Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece

Guha Jayachandran
Department of Computer Science, Stanford University, Stanford, California 94305

V. Vishal and Vijay S. Pande[a)]
Department of Chemistry, Stanford University, Stanford, California 94305

(Received 8 August 2005; accepted 20 February 2006; published online 24 April 2006)

We report on the use of large-scale distributed computing simulation and novel analysis techniques for examining the dynamics of a small protein. Matters addressed include folding rate, very long time scale kinetics, ensemble properties, and interaction with water. The target system for the study, the villin headpiece, has been of great interest to experimentalists and theorists both. Sampling totaled nearly 500 μs—the most extensive published to date for a system of villin's size in explicit solvent with all atom detail—and was in the form of tens of thousands of independent molecular dynamics trajectories, each several tens of nanoseconds in length. We report on kinetics sensitivity analyses that, using a set of short simulations, probed the role of water in villin's folding and sensitivity to the simulation's electrostatics treatment. By constructing Markovian state models (MSMs) from the collected data, we were able to propagate dynamics to times far beyond those directly simulated and to rapidly compute mean first passage times, long time kinetics (tens of microseconds), and evolution of ensemble property distributions over long times, otherwise currently impossible. We also tested our MSM by using it to predict the structure of villin *de novo*. © *2006 American Institute of Physics*. [DOI: 10.1063/1.2186317]

I. INTRODUCTION

Duan and Kollman's 1998 1 μs simulation of the villin headpiece heralded the potential of all-atom molecular dynamics to provide a high-resolution trajectory of a small protein.[1] Even today, with seven intervening years of Moore's law, such explicit solvent simulations remain a major challenge due to the required computational power.[2] Four years after the Kollman simulation, massively parallel simulations of villin were conducted using implicit solvent and yielded folding trajectories and an accurate prediction of the folding rate.[3] In the current work, we have utilized a distributed computing paradigm to realize tens of thousands of trajectories of villin in explicit solvent, totaling to a sampling of nearly 500 μs. This brings together the detailed model of Kollman's simulation with the statistical strength previously obtained only with implicit solvent.

One can use an ensemble of simulations to probe kinetic properties without directly simulating long folding pathways. For example, the calculation of the P_{fold} value of a conformation pools information from multiple, short simulations to yield the probability of that conformation folding before unfolding. In addition to testing putative transition state structures, P_{fold} analysis can provide an ordering of conformations along the folding pathway.[4] Each trajectory in a P_{fold} calculation can be relatively short, but many independent trajectories are needed. The analysis is thus naturally suited to distributed computing. Here, we demonstrate the use of similar, but more general, probabilities to probe the sensitivity of a system's kinetics to given perturbations.

While an ensemble of trajectories is vital for reliable statistics and for techniques such as P_{fold} analysis, it is also desirable to have predictions of dynamics occurring on the full folding time scale, rather than on the tens of nanosecond time scale that can typically be directly simulated with a detailed, explicit solvent model. This is not a challenge that can be addressed solely through running trajectories to orders of magnitude greater length unless we are able and willing to wait correspondingly longer times (or wait for innovation that makes computers significantly faster). As a single 10 μs trajectory of villin in explicit water would take years to achieve today, alternative approaches are clearly needed.

Recently, Markovian state models (MSMs) have been built for small peptides.[5–7] MSMs are defined by the transition probabilities between multiple states rather than just the two used in P_{fold} values. They hold the promise of being able to describe the transitioning of a system between states defined by conformational clusters,[8] with only a simple matrix multiplication operation required to propagate the system by a time step. A MSM approach could thus offer a solution to the conundrum of how to describe long time scale phenomena with simulation without waiting for the completion of very long trajectories.[9] Indeed, MSMs can be built from a set of relatively short trajectories and are therefore ideally suited to a cluster, grid, or distributed computing paradigm. While having a large number of trajectories yields benefits in and of itself (such as the ability to directly simulate folding and calculate rates), constructing MSMs lets one pool the infor-

[a)]Electronic mail: pande@stanford.edu

mation of multiple independent trajectories together in a probabilistically more descriptive picture of the dynamics. With grid computing becoming more and more popular in both academia and industry, we expect that this will be a powerful technique and we demonstrate here several possible applications.

The system used in this study, the 36-residue, alpha-helical villin headpiece, is considered a model system for both experiment and computation and has yielded significant results over the past several years.[1,3,10–12] It thus gives us a basis for validation of our simulations and is a natural system on which to demonstrate the applications of MSM methodology.

The main goals achieved in this work are as follows. First, we report on directly simulating folding trajectories of villin in explicit solvent starting from an unfolded conformation, a milestone in simulation. Second, we demonstrate a method of quantitatively assessing the sensitivity of dynamics to given system perturbations. Finally, we use the collected data to demonstrate the application of a MSM to a protein, presenting novel analyses reaching time scales not previously accessible and predicting quantities never before examined by atomistic simulation. Applications illustrated include computation of mean first passage times, evolution of property distributions over long time, and structure prediction. The techniques presented use villin for illustration but should be generally applicable to a broad range of problems involving long time scale dynamics.

II. METHODS

A. Molecular dynamics methods

We generated tens of thousands of independent trajectories for the 36-residue villin headpiece with molecular dynamics. The bulk of the simulations was performed on a subset of the nearly 200 000 processors around the world participating in our ongoing Folding@Home distributed computing project.[13] We adapted the GROMACS 3.1.4 molecular dynamics package to our distributed infrastructure.[4,14] Single precision computation was utilized, as was the case with previously published works with GROMACS and villin specifically,[14,15] and as with previous works with GROMACS more generally (free energy computation and protein folding kinetics).[4,16]

We largely followed the setup of Duan and Kollman as regards the temperature and pressure control algorithms, water model, box type, and time step.[1] As in that work, the villin sequence used was MLSDEDFKAVFGMTRSAFANLPLWKQQNLKKEKGLF protein data bank (PDB code 1VII), with N-acetyl and C-amino caps. The protein was solvated for all simulations in 5600–6000 explicit TIP3P water molecules in a truncated octahedron box, with periodic boundary conditions.[17] The minimum distance between a protein atom and the nearest image atom was 1 nm. Three sodium and five chloride ions were included to counter the protein's charge (~30 and 50 mM).

Simulations used a 2 fs integration step and 20 fs neighbor list update frequency, at 300 K temperature.[14] Berendsen temperature and pressure control were used, with coupling time constants of 0.1 and 1 ps, respectively.[18] Despite its unphysical scaling of kinetic energies, simulations with Berendsen control have previously yielded accurate folding rates on the microsecond scale.[4] The linear constraint solver (LINCS) algorithm[19] was used to constrain all bonds, rather than only bonds involving hydrogen atoms as in the Kollman simulation. The Garcia-Sanbonmatsu modified version of AMBER94 ("AMBERGS") served as the force field.[20,21] One to four van der Waals forces were scaled by 0.5 as in the base AMBER94; the data presented in Ref. 20 were with no such scaling.[22]

Two sets of calculations were run, each with a different treatment for long range electrostatics–particle mesh Ewald[23] (PME) or reaction field (RF).[24] Under RF, the Coulombic and van der Waals (vdW) neighbor lists went up to 10 Å with vdW interactions smoothed from 8 Å and an external dielectric of 80 was used. Under PME, the neighbor lists went up to 8 Å with vdW interactions smoothed from 6 Å. The grid spacing for Fourier transforms was 1.2 Å, the alpha parameter was 0.39 $Å^{-1}$, and the interpolation order was 4. The results presented below are from the more thoroughly sampled RF ensembles unless otherwise noted.

B. Characterization of the folded state

The simulated native state ensemble serves to validate native state stability in the simulation model and to characterize the native state and its intrinsic fluctuations. We ran 100 trajectories using each of the electrostatics treatments (with 2.8 μs total sampling using PME and 4.0 μs total sampling using RF). Due to the nature of the distributed computing environment, trajectories varied in length. In the PME set, the median trajectory length was 26 ns with 64 trajectories reaching 25 ns. In the RF set, the median trajectory length was 29.5 ns with 95 trajectories reaching 25 ns. A structure with 2.0 Å C_α root-mean-square deviation (RMSD) from the PDB structure 1VII, which resulted from brief equilibration of the PDB structure, was used to start all native state simulations.

It has been demonstrated that an averaging of folded structures can be more nativelike than any individual constituent of that ensemble.[25] Experiments such as NMR or x-ray crystallography make ensemble measurements. To make analogy using our single molecule trajectories, we computed the C_α–C_α distance matrix for each individual structure in the native simulation ensemble and computed the average across these at each time point.[25] The distance root mean square deviations (dRMSDs) of these "mean structures" with 1VII are shown in Fig. 1.[3] Figure 2 shows the dRMSD from 1VII seen among individual structures in the ensembles.[3] The C-terminal helix is key to experimental measurements such as tryptophan fluorescence,[26] and 91% and 95% of PME and RF conformations, respectively, have residues 24–33 helical.

Characterizing the native ensemble with global tertiary and per helix structural criteria suggested by the ensemble distributions (each of the three helices under 0.8 Å C_α-dRMSD and global C_α-dRMSD under 3.7 Å, all relative to the native 1VII structure), we can identify trajectories

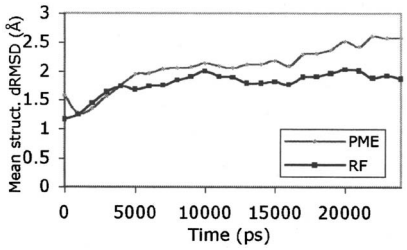

FIG. 1. Mean structure dRMSD vs time for PME and RF ensembles. The native ensembles remain close to the experimental structure over the time scale simulated. In particular, the mean structure under PME stays under 3 Å dRMSD through 25 ns and the mean structure under RF stays under 2 Å.

from the unfolded ensemble that reach the native state.[3] Though intended to be comprehensive, the criteria should not be considered a precise description of the model's native basin but a reasonable approximation; alternatives are discussed later.

C. Trajectories started unfolded

To obtain an unfolded starting structure for folding trajectories, we simulated a fully extended structure in the generalized Born with surface area (GB/SA) implicit solvent model in Tinker[27] at 1000 K and low viscosity (1 ps^{-1}) for 1 ns, to allow it to become more representative of the unfolded state and more amenable to solvation in explicit solvent. We began 10 000 independent simulations from that unfolded conformation. The trajectories were mostly 25 ns in length, with a median length of 25 ns, a mean of 25±2.7 ns, and a maximum of 50 ns. Though the lengths of individual trajectories were well under the experimentally observed microsecond-scale folding time, by probabilistic arguments we would expect the large ensemble of simulations to include a small number of folding trajectories.[3] We note that the more cumulative time simulated, the better the statistical measures. We also note that this distributed, parallel technique reduces the wall clock time rather than the cumulative computational time.

One might compare Monte Carlo simulation, with its strength in thermodynamics calculation, to a massive ensemble of molecular dynamics (MD) trajectories. Indeed, Monte Carlo simulation is an excellent technique for obtaining ensemble averages for thermodynamic quantities.[28] However, the kinetic interpretation of Monte Carlo simulations is unclear. Furthermore, molecular dynamics naturally exposes physical pathways, not requiring definition of a move set.

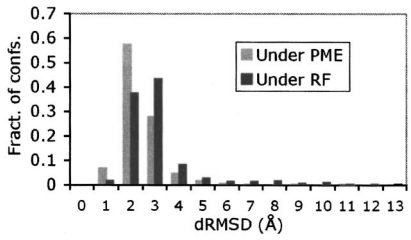

FIG. 2. Distribution of dRMSD (rounded to nearest angstroms) for the PME and RF native ensembles sample sampled within 15–25 ns. It shows that the native state is maintained on the simulated time scale and suggests the definition of the folded state described in the text.

D. MLE rate computation

A folding rate can be computed using the ensemble of simulations started unfolded.[29] Comparing this rate to experimental measurements can help us validate the simulation and sampling. When there is a single rate-limiting step, folding kinetics will be in the single exponential regime. We stress that this is a simplification—clearly there will be other, non-rate-limiting steps that would contribute to nonsingle exponential dynamics. Following Zagrovic and Pande,[29] we calculated the maximum likelihood estimator (MLE) for the rate assuming a single rate-limiting step (derivation presented below for completeness). Deviations from this single exponential regime can be addressed with MSM methods, as discussed in later sections.

Under a single exponential kinetics regime (single rate-limiting step), we can approximate the probability that a given trajectory becomes folded in the time between t and $t+dt$ by $P(t)dt=k\exp(-kt)dt$, where k is the rate constant. This is an approximation as there will be some small minimum lag time (t_{lag}) for our unfolded structure to reach the folded state; we approximate t_{lag} by the minimum folding time observed in the simulations.[4]

The definition of "folded" establishes which trajectories become folded (call them members of F) and at what times, and which trajectories do not fold (members of R) over their simulated lengths. Then, the probability of the observed data is

$$\prod_{i\in F} P(t_{f_i}) \prod_{i\in R} 1 - P(t_{e_i}), \tag{1}$$

where t_f is the time beyond t_{lag} at which a trajectory first meets the folded criteria and t_e is the total time beyond t_{lag} simulated for a trajectory (trajectories with length less than t_{lag} are neglected).[30] The value of k that maximizes this probability can be shown to be

$$|F|\left[\sum_{i\in F} t_f + \sum_{i\in R} t_{e_i}\right]^{-1} \tag{2}$$

($|F|$ denotes the number of members in set F). This method of calculating the rate can make more full use of collected simulation data than other methods used in the literature when trajectory lengths vary.[30] The Cramer-Rao lower bound on the variance of k is

$$\sqrt{|F|}\left[\sum_{i\in F} t_{f_i} + \sum_{i\in R} t_{e_i}\right]^{-1} \tag{3}$$

and the error in $\tau=1/k$ is the propagation of the k error.

E. Generalization of P_{fold} analysis: Calculation of $P_{X,Y}$ values

The P_{fold} value of a conformation is defined as the probability of its transitioning to the folded state before the unfolded state. Applications have included the ordering of conformations along the folding pathway and the identification

of the transition state.[4] In the present work, our interest is not especially in absolute P_{fold} values, but in probing sensitivity to certain system perturbations. We therefore propose computing more general commitment probabilities, which we denote $P_{X,Y}$,

$$P_{X,Y} = \frac{N(X)}{N(X) + N(Y)}, \tag{4}$$

where $N(X)$ is the number of trajectories that meet condition X before condition Y and $N(Y)$ is the number of trajectories that meet Y before X.

If a given perturbation has no impact on kinetics, then a necessary condition is that for any given XY, $P_{X,Y}(s) = P_{X,Y}(s')$ where s' is the perturbed form of system state s. A sufficient condition for a perturbation not affecting the system is that $P_{X,Y}(s) = P_{X,Y}(s')$ for all XY. One obviously cannot test all possible XY combinations but we do probe with multiple XY conditions based on protein conformation. Any XY yielding a difference in commitment probabilities under perturbed and unperturbed conditions would indicate the given perturbation affecting the protein's dynamics. To reduce statistical noise in the probabilities, X and Y should be sufficiently disjoint that there is some barrier between them.

To obtain such specific probabilities with high precision, we conducted simulations expressly for that purpose. We chose a variety of conformations from trajectories in the ensemble started unfolded and ran 100 trajectories from each of those points under both perturbed and unperturbed conditions, seeing what fraction of each set met the condition X before Y. The specific perturbations and XY tested are discussed in Sec. III.

F. MSM construction

In constructing a MSM, we largely followed the basic approach described by Singhal *et al.*[6] However, unlike in that work, where only data from trajectories started unfolded were utilized, here the data from the ensemble started unfolded, the native state ensemble, and the $P_{X,Y}$ simulation ensembles were all used. Especially as the $P_{X,Y}$ simulation ensembles include trajectories starting between the folded and unfolded states (the region containing the transition state), using all three sets of ensembles maximized converge of conformational space.

As the first step in construction of the MSM, all 4 509 355 conformations (taken at 100 ps intervals from each trajectory) were clustered into 2455 states. The k-means clustering algorithm was used, parallelized over a cluster of computers using a message passing interface (MPI) library.[31] The distance metric was dRMSD. After clustering, folded conformations were present in more than one cluster, as more than one cluster overlapped the native state. Such conformations were removed from their original clusters and segregated into a single new state s_f. Two states had no path connecting them to s_f (by the edges described in the next paragraph) and were removed so as to avoid having unfolded sinks, leaving a total of 2454 states.[6]

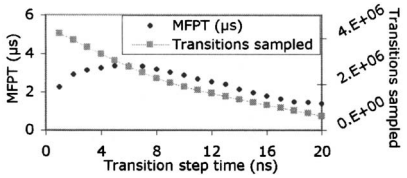

FIG. 3. Transition time of a MSM vs the MFPT from the extended to the folded state for that MSM. A condition for properly capturing the process underlying the data is that the MFPT levels as the transition time step is increased beyond some point (rather than grow linearly with time step as would be the case if the transition matrices did not reflect the increased time step). For the clustering described in the text, we see that the computed MFPT (solid) (uncorrected for viscosity) is near 3 μs over a range of transition time steps. The decline in MFPT at longer times is believed due to poor sampling there—sampling (dashed) decreases as transition time step increases, due to the limited lengths of trajectories.

Transition probabilities between states were computed based on transitions seen in the simulations. Given a transition time Δt, the probability of transitioning from a given state s_a to a given state s_b was given by the number of times, $T(s_a, s_b)$, a simulated trajectory was observed to be in state s_b time Δt after being in state s_a divided by the total number of Δt transitions from state s_a observed (including self-transitions),

$$P(s_a, s_b) = \frac{T(s_a, s_b)}{\Sigma_{\forall i} T(s_a, s_i)}. \tag{5}$$

The results below use a transition time step of 10 ns.

Validation of MSMs to check that the model accurately describes the process underlying the raw data is an active area of theoretical research.[5,32] In particular, methods to verify that a process is Markovian on a given time scale—that future moves depend only on the system's current state and not previous states—are being developed. In this work, an agreement between the model and direct examination of the simulation data for certain properties was considered a necessary but insufficient condition (as in Sec. III C). Another such condition, related to the steadiness of the mean first passage time (MFPT) to the folded state with regard to the transition time step, is shown in Fig. 3 (The MFPT from a state A to a state B is the expected time to first visit B starting from A—its computation is described in Ref. 6.) Further work in model validation, particularly with methods robust to the complex processes involved with real proteins, is ongoing and is expected to give it more theoretically complete grounding.

Given the large number of probabilities involved in a MSM, their statistical robustness is also a concern, even with the large amount of sampling in the present work. A bootstrap analysis was conducted. For each state of the above MSM, observed molecular dynamics transitions were randomly chosen (with replacement) from the set of all observed transitions from that state. As is often the convention, the number of transitions selected was equal to the total number of observed transitions from the state. The selected transitions were used to compute new transition probabilities from that state to the other states. Repeating this for each state resulted in the construction of a transition matrix. The procedure was repeated to construct 50 such matrices. The

transition with the largest standard deviation over the associated MSMs had a standard deviation of 0.059 in its probability. Over 99% of state-to-state transitions that were nonzero in at least one MSM had a standard deviation of under 0.015. Even numerically small error can be significant in impact. Here, considering the MFPT from the state including fully extended conformations to the folded state, the mean value over the models was 2.9 μs, with a standard deviation of 0.18 μs. This supports the value of 2.9 μs obtained from the MSM constructed with all the data. The MFPT is further discussed in the Results section.

A simple cross-validation test was also performed on the MFPT. We randomly divided the observed molecular dynamics transitions into ten disjoint sets (each of cardinality 188 836 or 188 837). Each of those sets was used to calculate transition probabilities for a MSM following the procedure described earlier. The average value over the models for the MFPT between the same two states as described in the previous paragraph was 2.9±0.24 μs, in agreement with the earlier values.

III. RESULTS

In Sec. III A below, we briefly discuss the observed folding trajectories and present rates computed using the MLE approach described earlier in Sec. II D. In the subsequent sections, the central feature of the analyses discussed is the use of transition probabilities. The ability to compute transition probabilities between conformational states is one of the key benefits of obtaining an ensemble of trajectories. In Sec. III C, we discuss how the computation of transition probabilities from single conformations ($P_{X,Y}$ values) may be used to probe the sensitivity of kinetics to given features of the system. We build on the method and obtained results in the final section, generalizing to the transition probabilities between conformational clusters and introducing several applications of a MSM.

A. Folding trajectories and rates

Below, we will present the folding rate computed from the ensemble of simulations started unfolded and compare it to experiments, but first, it is worthwhile to briefly consider trajectories from that ensemble individually and compare to previous computational studies of villin. The behavior of the residues 9–32, for example, has been typically examined in computational studies of villin (where numbering begins at 1 for the protein simulated, or at residue 41 of the NMR structure).[1,12] A rationale for examining this subset of the residues can be seen from the high variance (Fig. 4) in atom positions outside of this region even in the 29 NMR structures that were averaged to obtain 1VII.[33] The lowest C_α-RMSD for residues 9–32 ($RMSD_{9-32}$) seen under PME in the trajectories started unfolded is 1.50 Å and the lowest seen under RF is 1.29 Å. With the improved sampling, 89 PME trajectories and 135 RF trajectories see $RMSD_{9-32}$ values lower than the lowest 9–32 main chain RMSD, 3.0 Å, seen in the Kollman trajectory.[1] A similar number of our simulations reach lower values than the lowest 9–32 backbone RMSD reached in the Shen and Freed implicit solvent

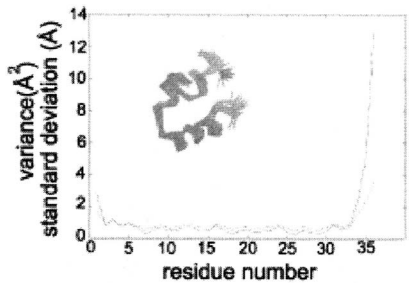

FIG. 4. Picture of all 29 McKnight structures overlaid and plot of the unbiased variance (dashed) and standard deviation (solid) in each C_α position based on the 29 structures. Plot values were obtained by first doing a C_α best fit of each of the 29 McKnight NMRs to 1VII and, for each of these, computing the Cartesian distance between each C_α and the corresponding C_α in 1VII. Both in the picture and graph, one can see that the ends—and especially the region after residue 32—are far less rigid than the 9–32 regions.

trajectory.[1,12] Also, 39 of our trajectories meet the dRMSD <2.85 Å folded criteria used in the analysis of an ensemble of implicit solvent simulations by Zagrovic *et al.*[3] The GB/SA solvent model used in that work led to excessive initial collapse, one of the motivations for examining explicit solvent. The lowest dRMSD seen in the current work is 1.74 Å. Figure 5 shows the time course of two trajectories reaching low RMSD.

Besides looking at folding trajectories individually, we also use the entire ensemble to compute the folding rate and compare it to experimental measurements. The folding rate of the villin headpiece has been experimentally measured by two methods: Kubelka *et al.* measured the folding rate of the villin headpiece mutant N28H to be 4.3±0.6 μs using laser temperature jump and Wang *et al.* obtained a 10 μs time scale using NMR line-shape analysis.[10,11] Applying the described MLE method to our villin ensemble (under RF), we obtain a raw value of 7.9±2.3 μs.

These results should be considered in the light of the nature of diffusion in the water model simulated. It is well known that the TIP3P water model[29] has an unphysically low viscosity.[34] To obtain a rough measure of this fact's kinetic impact, an equilibrated cubic water box with a 3 nm side was simulated under the same simulation model as described earlier. Sampling configurations at a 10 fs frequency and averaging over all molecules yielded a diffusion constant of 6.7×10^{-5} cm^2/s. The experimental diffusion constant of water at 298.2 K and 1.013 atm has been measured to be 2.23×10^{-5} cm^2/s. If we assume that the folding rate is linearly related with water's diffusion constant, this yields a correction factor of 0.33, or a folding rate of 23.7±6.9 μs. We emphasize that this is by no means a precise correction. Diffusion constants are highly dependent on factors such as density, which is itself conditioned on the presence or absence of protein, and even the conformation of the protein. Furthermore, it has not been established that the solvent's diffusion constant is linearly related with a protein's folding rate. With these issues in mind, the calculated rate is well within an order of magnitude of the experimental measurements.

Due to the dependence of the calculated rate on the specific criteria used to define the folded state, it is constructive

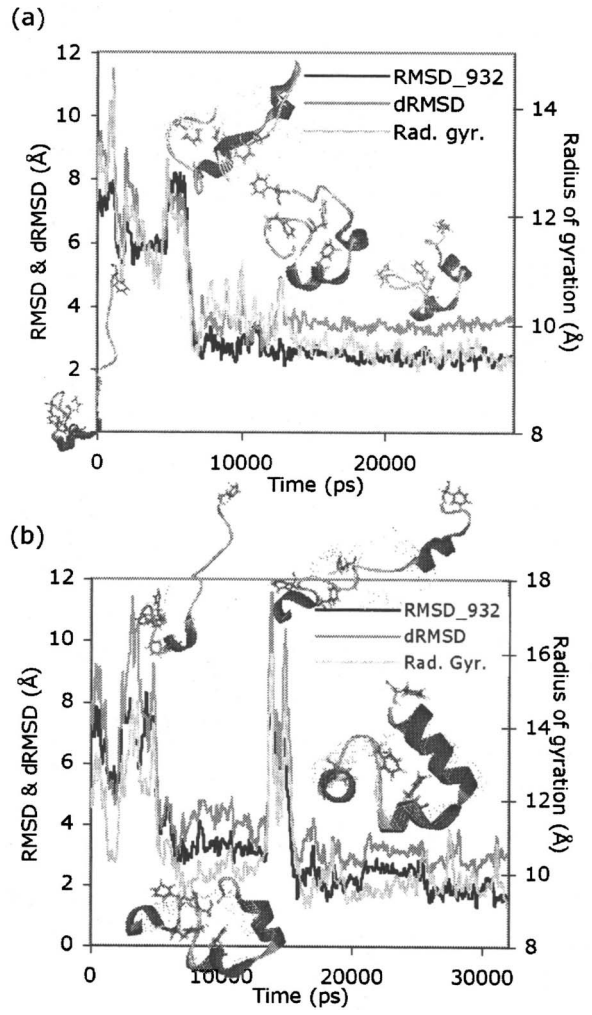

FIG. 5. The time course of two trajectories reaching lowing RMSD relative to the native structure. The dRMSD with 1VII, $RMSD_{9-32}$ with the 1VII, and radius of gyration (rad. gyr.) are shown.

to consider the variation of the predicted rate on the definition of the folded state chosen. In Table I, we examine several different criteria for the folded state, including typical criteria considered in computational studies and rough analogs to the experimental probes employed by Kubelka *et al.*[10] One would not expect an exact equivalence with experiments given imperfection in the force field and, more notably, the fact that the folded condition has to be characterized differently in simulation versus experiment, but all criteria considered here yield rates well within an order of magnitude of the experimental measurements (Table I). One should note that the systematic error in rate computation arising from the choice of definition is as great as the statistical error; folded definitions of the sort used here are but approximations to the folded state, recommending sensitivity analysis such as this.

While the maximum likelihood method used here and the fitting methods used in previous works[3,4] have a strong mathematical foundation when there is a basis for assuming single exponential kinetics, they cannot be extended to cases where the kinetic model is completely unknown. This is among the many issues that a MSM can address.

B. Kinetics sensitivity analyses

Before discussing applications of a MSM, which describes transitions between states defined by conformational clusters, we show how transition probabilities from individual conformations can be used to probe the sensitivity of dynamics. The results can help in constructing and understanding MSMs, particularly in definition of states and time steps. First, we show how we can probe whether protein dynamics are highly sensitive to water configuration through $P_{X,Y}$ analysis and discuss how the result aids MSM methodology. Then, we demonstrate how we can assess whether kinetics are sensitive to a given change in the simulation model. This technique may be useful in testing physical faithfulness of MSMs, and of molecular simulation in general. Finally, we discuss an extension of $P_{X,Y}$ analysis and present more on its relationship to MSMs.

1. Sensitivity to water configuration

A set of simulations, computing $P_{X,Y}$ values for protein structures in different water configurations, was undertaken to analyze the role of water in the folding process. For each starting protein conformation chosen, we ran 100 simulations of the conformation in each of two water configurations. The perturbed water configuration was generated by freezing the protein in the unperturbed form of the system and equilibrating the water for 500 ps.

In simulations of the 23-residue BBA5, water appeared not to play a significant structural role.[4] Here, we observe similar behavior. The best fit line on the plots of $P_{X,Y}$ values with and without reannealing of water is the identity line (Fig. 6). Formally, the commitment to X or Y of each trajec-

TABLE I. Sensitivity of computed rates (using MLE method, without viscosity correction) to the condition tested. Each row is for different folded criterion. The range of values highlights the systematic error from the choice of condition.

Condition	Notes	Trajs.	Min. folding time (μs)	k (μs^{-1})	τ (μs)
dRMSD $<$ 3.7 Å and local dRMSD of each helix $<$0.8 Å	Tertiary and secondary structures	12	16	0.13±0.036	7.9±2.3
$RMSD_{9-32}<3$ Å & dRMSD $<$ 4 Å and three helices	Local tertiary, global tetiary, and secondary structures	48	6.5	0.26±0.037	3.9±0.56
C-terminal helix (residus24–33 helical as judged by dssp)	To approximate experimental probe	282	3.7	13±0.077	0.78±0.046
C-terminal helix maintained continuously 1 ns	As above, with stability requirement	240	3.7	1.1±0.071	0.91±0.060
dRMSD $<$ 2.85 Å	Used in Ref. 3	32	10.4	0.022±0.038	4.6±0.82
dRMSD $<$ 3 Å	Looser than above, highlights sensitivity of condition	74	3.5	0.34±0.039	2.9±0.34

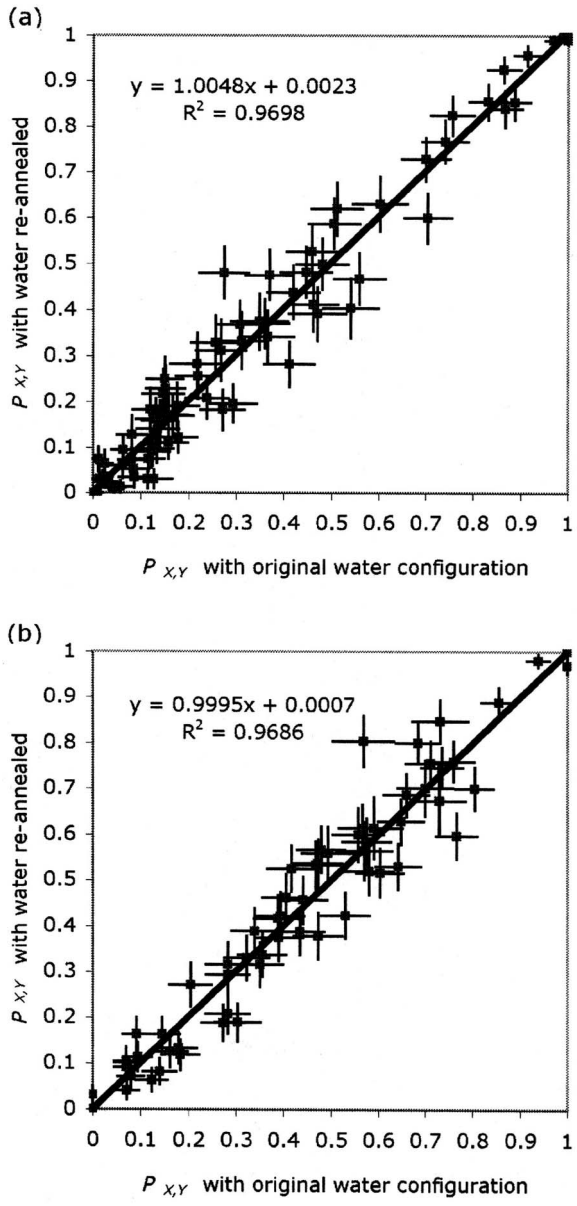

FIG. 6. Plots of $P_{X,Y}$ values for given structures with their original water configuration (meaning the configuration seen at the point the structure was sampled in the ensemble simulations) and with a reannealed water configuration. The best fit is the identity line. (a) X= "<20% native contacts formed." Y= ">40% native contacts formed." (b) X= "radius of gyration <10 Å." Y= "radius of gyration >13 Å."

tory is a Bernoulli trial and the estimated mean of the associated binomial distribution of many Bernoulli trials (after normalization) is $P_{X,Y}$. Assuming that if we repeated the entire set of Bernoulli trials many times the means observed would follow a normal distribution centered at the computed $P_{X,Y}$, the error bars in Fig. 6 about each $P_{X,Y}(s)$ data point are defined by the standard deviation of that normal distribution,

$$\sigma_{X,Y}(s) = \sqrt{\frac{P_{X,Y}(s)(1 - P_{X,Y}(s))}{N}} \tag{6}$$

[N is the total number of trajectories used to compute $P_{X,Y}(s)$]. Table II shows the RMSD values between $P_{X,Y}$ sets for different water configurations for several more XY conditions. These values are all under 0.06. In addition to the absolute RMSD, the table also shows the mean number of standard deviations comprising the absolute difference between two corresponding probabilities [using the fact that if $P_{X,Y}(s)$ and $P_{X,Y}(s')$ are independent, the variance in their difference is the sum of their respective variances],

$$\frac{1}{N}\sum_{s} \frac{|P_{X,Y}(s) - P_{X,Y}(s')|}{\sqrt{\sigma^2_{X,Y}(s) + \sigma^2_{X,Y}(s')}} \tag{7}$$

[N is the number of conformations, s is a conformation in one water configuration, s' is that conformation in a different water configuration, and $\sigma_{X,Y}$ is as defined in Eq. (6)]. None of these values exceeds 1.

These results agree with previous results on BBA5, which showed little impact on protein dynamics from perturbations to water configuration. The results indicate that, for villin, either the water configuration does not play a role in dynamics or water equilibration is so fast that water exists in a fairly equilibrated state around the protein. In the second case, the lack of sensitivity of villin's kinetics to the particular water configuration at a given time point is likely due to the water degrees of freedom annealing quickly relative to protein motion.

Besides its relevance to the role of water in folding, the result is also encouraging for the construction of MSMs. In the MSM that we will present, states are defined purely from the protein coordinates. If the water coordinates played a role in the above $P_{X,Y}$ values, that result would imply that water coordinates are important on the several nanosecond time scale and that a MSM omitting water from its state definitions could accurately represent the simulations only on a coarser time resolution than the nanosecond length of the $P_{X,Y}$ trajectories and/or a state partitioning coarse enough to lump conformations from X and Y into the same state. Both restrictions would limit the insights available from the model (the limiting case would be one state spanning the entire space or an infinite time step). Also, as the restriction on time resolution would require transitions be sampled over longer time steps, it would have the effect of requiring that longer trajectories be simulated. We reiterate that the $P_{X,Y}$ analysis shown here, though satisfying necessary conditions for omitting water coordinates from state definitions, is not sufficient to prove them unnecessary. In a later section, we describe a generalization of $P_{X,Y}$ and other tests that could be more probative.

2. Sensitivity to electrostatics treatment

$P_{X,Y}$ analysis can be used not just to probe water, but any factor in the system, including aspects of the simulation model. We illustrate its use for probing sensitivity to long range electrostatics treatment. If transition probabilities and other kinetic properties obtained from MD are highly dependent on a certain aspect of the model, then that aspect likely deserves further study. Just as with a single MD trajectory, the ability of a MSM to predict experiment is dependent on the validity of the simulation model. While an examination of all possible molecular dynamics methods is clearly beyond the scope of this paper, we present an example of such sensitivity analyses, for a case where there is still a variety of

TABLE II. The RMSD between a set of $P_{X,Y}$ values for a set of 88 conformations in water and a set of $P_{X,Y}$ values for those conformations in reequilibrated water configurations. The comparison is shown for various XY. The values do not show a major effect on the probabilities from the specific water configuration.

Condition X	Condition Y	RMSD	Eq. 7 (mean std. devs. in abs. difference)
dRMSD <3 Å	dRMSD >7 Å	0.03	0.8
Helix 1 dRMSD <1.6 Å	Helix 1 dRMSD >3.5 Å	0.04	0.9
Helix 1 dRMSD <1.8 Å	Helix 1 dRMSD >3.7 Å	0.05	1
Helix 3 dRMSD <1.6 Å	Helix 3 dRMSD >3.5 Å	0.03	0.8
Helix 3 dRMSD <1.8 Å	Helix 3 dRMSD >3.7 Å	0.02	0.9
Fract. native contact <0.2	Fract. native contacts >0.4	0.05	0.9
Fract. native contacts <0.3	Fract. native contacts >0.6	0.03	1
Rad. gyr. <10 Å	Rad. gyr. >13 Å	0.06	0.9
Rad. gyr. <10 Å	Rad. gyr. >15 Å	0.06	0.8
Rad. gyr. <12 Å	Rad. gyr. >17 Å	0.03	0.8
$\mathrm{RMSD}_{9-32}<3$ Å	$\mathrm{RMSD}_{9-32}>6$ Å	0.04	0.8

methods currently employed in the literature. In Sec. III C, we discuss how a MSM may be used to more efficiently test and develop new models.

As we conducted simulations both with PME and RF, we can examine the sensitivity to change between these long range electrostatics methods. In comparing the performance of PME and RF, speed and reproduction of experimental observations are the two main considerations. Algorithmically one would expect a force calculation with RF to be faster than with PME and in the GROMACS MD implementation, a force calculation with RF is close to 20% faster than one with PME. Simulations were run where one ensemble of trajectories with a given starting structure used PME and another ensemble with the same starting structure used RF. Plots, across 112 structures, of $P_{X,Y}$ for the structure under PME versus $P_{X,Y}$ for that structure under RF indicate differences (Fig. 7). The nonunity slopes indicate that there is sensitivity to the choice; the absolute position of a structure along a pathway depends on it. Absolute RMSDs for the conditions shown in Table II range up to 0.16 in this case and the mean number of standard deviations in the differences ranges to 2.7. More comprehensive tests of this type may help elucidate differences from different simulation methods or different simulated physical conditions. A theory to explain the significance of such shifts has been recently explored by Rhee and Pande.[35]

We stress that the analysis here does not deem one method of superior accuracy, since accuracy is to be judged with comparison to experiment and both methods reasonably reproduce rates and maintain native structure. Furthermore, the parameters used for each algorithm (such as cutoffs) can make a significant difference and only set was used for each here. What this analysis does offer is the ability to address whether the choice between PME and RF has any major impact on $P_{X,Y}$ values, in a parallelized manner that may be generally applied for other such methodological comparisons. This is a more stringent test than solely comparing folding rates, given the fact that two simulation methods may produce identical rates with different mechanisms.[36]

3. Extension and relation to MSMs

A MSM is similar in spirit to the $P_{X,Y}$ technique in its consideration of transitions between states: A $P_{X,Y}$ value gives the transition probabilities from one state to each of two other defined (large) states, while a MSM is defined by a transition matrix that specifies the transition probabilities between potentially many different conformational states. We point out that a comprehensive extension of the described $P_{X,Y}$ analysis method would be to define a large diversity of conformational states X_i and compute $\boldsymbol{P}_{\mathrm{neighbor}}$ vectors, where $\boldsymbol{P}_{\mathrm{neighbor}}(s)$ is the vector of probabilities such that element i of the vector is the probability that conformation s reaches state X_i before reaching any other defined state X_j. Comparing $\boldsymbol{P}_{\mathrm{neighbor}}$ vectors could be a stronger test than comparing $\boldsymbol{P}_{X,Y}$ values because, due to the smaller size of the states, it would be more sensitive to pathway differences. Unfortunately, smaller states also mean that more sampling is required to obtain statistically significant $\boldsymbol{P}_{\mathrm{neighbor}}$ vectors and its computation is beyond the scope of the current work. Sufficient sampling to measure $\boldsymbol{P}_{\mathrm{neighbor}}$ could also yield precise transition probabilities for a MSM, though how to choose states ahead of time is not clear, a problem avoided by the construction method described earlier.

In the opposite direction, the P_{fold} of a given state in a MSM can be computed analytically from the MSM.[6] However, with the clustering used in the construction of our MSM, the P_{fold} would not be associated with a single, well-defined conformation. Furthermore, probing the effect of varying system parameters would still require simulating batches of trajectories from a number of the states under the perturbed conditions.

C. Applications of a MSM

We now demonstrate several possible applications of the MSM (described in the Methods section) built for villin from all of the explicit solvent simulation data. These include an examination of long time scale folding, computation of ensemble property distributions, structure prediction, and model refinement. We conclude with a discussion of increasing the efficiency of MSM analysis.

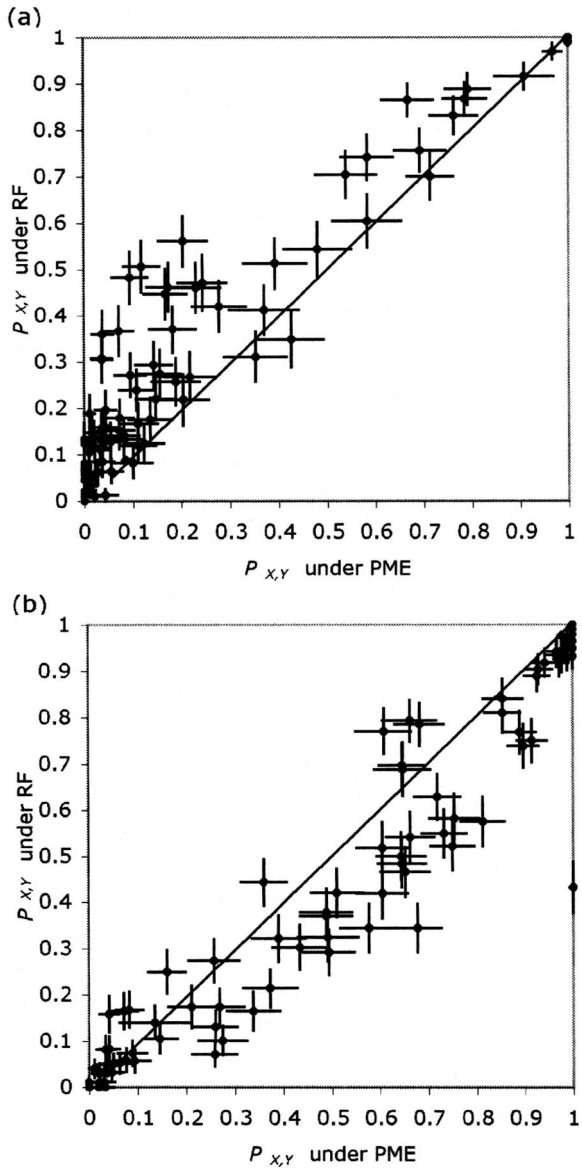

FIG. 7. Plots of $P_{X,Y}$ values for given structures when simulated under PME and RF. (a) X= "<20% native contacts formed." Y= ">40% native contacts formed." (b) X= "$RMSD_{9-32}<3$ Å." Y= "$RMSD_{9-32}>5$ Å."

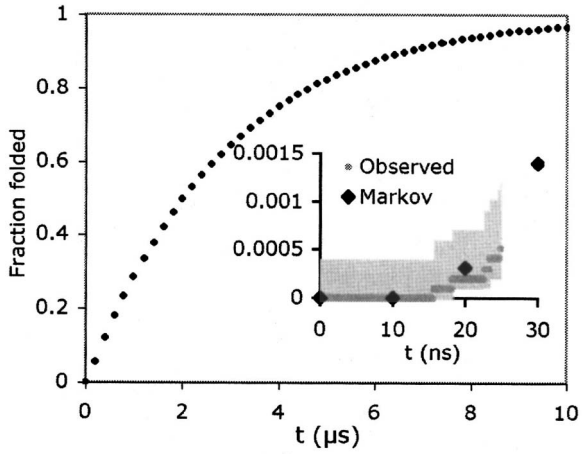

FIG. 8. Cumulative distribution of first folding times. The main plot shows the fraction computed by propagating state transition probabilities with the MSM. The single exponential, $1-\exp(-kt)$, with $k=0.34$ (the reciprocal of the computed MFPT), fits the curve ($R^2=1.0$). In the inset is a magnification of the same data through only 25 ns, along with the fractions obtained directly from the simulations. The deviation between the curves is within the statistical error in the observed fractions and thus deemed as meeting a necessary condition for the validity of the MSM. The error shown for each observed fraction is the 95% confidence interval assuming a beta distribution (Ref. 39).

1. Long time scale folding

In previous works,[3,4] we have reported the cumulative distribution of first folding times observed in the simulated ensemble. With a MSM, we can extend that distribution to longer times, as one can quickly and easily compute the evolution of the system over arbitrary time lengths. The MSM describes dynamics with a transition matrix $\mathbf{P}$, where element $\mathbf{P}(i,j)$ is the probability of transiting to state j in the next move given that one is currently in state i, and where the self-transition probability of the folded state is set to 1 so that we will only capture *first* folding times. Then, the fraction of trajectories in each state after n propagation steps will be in the row vector $\pi(n)=\pi(0)\ \mathbf{P}^n$, where $\pi(0)$ is a row vector with the starting fractional populations.

With this model, the fractional folded population at a given time can be computed. Figure 8 shows the fractions folded observed directly in the simulated unfolded ensemble, through 25 ns, and the computed fraction folded from the MSM to 10 μs. The MSM's first folding time cumulative distribution curve is single exponential, of the form $1-\exp(-kt)$, suggesting a single rate-determining barrier along the folding pathway (or a number of barriers of equal height). The curve agrees, over the simulated time, with the cumulative distribution of first folding times observed directly in the simulated ensemble. This agreement is a necessary condition for considering the MSM a valid representation of the simulated data.

The MFPT from a given state to the folded state can also be easily calculated from the transition matrix.[6] We obtain a value of 8.7 μs (including the approximate correction for the solvent's anomalous viscosity discussed earlier) from the state that includes the initial unfolded structure of our simulations (the average MFPT from all states with a mean dRMSD>7 Å is also equal to that). This is in reasonable agreement with both experiment and the maximum likelihood method described earlier. We note that the MFPT calculation requires no assumption about the overall kinetic model of the protein, making it ideal for more complicated systems. A formalism for computing the error in the MFPT yielded by a given MSM is being developed.[37]

2. Evolution of ensemble properties

As we can obtain the time evolution of per-state populations from the MSM, it is also possible to obtain the time evolution of ensemble distributions of various properties, such as RMSD or helical content. In addition to probing observable behavior over long time, these may be useful for comparison with experiments that yield ensemble averaged results and therefore may be both another tool in the testing of simulation methods and a tool for decomposing ensemble averages to distributions.

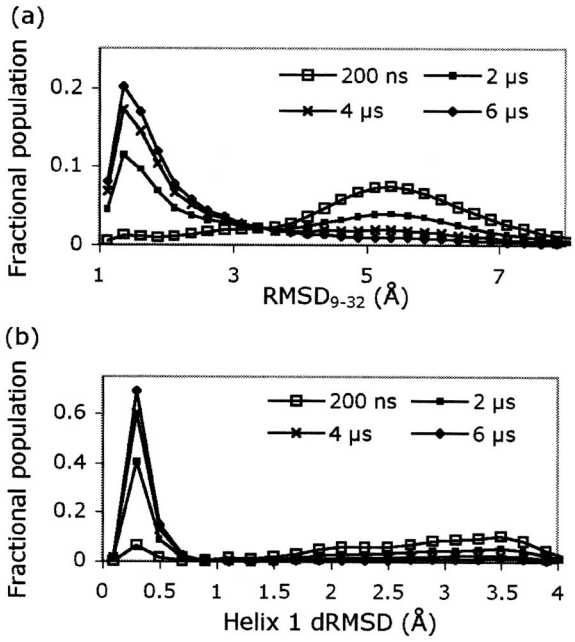

FIG. 9. Evolution of ensemble distributions of $RMSD_{9-32}$ and helix 1 dRMSD computed from the MSM.

For a given simulation observable of interest, **m**, we first characterize each state s_i of the MSM by the distribution of **m** among the conformations comprising s_i (call the associated normalized histogram $\mathbf{D}_{s_i}^m$). Note that if sampling of a state s_i is poor, $\mathbf{D}_{s_i}^m$ may not accurately represent the distribution of **m** expected for that state under a Boltzmann distribution. We later suggest methods, besides an increased simulation, for addressing this. The folded state s_f is treated as a sink in this demonstration to more clearly see the progression of states from unfolded to the folded state. Future applications of this analysis technique may omit this approximation without any other change in procedure. Given the population of each state at a number of time points t_j, determined as described earlier, we can now use the per-state histograms $\mathbf{D}_{s_i}^m$ to build an overall distribution of **m** for each time point [call the histograms $\mathbf{D}^{\mathbf{m}}(t_j)$]. We build a histogram $\mathbf{D}^{\mathbf{m}}(t_j)$ by taking the sum of the per-state histograms $\mathbf{D}_{s_i}^{\mathbf{m}}$ weighting each bin of $\mathbf{D}_{s_i}^{\mathbf{m}}$ by $\pi_{s_i}(t_j)$, the fractional population of s_i at t_j,

$$\mathbf{D}^{\mathbf{m}}(t_j) = \sum_i \mathbf{D}_{s_i}^{\mathbf{m}} \pi_{s_i}(t_j). \tag{8}$$

Figure 9 shows the outcome of applying the above method for villin. With such an analysis, we can assess mechanisms unconstrained by the actual simulated length of the trajectories (within the limit of there having been sufficient sampling to construct the model). This will be especially important for even more complex, slower evolving proteins.

3. Structure prediction

The stationary distribution of a MSM for a protein has a special significance. Given the transition matrix **P** defining a MSM, the stationary distribution of the MSM is defined by the vector **s** that satisfies the equation $\mathbf{P}=\mathbf{sP}$—it is the eigenvector of **P** associated with the eigenvalue 1. The vector **s** is therefore the equilibrium distribution of the process.

This suggests that MSMs may be useful for prediction of unknown protein structure. The stationary distribution indicates free energies, so perhaps conformations in free energy minima (native ensemble) can be identified from the MSM. A challenge to such an application, besides force field accuracy and sampling hurdles, is that states must be small enough to provide a precise description of the native state (something not achieved here), while being large enough to have adequate observations for accurate computation of transition probabilities. Furthermore, the clustering metric must be such that native conformations do not get divided amongst too many different neighboring clusters. With these issues in mind, the ability to predict the native structure tests the accuracy of a MSM.

Illustrating these points, a MSM was constructed and validated for villin as described previously, except without segregation of folded conformations and using only the trajectories started extended (trajectories from the $P_{X,Y}$ analysis and from the native state were not used so as to avoid any inclusion of native state knowledge). The MSM's stationary distribution's most probable state (1.3%, 16 times the mean probability of 0.08%) has a mean dRMSD of 3.8±0.2 Å and a mean $RMSD_{9-32}$ of 2.7±0.4 Å. This state does not only contain folded conformations, since there are conformations from within and without the native basin that are yet similar enough to be clustered together at the utilized clustering granularity. Also, it is not the only state including folded conformations, as they are divided amongst neighboring clusters, making the interpretation of the state probabilities in the context of native and nonnative difficult. The 181 states containing at least one conformation meeting the dRMSD aspect of our folded criteria have a total probability of 0.29. That the native state is not more favored may indicate an error in the force field as well as errors in the MSM especially due to insufficient sampling. Still, further investigation along these lines seems warranted given that the single most probable state, described above, is nativelike and suggests that practical insights may arise even short of a completely accurate equilibrium distribution. It may also be possible to use high probability states to guide further sampling, perhaps with a more detailed and more accurate model, such as a polarizable force field.

Finally, while molecular dynamics combined with MSM methods will generally not be the most efficient way to predict the folded structure, native structure is an important additional test of methodology. Structure prediction from molecular dynamics is difficult due to the sampling involved as well as possible issues in the force fields. A MSM offers an excellent tool for use in assessing and improving methodology. If a force field is modified, for example, simulations using that modified force field can be started from each of an existing MSM's states; by starting trajectories from throughout conformational space, new transition probabilities can be determined much more efficiently, and with shorter trajectories, than they were for the initial model.

4. Greater efficiency

Clearly, greater numbers of trajectories to greater lengths would yield better sampling and thus more accurate MSMs. However, even with the same amount of simulation, one may benefit from dividing the trajectories to start among a more diverse set of initial structures. This should result in coverage of portions of conformational space that would require more time to reach if all trajectories were started at a single point. No additional steps need be taken in computation of transition probabilities, since those are purely based on local transitions between states and are normalized by the number of transitions seen from a state. We also point out that regardless of the distribution of starting points for the trajectories used to construct the MSM, the model can be used to realize pathways starting from any given state or combination of states. The diversity of starting points would be particularly important for proteins with more complicated kinetics, such as those with intermediate states. For such proteins, the important regions to start simulations might be between each intermediate state, running the trajectories to just long enough length that the sampling of the states overlaps. This is motivated in a similar manner to transition path sampling methods.[38]

IV. CONCLUSION

The main constraint on computational studies of protein dynamics, and of folding in particular, has been the inaccessibility of long time scales and ensemble statistics. Here, we have demonstrated the power of massively parallel simulation and analysis tools such as MSMs to help overcome this barrier. Simulations of large ensembles, previously accomplished for villin only in implicit solvent and only for smaller systems in explicit solvent, allowed the computation of rates and examination of folding trajectories.

$P_{X,Y}$ analysis also made use of parallelized simulation data. Here, its use to probe the role of water and to compare different simulation methodologies was demonstrated. The observed insensitivity to water configuration would be good to test for larger proteins, and the methodology test would ideally be applied to a large range of common simulation methods. In the case of a discrepancy in the latter, the development of new experimental tests may be required to distinguish which, if any, is correct.

However, simulations can perhaps also now move towards what experiments can measure, as accurately built MSMs can propagate ensemble data over long times for each model. The use of MSMs built from the data permits extrapolation to time scales far beyond those directly simulated. Applications include consideration of long time scale kinetics, examination of the evolution of ensemble properties, and perhaps even free energy computation and structure prediction. They can be used to help refine simulation models to obtain a higher native state stability, for example.

We believe that much work remains to be done in developing techniques for analyzing the data obtained from massively parallel simulation. Further development of MSMs should, for example, include the introduction of better statistical measures of error and techniques for choosing clustering metrics. We believe that efficiency will greatly increase from the use of simulation data from even more starting conformations and that this will be crucial for larger systems with more complex, slower kinetics. This work shows, however, that techniques for examining the dynamics of a protein, in a detailed model on a long time scale, are today within our reach.

ACKNOWLEDGMENTS

We thank Folding@Home participants worldwide (http://folding.stanford.edu). This work was supported by a grant from NSF Molecular Biophysics.

[1] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[2] V. S. Pande, I. Baker, J. Chapman *et al.*, Biopolymers **68**, 91 (2003).
[3] B. Zagrovic, C. D. Snow, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 927 (2002).
[4] Y. M. Rhee, E. J. Sorin, G. Jayachandran, E. Lindahl, and V. S. Pande, Proc. Natl. Acad. Sci. U.S.A. **101**, 6456 (2004).
[5] W. C. Swope, J. W. Pitera, F. Suits *et al.*, J. Phys. Chem. B **108**, 6582 (2004).
[6] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[7] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy, Proc. Natl. Acad. Sci. U.S.A. **102**, 6801 (2005).
[8] M. E. Karpen, D. J. Tobias, and C. L. Brooks III, Biochemistry **32**, 412 (1993).
[9] W. Huisinga, S. Meyn, and C. Schutte, Ann. Appl. Probab. **14**, 419 (2004).
[10] J. Kubelka, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **329**, 625 (2003).
[11] M. Wang, Y. Tang, S. Sato, L. Vugmeyster, C. J. McKnight, and D. P. Raleigh, J. Am. Chem. Soc. **125**, 6032 (2003).
[12] M.-Y. Shen and K. F. Freed, Proteins: Struct., Funct., Genet. **49**, 439 (2002).
[13] M. R. Shirts and V. S. Pande, Science **290**, 1903 (2000).
[14] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[15] D. van der Spoel and E. Lindahl, J. Phys. Chem. B **107**, 11178 (2003).
[16] H. Fujitani, Y. Tanida, M. Ito, G. Jayachandran, C. D. Snow, M. R. Shirts, E. J. Sorin, and V. S. Pande, J. Chem. Phys. **123**, 084108 (2005); E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
[17] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[18] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[19] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[20] A. E. Garcia and K. Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. **99**, 2782 (2002).
[21] W. D. Cornell, P. Cieplak, C. I. Barly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[22] A. E. Garcia (private communication).
[23] T. Darden, D. York, and L. Pederson, J. Chem. Phys. **98**, 10089 (1993).
[24] M. Neumann and O. Steinhauser, Mol. Phys. **39**, 437 (1980).
[25] B. Zagrovic, C. D. Snow, S. Khaliq, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 153 (2002).
[26] M. Buscaglia, J. Kubelka, W. A. Eaton, and J. Hofrichter, J. Mol. Biol. **347**, 657 (2005).
[27] http://dasher.wustl.edu/tinker/
[28] A. Leach, *Molecular Modeling*, 2nd ed. (Prentice-Hall, Englewood Cliffs, NJ, 2001).
[29] B. Zagrovic and V. S. Pande, J. Comput. Chem. **24**, 1432 (2003).
[30] B. Zagrovic and V. S. Pande, J. Comput. Chem. **24**, 1432 (2003).

[31] http://www-unix.mcs.anl.gov/mpi/
[32] W. C. Swope, J. W. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[33] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, Nat. Struct. Biol. **4**, 180 (1997).
[34] M.-Y. Shen and K. F. Freed, Biophys. J. **82**, 1791 (2002); M. W. Mahoney and W. L. Jorgensen, J. Chem. Phys. **114**, 363 (2001).
[35] Y. M. Rhee and V. S. Pande, Berkeley Mini Statistical Mechanics Meeting, Berkeley, CA, 2005 (private communication).
[36] I. C. Yeh and G. Hummer, J. Am. Chem. Soc. **124**, 6563 (2002).
[37] N. Singhal and V. Pande, J. Chem. Phys. **123**, 204909 (2005).
[38] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[39] C. D. Snow, Y. M. Rhee, and V. S. Pande, Biophys. J. (in press).