Euclidean distance to measure similarity. K-means refinement[40] was performed for 20 iterations using the greedy K-centers cluster assignments as input.

Rmsd. Cα coordinates were clustered using greedy K-centers clustering and root-mean-squared-deviation (rmsd) as the distance measure as described previously.[27,41]

MSM Model Construction. Transition matrices were constructed with a lag time of 100 ns, as determined from convergence of implied time scale plots[26] (Supporting Information, Figure S1). Macrostates were obtained from the eigenvectors of the transition matrices using Perron clustering[42] as described previously (see Supporting Information).

Log-Likelihood Metric. We define the likelihood of a Markov Model given a set of trajectories to be the probability of observing the trajectories given the model. Given a set of assignments of configurations $\{c_t\}$ to states s_t, the probability of a particular configuration trajectory is given by a two-level Markov Model

$$p(\{c_t\}) = p(\{s_t\})p(\{c_t\}|\{s_t\}) = p(s_1)\prod_{i=1}^{N-1} p(s_{t+1}|s_t)\prod_{t=1}^{N} p(c_t|s_t)$$

where s_t is the state assigned to configuration c_t occupied at step t in the trajectory, $p(s_t)$ is the probability of state s_t, $p(s_{t+1}|s_t)$ is the probability of transitioning to state s_{t+1} from state s_t, and $p(c_t|s_t)$ is the probability of configuration c_t given that the system s in state s_t. $p(\{c_t\})$ is the probability of the configuration trajectory, $p(\{s_t\})$ is the probability of the state trajectory, and $p(\{c_t\}|\{s_t\})$ is the probability of the configuration trajectory given the state trajectory.

Computation of the likelihood requires estimation of the above quantities based on the training data. $p(s_1)$ is well modeled by the frequency of state s_1, n_1/N_{total}, where n_1 is the number of configurations in state 1 and N_{total} the total number of configurations. $p(s_{t+1}|s_t)$ is well modeled by the transition frequency from state t to state $t+1$ observed in the trajectories. A simple choice for $p(c_t|s_t)$ that follows from the assumption that configurations are uniformly distributed within the states is the inverse of the phase space volume spanned by state s_t. However, it is difficult to compare cluster volumes between the different structural representations (rmsd, contact map, strand pairings). Instead, we chose to estimate $p(c_t|s_t)$ as $1/n_t$, where n_t is the number of configurations assigned to state s_t. This choice has the obvious benefit that models in different representations can be readily compared, and the further advantage that the likelihood of the null model (see below) is independent of the number of states. However, the implicit assumption that configurations are uniformly distributed in phase space (so the volume becomes proportional to the number of configurations in the cluster) is clearly an oversimplification. Improving the estimate of $p(c_t|s_t)$ is an important area for future work.

We normalize by computing the likelihood of a null model in which all states have equal size and the probability of all transitions are equal. For the null model with m states and N_{total} total configurations, the probability of observing a configuration given a state is m/N_{total} and the probability of transitioning to any state is $1/m$, so the probability at each step of the null-model trajectory is $(m/N_{\text{total}} \times 1/m) = 1/N_{\text{total}}$ as it should be since all configurations are equally probable. To avoid round-off errors, we compute the log-likelihood instead of the likelihood, and subtract the log-likelihood of the null model, yielding the final form of the log-likelihood shown in the figures:

$$\log(p(\text{traj}|\text{MSM})) = \log\left(\frac{n_1}{N_{\text{tot}}}\right) + \sum_{t=1}^{N_{\text{test}}-1} \log(T(t, t+1)) + \sum_{t=1}^{N_{\text{test}}} \log\left(\frac{N_{\text{tot}}}{n_t}\right)$$

Cross-Validation Procedure. We experimented with a number of ways to cross-validate Markov State Models using independent trajectory data. The most rigorous is to construct an MSM purely from the training set data, and then assign each test set configuration to a training set microstate based on geometric similarity. This turned out to be computationally intractable since obtaining good statistics required many repeated cross-validation calculations with different randomly selected training set/test set partitions, and it was not feasible to carry out the microstate clustering large numbers of times for each parameter set and representation considered. We settled on a considerably more tractable approach in which the entire data set is clustered to obtain microstate definitions, but the transition matrix construction and subsequent spectral clustering to obtain macrostate definitions are based on training set data only. Microstates observed in the test set but not in the training set are reassigned to the closest microstate in the training set based on geometric similarity. The model quality metric, which we refer to throughout the rest of the paper, is the log-probability of observing the test set data given the model constructed from the training set data. Cross-validation was performed a total of 1000 times for each model. One randomly selected segment of the data comprising a total of 1% of the total data was removed prior to transition matrix construction and spectral clustering but after the initial geometric clustering step, and the 99% which remained was used to compute the transition matrix and state occupancies. To reduce sensitivity to the original clustering results for models with less than 10 000 microstates, the cross-validation procedure was repeated for 4−10 different initial clusterings obtained using different random seeds.

Likelihoods of microstate models were computed using the microstate occupancies and microstate transition probabilities computed from the training set. Similarly, likelihoods of macrostate models were computed using macrostate occupancies and macrostate transition matrices computed solely from the training set.

RESULTS

We begin by briefly describing the steps involved in constructing a Markov State Model (MSM) from molecular dynamics or Monte Carlo trajectories.[32,33,43] The first step involves discretizing the simulation; the individual configu-

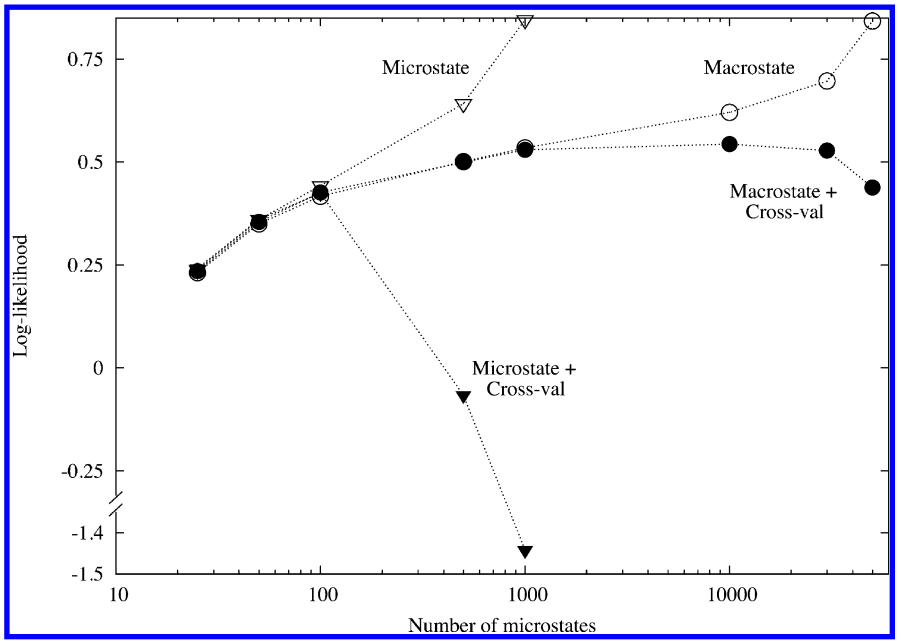


Figure 1. Assessment of rmsd-based MSMs using the likelihood metric. Triangles, microstate models; circles, macrostate models. Open symbols represent training set likelihoods, closed symbols, test set (cross-validated) likelihoods. The number of microstates is indicated on the *x*-axis; the clustering radius decreases from 8.1 Å for the 20-microstate model to 4.4 Å for the 5×10^4-microstate model. For small numbers of microstates all models have similar likelihoods. For more than 100 microstates, the cross-validated likelihood drops steeply but is rescued by lumping of the microstates into macrostates. To generate the macrostate models, microstates were lumped into 20 macrostates using Perron clustering.

rations representing trajectory snapshots are clustered based on geometric similarity without consideration of their kinetic proximity. The resulting clusters are called "microstates". The frequencies of transitions within and between microstates are computed, populating a count matrix. Groups of kinetically related microstates, called macrostates, are then identified from the eigenvectors of the resulting transition matrix by Perron clustering.[42] The macrostates and transitions among them constitute a reduced complexity description of the longer time scale dynamics of the original system.

The process of discretizing a simulation is a nontrivial task and involves a number of critical choices.[44] If the clustering is too coarse (too few clusters), the resulting assignments will likely contain configurations separated by large kinetic barriers. On the other hand, if clustering is too fine (too many clusters), the resulting transition matrix will quickly become sparse, resulting in bad statistics and reduced generalizability. In the limit, each configuration is in its own microstate and the resulting model has no predictive power. In order to assess the choices made during assignment, a metric is needed for assessing the extent to which the resulting MSM accurately represents the dynamics of the system.

We have found that a simple likelihood statistic coupled with cross-validation provides a very useful model-quality metric (see Methods). Given a choice of distance metric and geometric clustering threshold (number of microstates), we partition the data into training and testing sets and compute the log-probability of observing the test set data given the training set data using transition matrices compiled exclusively from the training set (see Methods).

To investigate the utility of the log-likelihood statistic to guide MSM construction and evaluation, we used the rmsd-based clustering approach pioneered by the Pande group[22,41,42,45,46] and others[32,47,48] to build MSMs from the 200 μs MD trajectories of the WW domain from the DE Shaw group.[1] We compared the likelihood of the microstate-based models resulting from the initial geometric clustering step to the likelihood of macrostate models produced by lumping microstates together using Perron clustering. Without cross-validation, the likelihoods of both the microstate and macrostate models increase monotonically with increasing numbers of microstates (open symbols, Figure 1). With cross-validation, the likelihood decreases sharply for higher number of microstates as the predictive power of the model (generalizability) decreases (closed symbols, Figure 1). The macrostate models retain high likelihoods, indicating greater generalizability even with higher numbers of microstates (compare closed circles and triangles, Figure 1). This result reinforces the idea, central to the motivation for MSMs, that grouping based on kinetic connectivity is superior to clustering based solely on geometric similarity.

The cross-validated likelihood statistic provides a metric for evaluating alternative ways to construct macrostate MSMs. Given a fixed number of final macrostates, the coarseness of the initial geometric clustering is controlled by varying the number of microstates. With increasing number of microstates, the geometric lumping is less coarse, and correspondingly, the final macrostate composition is more dominated by kinetic connectivity. As shown in Table 1, all macrostate models have higher log-likelihoods than the corresponding microstate model with the same final number of states (Table 1, "rmsd"). The cross-validated log-likelihood of the rmsd-based macrostate model reaches a maximum between 1000 and 10 000 microstates, which correspond to 6.5 and 5.3 Å rmsd cluster radii, respectively (Figure 1, closed circles). In practice, for such a small protein system, configurations with significant differences (i.e., strand register shifts) can have rmsds much less than 5.3 Å rmsd and hence be assigned to the same microstate even though this violates the assumption that geometric similarity implies kinetic similarity. Rmsd-based clustering thresholds

dx.doi.org/10.1021/jp3044303 | J. Phys. Chem. B 2012, 116, 11405−11413

Table 1. Trade-off between Kinetic versus Geometric Lumping As Measured by Log-Likelihood[a]

representation	no. initial states	no. final states	log-likelihood
rmsd	20	20	0.21
	100	20	0.43
	1000	20	0.53
	10000	20	0.55
	175	20	0.68
secondary structure	175	10	0.56
	175	175	0.54
	20	20	0.05
K-centers contact map	100	20	0.25
	1000	20	0.52
	10000	20	0.31
	20	20	0.84
K-means contact map	100	20	0.68
	1000	20	0.63
	10000	20	0.36

[a]For each representation (column one), the log-likelihood (column four) is measured as a function of the number of initial geometrically defined states (column two) and the number of kinetically clustered final states (column three). If the number of initial states is equal to the number of final states, the model was constructed purely using geometric clustering. The more initial states, the finer the partitioning of space before the kinetic-clustering step.

below 3 Å may well be necessary to preserve kinetic relationships, but partitioning the data this finely would result in orders of magnitude more microstates and a poorly determined transition matrix, decreasing the generalizability of the model (this may be responsible for the small decrease in likelihood for 50 000 microstates evident in Figure 1).

We reasoned that more economical models might be obtainable using geometric similarity measures other than rmsd to group configurations into microstates. With the aim of building discrete-state models for much larger systems, we first considered a variety of reduced representations in which configurations are described by their secondary structure pairings (see Methods section and Supporting Information). With orders of magnitude less initial states, the resulting macrostate models (same number of final states) had significantly higher likelihoods (Table 1, "ss-pair") than the corresponding rmsd-based models (Table 1, rmsd). The secondary structure pairing representation captures key features of the dynamics and thus provides more kinetically relevant microstate definitions—this is not surprising as previous descriptions of WW domain folding have focused on the formation of the two hairpin structures.[1,27,49,50]

While the secondary structure-based models appear to be more economical in representing the dynamics, they are unable to describe contributions from more general interactions, like hydrophobic core formation.[35] On the other hand, rmsd-based models are more general and the resolution of clustering is easily controlled, but kinetically irrelevant structural elements such as flexible loops and termini may contribute significantly to state assignment whereas formation of hydrogen bonds might be missed, resulting in the need for large numbers of

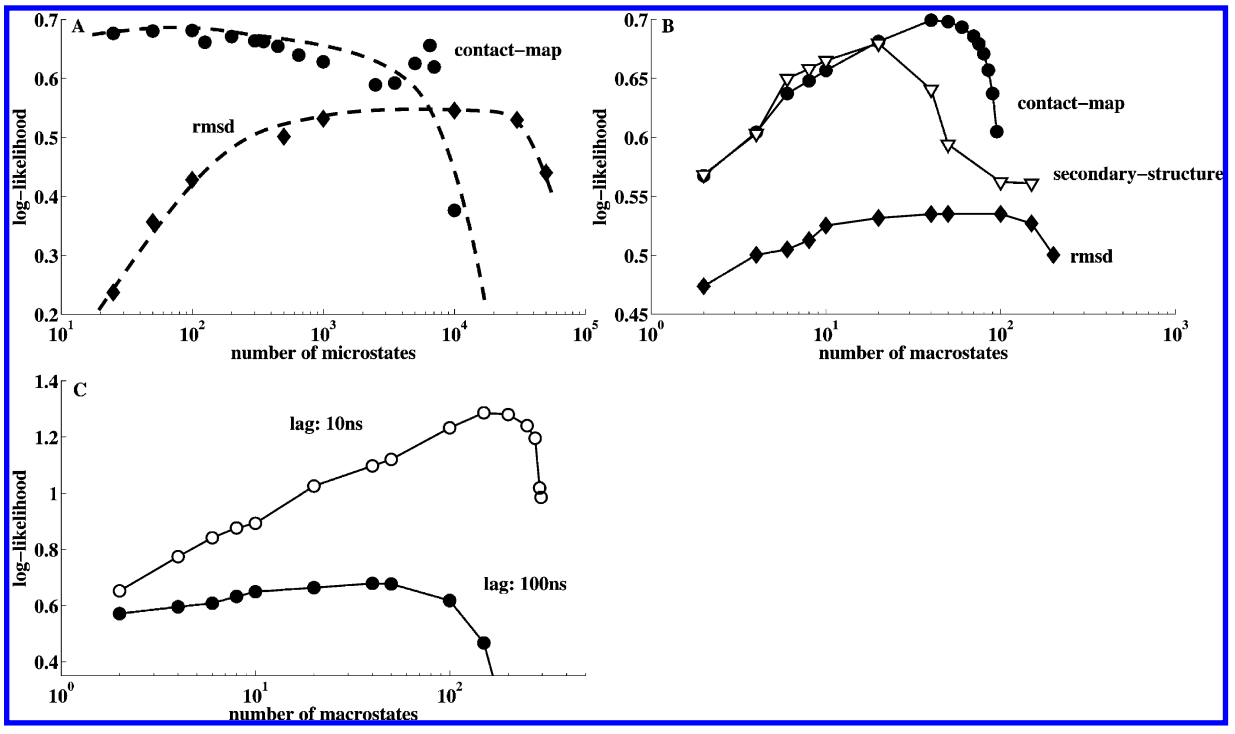


Figure 2. Evaluation of MSMs using the likelihood metric. (A) Dependence of likelihood on the number of microstates. Configurations were first grouped into the indicated number of microstates and these were then further lumped into 20 macrostates using Perron clustering. Diamonds, rmsd-based microstate assignments; circles, K-means contact map-based assignments. Lines are manually drawn to guide the eye. (B) Dependence of likelihood on number of macrostates. For each representation, models were constructed using the number of microstates producing the highest likelihood in panel A (diamonds, rmsd 1000; circles, K-means contact map 100; and triangles, secondary structure pairing 175). (C) More states are required to model short time scale dynamics. For the K-means contact map representation with 300 microstates, the dependence of the log-likelihood on the number of macrostates is shown for a lag time of 100 ns (filled triangles) and 10 ns (open circles). The optimal number of macrostates is higher at shorter lag times.

dx.doi.org/10.1021/jp3044303 | J. Phys. Chem. B 2012, 116, 11405−11413

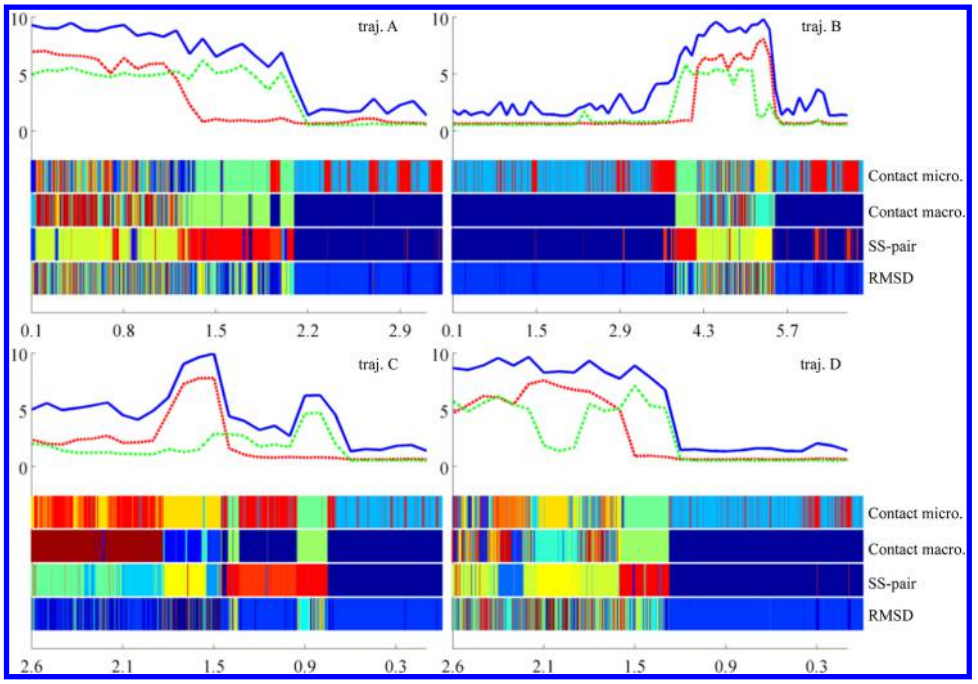


Figure 3. Comparison of microstate and macrostate trajectories in the three representations. The panels show portions of trajectories bracketing four folding/unfolding transitions. Upper portion of panels: blue, rmsd to native structure over residues 1−34; red, rmsd to native over strand one; and green, rmsd over strand 2. Lower portion of panels: macrostate trajectories for (first row) a 25-state K-means refined contact map microstate model, (second row) a 20-macrostate contact map-based model (100 microstates), (third row) a 20-macrostate secondary structure-pairing-based model (175 microstates), and (fourth row) a 20-macrostate rmsd-based model (1000 microstates). Each color represents a different macrostate.

states to accurately describe the dynamics. To combine the advantages of the secondary structure and rmsd-based representations, we explored a contact map-based representation (see Methods). We reasoned this representation would capture strand pairings and registers, like the secondary structure model, while remaining general by also capturing important long-range contacts not necessarily involving hydrogen bonding. At the same time, such a representation would be less sensitive to loop and termini fluctuations than rmsd-based models.

We experimented with two clustering methods for grouping configurations into microstates using the contact map-based representation. The likelihoods of models constructed from microstates obtained using the very fast greedy K-centers method (Table 1) were similar to those of the rmsd-based models, but not as high as the secondary structure pairing models. We reasoned that more accurate clustering of configurations into microstates could produce better models, and took advantage of the fact that the contact map representation allows a simple definition of the cluster center—the average of all the contact maps in a cluster—and used K-means optimization to refine the cluster boundaries (Table 1, Figure 2A). The likelihoods of models produced using K-means clustering were considerably higher than models created with K-centers clustering (Table 1). Improved state definitions, as expected, yield more predictive models.

The optimal number of microstates differs considerably in the different representations. For the rmsd and K-centers contact map representations, the highest likelihoods are obtained for models with 1000−10 000 microstates. In contrast, for the secondary structure pairing and K-means refined contact map microstate definitions, the optimal number of microstates is between 100 and 200, and the initial partition into microstates represents the dynamics better for the latter models than the former (compare Table 1 "rmsd" and "K-means contact map"; the log-likelihoods of the K-means contact map microstate models are similar to those of the K-means contact map macrostate models with the same number of states). The large numbers of microstates for the best rmsd-based models is likely necessary to avoid excessively large cluster radii (∼8 Å in the 20-microstate model) (see Figure 1 legend). The contact map and secondary structure pairing models have significantly higher likelihoods while requiring many fewer microstates probably because clustering in these representations better preserves kinetic connectivity. Improved clustering methods can yield significantly higher likelihood models as illustrated by the difference in likelihood of the contact map K-centers and K-means based models. Thus, improved methods of rmsd-based cluster refinement[45] could perhaps yield significantly higher likelihood models than the rmsd models constructed in this study. The relatively small number of microstates required to build a predictive model using the contact map and secondary structure pairing representations (100−300) suggests that building accurate models of the folding of larger proteins with these approaches should be feasible.

We next considered the dependence of the likelihood on the number of macrostates. We anticipated that the log-likelihood would initially increase with number of macrostates due to the improved representation of folding dynamics but subsequently decrease due to overfitting to the training data. This was indeed observed. As shown in Figure 2B (circles), the likelihood peaks at 40 macrostates for K-means-refined contact assignments, 40−100 for rmsd assignments, and 20 macrostates for secondary structure pairing-based assignments. These results suggest that the most predictive, and hence in a sense most accurate model of WW domain folding involves less than 100 discrete states. To investigate what effect the lag time has on

dx.doi.org/10.1021/jp3044303 | J. Phys. Chem. B 2012, 116, 11405−11413

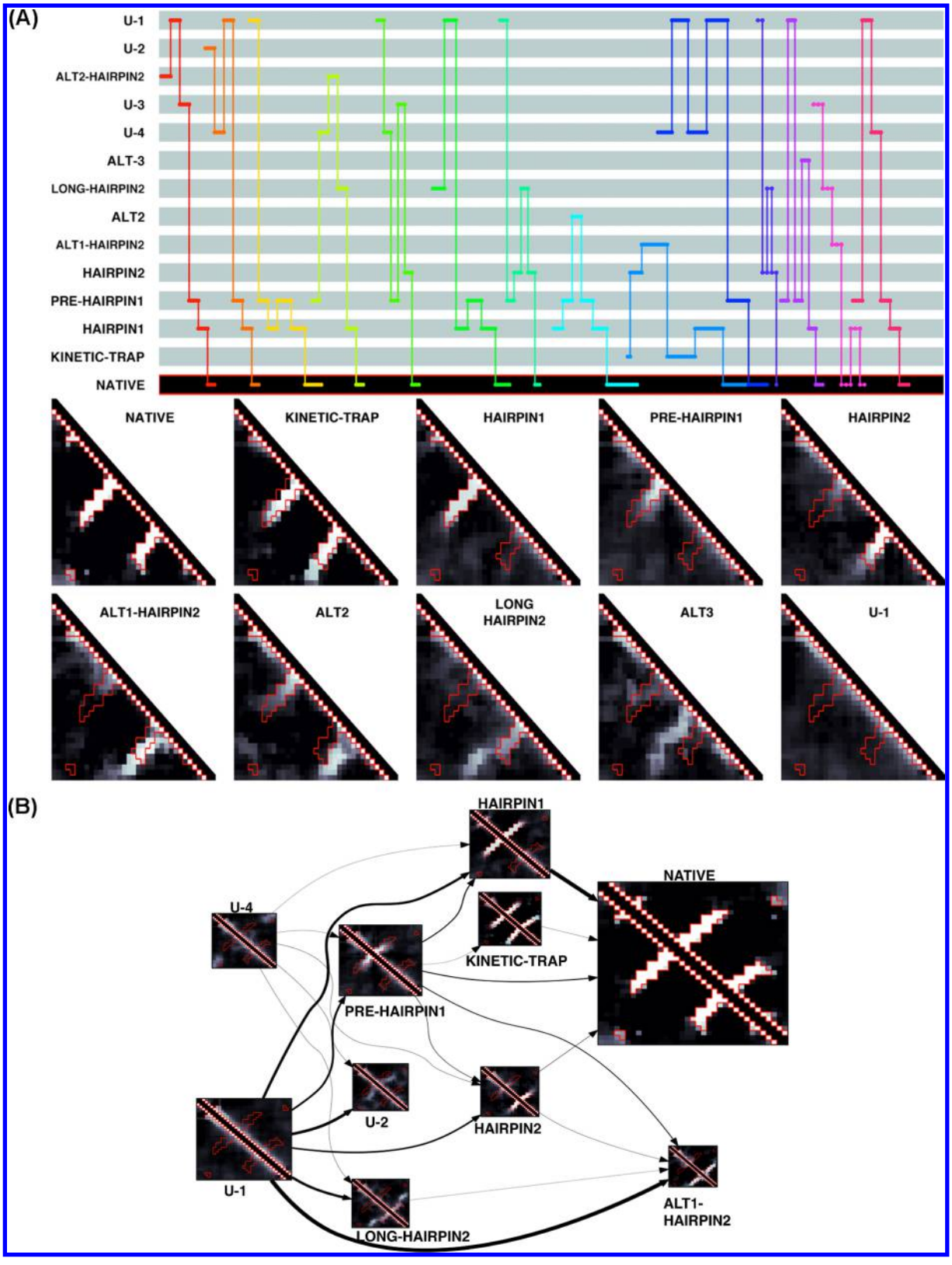


Figure 4. (A) Analysis of folding transitions. Discrete state representation of the 14 folding/unfolding events in the long time scale WW domain folding simulations. In the top panel, each color represents one of the 14 unfolding−folding transitions, and the lines depict the sequence of states visited enroute to the native state. The lower panels show contact maps of the most commonly visited states in these folding transitions. While there is considerable heterogeneity early on, in most of the trajectories the "hairpin1" state immediately preceeds the transition to the native state. ALT abbreviations denote alternative structures which are non-native, ALT-HAIRPIN describe alternative macrostates which are similar to either native hairpin. U abbreviations denote macrostates with little to no regular structure. (B) Flux diagram of macrostate model constructed only from folding-transition regions. Model shown is for the 100-microstate K-means-refined contact map microstate model which is kinetically lumped into a 20-macrostate model (constructed from a 10 ns lag time). The individual macrostate names are the same as in part A. Width of arrows represent flux.

dx.doi.org/10.1021/jp3044303 | J. Phys. Chem. B 2012, 116, 11405−11413

the optimal number of macrostates, we decreased the lag time from 100 to 10 ns (Figure 2C). The maximum in the log-likelihood shifts to higher numbers of macrostates as expected since additional states are required to model the short-time dynamics of the system.

To obtain an intuitive feeling for the usefulness of the different models in providing descriptions of the folding kinetics, we compared the descriptions of the folding transitions observed in the MD simulations provided by the contact map, rmsd, and secondary structure-based models. For reference, we computed for each folding transition the rmsd to native for the global structure, the first hairpin, and the second hairpin,[1,2] and compared this to a contact map-based microstate trajectory and macrostate trajectories produced by the highest log-likelihood rmsd, contact map, and secondary structure pairing models (Figure 3). In the contact map representation, there are macrostates with either the first hairpin, the second hairpin, or both (Figure 4), and the formation of the first hairpin and the second hairpin is clearly evident in the contact map macrostate trajectory (Figure 3, second row color bar). The contact map-based microstate trajectory is very similar to the macrostate trajectory (compare top color row with rows 1−2) as expected given its similar log likelihood (Table 1). The secondary structure pairing-based MSM trajectory (Figure 3, third color bar) yields similar qualitative results, but assigns a larger portion of the unfolded state to a single macrostate. The rmsd MSM trajectory, constructed using the maximum log-likelihood microstate model of 1000 microstates, does not clearly identify either of the intermediates with just 20 macrostates (Figure 3, fourth color bar), although a previous study identified these intermediates in a 200-macrostate model and approximately 26 000 microstates.[26] The full set of comparisons are displayed in the Supporting Information, Figure 2.

The contact map and strand macrostate models reveal the existence of a "kinetic-trap" state, which has a global rmsd to native of approximately 5 Å (Figure 3C color bar rows 1−3). The rmsd macrostate model, constructed from 1000 microstates, does not detect this state; instead it assigns this state to the native macrostate (Figure 3C, bottom color bar). Further analysis of this state shows that it persists for approximately 2 μs (Supporting Information, Figure 6B, yellow-green-colored state at approximately 25 ms), and consists of a shifted strand-one register but a correct strand-two register (Supporting Information, Figure 4b). The contact map and secondary structure pairing-based macrostate models evidently identify metastable states that cannot easily be detected using rmsd metrics.

We next considered how to obtain a comprehensive overview of WW domain folding from the contact map-based macrostate model. Previous work with MSMs utilized overall flux diagrams showing the flow of probability density along the model. We constructed such a model (Supporting Information, Figure 5) built from the entire set of simulation data, which is similar to that reported previously for the WW domain based on the same simulation data by Pande and co-workers with multiple paths to the native state.[26] A flux diagram focused on the transition regions (Figure 4B) had fewer connections to the native state (4 instead of 8) which is surprising given that this subset of the data includes all transitions between unfolded and native states.

To investigate this further, we examined the individual folding/unfolding transitions in the context of the contact map-based model. Figure 4 exhibits the 14 folding/unfolding transitions observed in the trajectories in the contact map-based models (details in the Supporting Information). We find that many states with residual structure are traversed before folding to native (Figure 4A). These states contain some helix or strand secondary structure and many are compact (Supporting Information, Figure 4). One state corresponds to the extension of the first hairpin to pair with the N and C-termini of the protein (Supporting Information, Figure 4M). Other states correspond to altogether non-native strand pairings (Supporting Information, Figure 4F,G,I,K,L). In accordance with this variety of non-native alternative states, we find that each non-native alternative is visited only once or twice. For 10 of the 14 transitions (Figure 4A), we find that folding proceeds through the first hairpin and for four of the pathways, folding proceeds through formation of the second hairpin intermediate.

Whereas the flux diagram represents a prediction of the folding dynamics based on the MSM transition matrix and computed p-fold values,[47] the folding-pathway reconstruction only uses the state assignment of the MSM model for analysis and otherwise relies solely on the actually observed transitions in the trajectories. While the conclusions drawn from the flux diagram and the folding-pathway reconstruction are qualitatively similar (folding proceeds along a single dominant route with formation of hairpin one), small discrepancies (the kinetic-trap state is kinetically related to the native state in the flux diagram but not the folding-pathway reconstruction) remain. These discrepancies may arise because sampling is too limited, even with the ground-breaking computing resources used for the simulations,[1] to put meaningful statistical weights on the different folding pathways or because of errors in MSM construction resulting from microstate misassignments (grouping of configurations separated by large kinetic barriers in the same microstate).

DISCUSSION

We show that the likelihood of an independent test set is a powerful metric for assessing alternative discrete state models of protein folding based on molecular dynamics simulations. As suggested by previous work with MSMs, we find that, when using rmsd, grouping configurations based on their kinetic connectivity is considerably more effective than grouping based on geometric similarity to obtain microstates. The highest likelihood models require on the order of 1000 to 10 000 microstates, in the range used in previous studies. We find that contact map and secondary structure-pairing representations are considerably more economical, achieving higher likelihoods with many fewer microstates (100 and 175, respectively). Furthermore, the K-means refined contact maps and strand-pairing representations are more effective at preserving kinetic connectivity and identify intermediates clearly.

In this paper, we used the very long MD simulations of WW domain folding to evaluate our methodological developments. Using the improved combination of contact map and clustering procedure or strand-pairing-based representations, we identified a kinetic-trap state that is clearly metastable and is not easily detected by either simpler rmsd-based metrics[1] or rmsd-based macrostate models.[26] Comparing individual folding pathways, we find considerable heterogeneity at the beginning of the transitions, but convergence in the late stages of folding, with the majority of the folding transitions involving formation of the first hairpin as suggested in the initial study.[1] Still more sampling (longer MD simulations) would be required to assign

statistical weights to the different observed folding transition pathways.

The methods described here should be generally useful for building discrete state models of folding dynamics from long time scale molecular dynamics simulations. The likelihood measure provides a means to assess alternative model formulations, and the combination of contact map representation and cluster refinement strategy should scale considerably better than rmsd-based clustering to larger systems. It will be particularly interesting to use the approach outlined here to build discrete state models based on the long time scale simulations recently reported for a number of larger proteins by Shaw and co-workers[2]—we anticipate these will provide more new insight than the model obtained here for the very simple WW domain.

■ ASSOCIATED CONTENT

Supporting Information

Methodological details on microstate assignment, algorithm implementation, and MSM construction and analysis. Additional supporting figures and tables show implied time scales for each featurization method, the full set of folding transitions depicted according to Figure 3, the log-likelihood as a function of macrostate number for the MSM constructed from folding-transition data only, structural representatives of each macrostate, and additional flux diagrams. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone (206) 543-1295. Fax (206) 685-1792. E-mail: dabaker@u.washington.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Greg Bowman and T. J. Lane for many helpful conversations and help with constructing MSMs, and we thank T. J. Lane and Vijay Pande for generously sharing data prior to publication. We thank David Shaw and his group for making their very long time scale simulations available. This work was supported by DFG grant LA 1817/3-1 (to O.F.L.),

■ REFERENCES

(1) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(2) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.
(3) Chodera, J. D.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2011** , *108*, 12969−12970.
(4) Rao, F.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9152−9157.
(5) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. *J. Chem. Phys.* **2011**, *134*, 135103.
(6) Garcia, A. E. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.
(7) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins* **1993**, *17*, 412−425.
(8) Beck, D. A.; Daggett, V. *Biophys. J.* **2007**, *93*, 3382−3391.
(9) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088−1093.
(10) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732−6737.
(11) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 586−591.
(12) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885−9890.
(13) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331−339.
(14) Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246−4257.
(15) Prentiss, M. C.; Wales, D. J.; Wolynes, P. G. *J. Chem. Phys.* **2008**, *128*, 225106.
(16) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248−251.
(17) Weinkam, P.; Romesberg, F. E.; Wolynes, P. G. *Biochemistry* **2009**, *48*, 2394−2402.
(18) Stamati, H.; Clementi, C.; Kavraki, L. E. *Proteins: Struct. Funct. Bioinf.* **2010**, *78*, 223−235.
(19) Ferguson A., P. A.; Kevrekidis, I.; Debenedetti, P. *Chem. Phys. Lett.* **2011**, *509*, 1−11.
(20) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.
(21) Huang, X.; Yao, Y.; Bowman, G. R.; Sun, J.; Guibas, L. J.; Carlsson, G.; Pande, V. S. *Pac. Symp. Biocomput.* **2010**, 228−239.
(22) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(23) Bowman, G. R.; Huang, X.; Pande, V. S. *Cell Res.* **2010**, *20*, 622−630.
(24) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890−10895.
(25) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 664−667.
(26) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.
(27) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.
(28) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(29) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.
(30) Zhuang, W.; Cui, R. Z.; Silva, D. A.; Huang, X. *J. Phys. Chem. B* **2011**, *115*, 5415−5424.
(31) Buch, I.; Giorgino, T.; De Fabritiis, G. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 10184−10189.
(32) Noe, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.
(33) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods* **2010**, *52*, 99−105.
(34) Bacallado, S.; Chodera, J. D.; Pande, V. *J. Chem. Phys.* **2009**, *131*, 045106.
(35) Kortemme, T.; Morozov, A. V.; Baker, D. *J. Mol. Biol.* **2003**, *326*, 1239−1259.
(36) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Methods Enzymol.* **2004**, *383*, 66−93.
(37) Blum, B.; Jordan, M. I.; Baker, D. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1583−1593.
(38) Lange, O. F.; Baker, D. *Proteins* **2012**, *80*, 884−895.
(39) Gonzalez, T. F. *Theor. Comput. Sci.* **1985**, *38*, 293−306.
(40) MacQueen, J. *Proc.5th Berkeley Symp. Math. Statistics Probability* **1967**, *1*, 281−297.
(41) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197−201.
(42) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(43) Prinz, J. H.; Keller, B.; Noe, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912−16927.
(44) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(45) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory. Comput.* **2011**, *7*, 3412−3419.
(46) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415−425.

(47) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(48) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.
(49) Jager, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M. *J. Mol. Biol.* **2001**, *311*, 373−393.
(50) Deechongkit, S.; Nguyen, H.; Powers, E. T.; Dawson, P. E.; Gruebele, M.; Kelly, J. W. *Nature* **2004**, *430*, 101−105.