THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

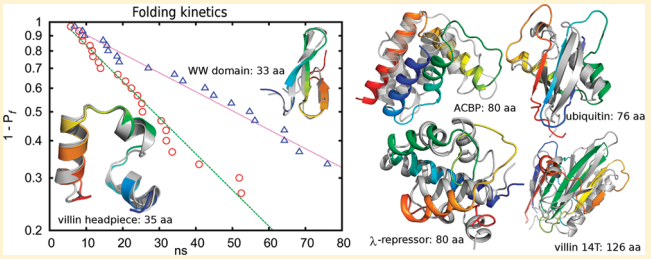
Discrete Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization

David Shirvanyants,[†] Feng Ding,[†] Douglas Tsao,[‡] Srinivas Ramachandran,[†] and Nikolay V. Dokholyan*[,†]

[†]Department of Biochemistry and Biophysics, School of Medicine, and [‡]Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599, United States

Supporting Information

ABSTRACT: Until now it has been impractical to observe protein folding in silico for proteins larger than 50 residues. Limitations of both force field accuracy and computational efficiency make the folding problem very challenging. Here we employ discrete molecular dynamics (DMD) simulations with an all-atom force field to fold fast-folding proteins. We extend the DMD force field by introducing long-range electrostatic interactions to model salt-bridges and a sequence-dependent semiempirical potential accounting for natural tendencies of certain amino acid sequences to form specific secondary structures. We enhance the computational performance by parallelizing the DMD algorithm. Using a small number of commodity computers, we achieve sampling quality and folding accuracy comparable to the explicit-solvent simulations performed on high-end hardware. We demonstrate that DMD can be used to observe equilibrium folding of villin headpiece and WW domain, study two-state folding kinetics, and sample near-native states in ab initio folding of proteins of ∼100 residues.



INTRODUCTION

Uncovering the relationship between protein structure and its sequence is the cornerstone problem of biophysics. The structure−sequence relationship is an inherent component of the protein folding problem and of many important biological processes involving conformational transitions in proteins. Our understanding of protein conformational behavior has greatly benefited from computer simulations. Computer simulations have played an instrumental role in biophysics due to the development of high-performance sampling algorithms and accurate potential functions (also known as force fields).[1−6] Recently, molecular simulations have made immense progress in both directions, allowing probing of milliseconds-scale dynamics of explicitly solvated and charged biopolymers.[7,8]

The ab initio folding of proteins (deducing the native fold relying solely on physics of interactions) has long been the holy grail of protein simulations.[9−11] There has been notable success in the ab initio folding of short (<50 residues) polypeptides. Several studies have been able to sample folding to near-native structures (those that are close to native structure with high statistical significance[12,13]) of villin headpiece,[7,14,15] WW-domain,[15−17] and Trp-cage,[15,18−22] at least as isolated events. A recent study[7] has succeeded in producing simulation trajectories with well populated native states for villin headpiece and WW-domain. Many of these successes have only been achieved due to advanced rapid-sampling protein simulations, which still belong to the realm of large-scale computer clusters[1−4,23,24] or powerful dedicated supercomputers.[25] The time-scale that can be reached by large-scale computer clusters and dedicated supercomputers is within the submillisecond range. While only a few of the fastest folding proteins fold within the millisecond time scale,[26−28] the folding of larger proteins still remains a distant aim. To date, there are no published studies of sampling near-native conformations of proteins with sequence lengths larger than 80 amino acids in ab initio computer simulations.

Coarse-grained methods have been proposed to optimize the computational resource utilization. These methods make use of the time-scale separation that exists in many systems between relatively slow processes of physical interest (such as protein conformational changes) and fast processes (such as atomic bond and valence angle vibrations, or water diffusion) that can be neglected in the studies of long-time scale processes. The obvious challenge of coarse-grained methods is properly selecting the level of detail to preserve the phenomena of interest while avoiding unnecessary computations. In this study, we focus on detailed modeling of proteins using the recently developed approach of discrete molecular dynamics (DMD),[15,29−32] which uses the *implicit solvent* model combined with *atomic-level* details of the protein macromolecule. Previously, we constructed the DMD force field using the CHARMM effective solvation model by Lazaridis and Karplus[33] to model the electrostatic interactions with the solvent and explicit modeling of hydrogen bonds to model the

Special Issue: B: Macromolecular Systems Understood through Multiscale and Enhanced Sampling Techniques

Received: November 28, 2011
Revised: January 24, 2012

 © XXXX American Chemical Society dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX−XXX

electrostatic interactions between polar/charged atoms. We have applied DMD methods for simulations of biopolymers and have demonstrated its ability to reproduce proteins equilibrium dynamics with accuracy comparable to the accepted MD methods.[9,15,29−32,34] However, as the protein length increases, the accessible conformational space grows exponentially which requires adequately longer sampling and results in the accumulation of the inherent inaccuracies of the force field, thus limiting the ability of current methods to achieve native folding of large proteins. An improved conformational sampling can be achieved with replica exchange simulations.[35] The replica exchange approach has allowed us to observe the folding of several small fast-folding proteins to their near-native states.[15] However, it is not straightforward to extract the folding kinetics from replica exchange simulation trajectories since the temperatures in each replica follow a random walk. Therefore, in order to study folding of larger proteins, and especially, the kinetics of folding process, it is necessary to improve the computational sampling methods and force field accuracy.

Here, we extend our approaches in order to access longer time scales and also larger systems. We extend the DMD force field by introducing long-range electrostatic interactions, which allow us to model salt-bridges. We also include a sequence-dependent semiempirical potential accounting for natural propensities of certain amino acid sequences to form specific secondary structures. We also enhance the computational performance by parallelizing the DMD algorithm. We focus on practical applications of our method such as real time performance and its scaling ability. We benchmark our model by studying folding equilibrium and kinetics for the group of fast-folding proteins. We also test our DMD method on the folding of larger proteins ranging from 60 to 120 amino acids. To our knowledge, this is the first study of the computer simulation sampling of near-native conformations of proteins >80 residues long using up to 32 computer processors, which is a very modest amount of commonly available computer hardware. Using a small number of commodity computers, we are able to achieve sampling quality and folding accuracy comparable to the explicit-solvent simulations running on high-end computer hardware. We believe that this study clearly demonstrates feasibility of protein folding and its related tasks using commodity computers.

METHODS

Discrete Molecular Dynamics Simulation of Proteins. The DMD method[15,36−38] is an event driven simulation method using a discrete potential energy function ("force field"). It is numerically equivalent to traditional MD up to the discretization step. In the limit of small potential energy discretization step Δx, DMD will produce trajectories identical to traditional MD in the limit of small time step Δt for the same force field.

DMD features a reduced amount of calculations compared to traditional MD, as there is no need to compute forces and accelerations. Instead, DMD consists of a sequence of atomic collisions. In MD, atoms move with constant accelerations during the integration step. In DMD, atoms move with constant velocities between collision events. The benefit of using a discretized potential function in DMD is similar to MD with an adaptive time step,[39−42] where slower motions (shallow potential wells in MD, wide potential steps in DMD) are computed with larger time step Δt than the high frequency oscillations (sharp potential wells in MD, narrow steps in DMD) such as bonded interactions. The earlier DMD implementations faced challenges of complex event scheduling algorithms,[36] high memory usage,[36] and difficulties of parallelization.[43] However, with advances in computer technology, event driven simulation algorithms[43−48] have overcome these earlier problems. In addition to the computational efficiency of DMD, its event-driven nature allows flexible modeling of specific interactions that define the structure and dynamics of biomolecules.[15]

In this study, we use the *all-atom* protein model developed by Ding et al.[15] that has been extended to account for long-range charge−charge interactions and sequence-dependent local backbone interactions. The all-atom protein model[15] is based on the CHARMM19 energy function along with EEF1 solvation model[33] and an explicit hydrogen bonding potential. The discrete representation of DMD potential allows simple and efficient implementation of hydrogen-bond properties of directionality and saturation, as it permits instantaneous switching of interaction potentials between the atoms when bonds are formed. Here we extend the DMD force field to take into account long-range charge−charge interactions in addition to the short-range interactions of polar groups with each other (the formation of hydrogen bonds) and with the solvent (provided by EEF1 solvation model). Long-range electrostatic interactions stabilize the native state of the protein,[49−51] and in our simulations of short proteins, we observe higher populations of near-native states when long-range interactions are included (Supporting Information Figure S1), despite the simplistic representation of electrostatics in our simulations. We observe an even higher population of near-native states when an additional force field term that accounts for sequence-dependent backbone interactions is included (Supporting Information Figure S1). This sequence-dependent force field correction accounts for subtle differences in short-range interactions between backbone atoms of different amino acids. These sequence-dependent interactions result in different propensities toward certain secondary structures for different amino acid sequences.

Parallel Discrete Molecular Dynamics. DMD is traditionally considered to be intrinsically difficult for parallel implementation. The reason for this difficulty is that in the sequence of DMD events, every subsequent event is computed from the current atom positions and velocities, which themselves result from a preceding chain of events. DMD events include atom collisions, as well as noncollision events needed to model thermostat, hydrogen bonding and to keep track of the atom's nearest neighbors.[52] Any two events in DMD are potentially coupled; that is, the outcome of a preceding atomic collision may affect the time and place of the subsequent events. The common conclusion is that it is impossible to predict many collisions in parallel, since after the first collision other predictions may become invalid. However, there is a workaround for this challenge, if we note that coupling of collisions is limited in time and space. When a certain collision between atoms i and j takes place, its effect propagates through the system with a finite average speed. Therefore many of the earlier collision predictions will remain valid if the participating atoms k and l are located sufficiently far from both i and j and the k−l collision takes place within a short time period after the i−j collision. The feasibility of the event-based parallelization approach has been recently

dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX−XXX

demonstrated by Khan and Herbordt[48] using a scalable implementation on up to 8 CPUs in shared-memory system.

The parallelization approach described in Khan and Herbordt[48] splits the DMD simulation cycles into several stages. First, every collision event is predicted based on the current atoms positions and velocities. Using the predicted collision time, DMD computes new atoms coordinates and velocities. However, unlike the regular DMD algorithm, in parallel DMD, the atoms' state is not immediately updated. Instead, results of the collision evaluation are stored at a temporary memory location (Supporting Information Figure S2). Then every event is tested to exclude collisions that have been superseded by an earlier collision of participating atoms (effect of coupling). Finally, events that have not been excluded are "committed", that is, results previously stored at temporary location are copied to the primary storage of atom properties. Certain stages, such as collision prediction, evaluation, and testing for coupling, can be performed simultaneously for most of events, while the committing stage is executed only serially. The intermediate temporary storage of predicted atom coordinates is required for speculative and parallel processing of predicted collisions. When many collisions are analyzed in parallel, the new atom coordinates, as well as newly predicted collisions are stored in temporary variables. If execution of an event results in cancellation of one of the following events due to coupling, the canceled events will be discarded together with the temporarily stored evaluation results. In a typical DMD simulation, event prediction is the most computationally intensive component, thus its parallelization produces the largest performance gain.

DMD performance depends on the average number of interacting neighbors around an atom. Generally, DMD simulations of compact objects, such as collapsed globular proteins are more computationally costly than simulations of dilute systems such as unfolded protein. DMD simulations of larger compact proteins are slower due to the lower ratio of surface to buried atom number, since buried atoms have on average more neighbor atoms and require more intensive calculations to predict the collision. Performance of the parallel DMD also depends on the fraction of coupled events. Parallel processing of coupled events is impossible, as execution of an earlier event invalidates results of evaluation of the latter event. However, the probability of event coupling decreases as the system size grows (Supporting Information Figure S3). Due to lower rate of the coupled events, efficiency of the event-based parallelization approach increases for large systems and partly compensates the slowdown due to larger fraction of buried atoms. This compensation results in nearly linear dependence of simulation time on protein length for parallel DMD (Figure 3).

Thread synchronization is the most important step in parallel DMD (pDMD) simulation, which is not present in serial algorithm. We need to ensure that two or more threads never simultaneously modify the same shared data. The result of such unsynchronized data access is unpredictable. The synchronization is usually preformed by introducing the so-called "lock mechanism", which allows one thread to access data and make the other thread wait until the first thread is no longer accessing the data. We also detect coupled events and ensure that they are processed in a serial manner. Thread locking and coupled events lead to wasted CPU cycles with adverse effects for parallelization efficiency. Performance of thread synchronization strongly affects overall pDMD performance as handling of

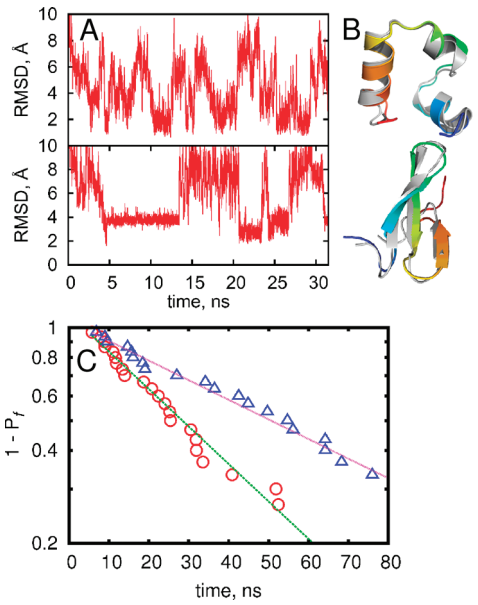


Figure 1. Folding kinetics of the short proteins. (A) Root-mean-square deviation of conformations from crystal structure in the representative trajectory for villin headpiece (upper panel) and WW-domain (lower panel). (B) Near-native structure of villin headpiece (upper panel) and WW-domain (lower panel) observed during simulations displayed using cartoon representation in PyMol. Crystal structures are shown in gray. (C) Probability to observe folding event as function of time for villin headpiece (Δ) and WW-domain(○). Dotted lines indicate exponential fitting.

every collision requires at least one synchronization point using a blocking lock mechanism, and may cause threads to waste time waiting for one another. This problem intensifies for our all-atom force field for DMD simulations of proteins. Compared to the model of a homogeneous fluid with single well interaction potential,[48] DMD of proteins produces more frequent collisions (Supporting Information Figure S4), as it employs complex multiwell potentials, includes a thermostat, and allows for dynamic changes of atomic interactions to simulate chemical reactions and noncovalent reversible bonding, such as hydrogen bonds and salt bridges. In order to minimize the locking overhead, we have developed the parallel DMD algorithm using only nonblocking locks. Nevertheless, the high rate of data exchange between threads makes our implementation of parallel DMD highly dependent on the speed of memory access. On modern processors, such the Intel Xeon or AMD Opteron, the highest exchange rate is achieved between the cores of a single multicore CPU. Therefore, the best scaling of the parallel DMD is achieved when all threads run on CPU cores on the same dye (Supporting Information Figure S5).

Folding Kinetics of Small Fast-Folding Proteins. Starting from a fully extended conformation, we generate 30 independent trajectories of 0.5 μs each at a constant temperature of 300 K. We evaluate the accuracy of folding by observing the root-mean-square deviation (rmsd) of α-carbon atom positions from the crystal structure and fraction of native contacts[53] (Q-value). We use rmsd, Q-value, and internal energy as state variables to construct density of states diagrams in order to analyze sampled conformations (Supporting Information Figures S6 and S7). In the case of WW-domain, we computed rmsd and Q-values for the chain segment

dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX–XXX

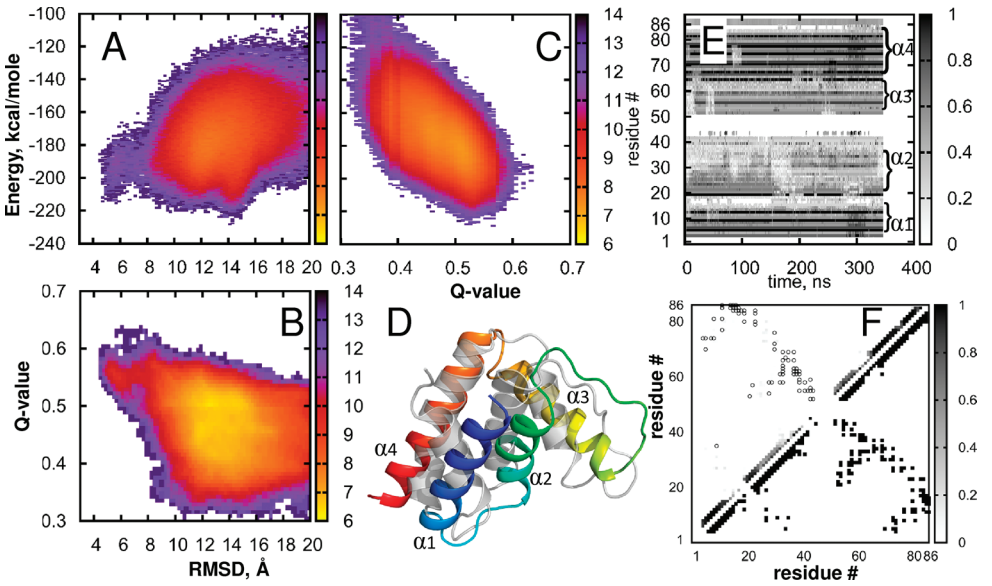


Figure 2. Acyl-coenzyme A binding protein (ACBP). States density maps: (A) energy vs rmsd; (B) Q-value vs rmsd; (C) energy vs Q-value; (D) best fit of simulated (rainbow) to native structure (gray) displayed using the cartoon representation in PyMol; (E) per-residue native contact frequency in a sample trajectory; (F) density of native contacts. The lower triangle shows native contacts, gray squares in the upper triangle indicate the probability to observe native contacts in our simulations, and open circles show native contacts with probability to observe below $P_{threshold}$ = 0.03. More details are provided in the Supporting Information.

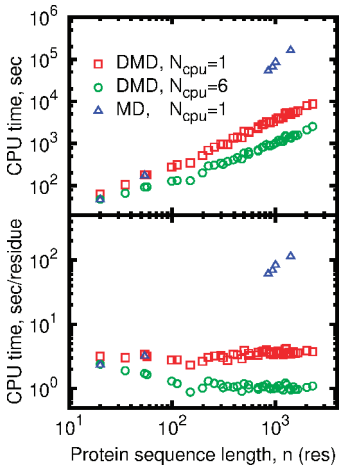


Figure 3. Wall-clock time needed to progress DMD simulation by 100 ps as a function of protein length. (A) Absolute time in seconds of runtime. (B) Normalized time in seconds of runtime per one residue. MD performance was evaluated with Gromacs 4.0.5 [2] using SPC/E water model and time step of 4 fs.

between the conserved W11 to W34, and in villin headpiece, we analyzed the segment between S43 and L75. We have excluded the unstructured protein segments from our analysis in order to minimize the effect of the random fluctuations of these segments on our structural studies.

Equilibrium Protein Folding: Sampling the near-Native States of Larger Proteins. Similar to short proteins, we start simulations from a fully extended conformation and generate 32 independent trajectories at constant temperature of 300 K for each of the test proteins, listed in the Supporting Information Table S1. Using state diagrams derived from rmsd, Q-value, and potential energy (Figure 2, Supporting Information Figures S8–S11), we characterize the quality of sampling and accuracy of force field. We compute the time-dependent fraction of native contacts per residue and matrix of native contact formation probabilities to analyze the propensity of protein structural elements to the native conformation. Additionally, we estimate the structure predictive ability of the DMD force field based on the commonly used measure of global distance test.[54] To minimize contribution of random fluctuations, we have excluded highly mobile residues from our rmsd calculation. These excluded regions are five residues at C-terminal in villin 14T, three residues at N-terminal, and six residues (M46–G51) in the unstructured connecting segment in ACBP.

We use the ratio of the cumulative length of N simulation trajectories of length τ_{max} to the experimental folding time $\zeta = N\tau_{max}/\tau_f$ as a rough estimate of the quality of sampling. Mainly due to the use of an implicit solvent model, folding times of short proteins in DMD simulations are approximately 50 times shorter than experimental folding times according to our folding kinetics study of small, fast-folding proteins. In other words, sampling required to observe one folding event of villin headpiece or WW domain on average is $\zeta \sim 0.02$. Assuming that this ratio holds for longer proteins, for a protein with an experimental $\tau_f \sim 2$ ms (such as ubiquitin), we can observe on average one folding event in a single 40 μs long trajectory. In practice, 32 trajectories of 0.3–0.5 μs long add up to a cumulative length of 9–15 μs (Table 1). The achieved sampling is less than needed for observing one or more of folding events, but it is sufficient to evaluate the performance and application of the force field and the DMD simulation algorithm.

We can estimate DMD sampling efficiency and performance of the force field by characterizing the sampled structures with the smallest rmsd to the native state. Strictly speaking, the smallest rmsd has the nature of an extreme value and does not measure the force field ability to correctly reproduce the entire potential energy landscape. Nevertheless, considering an innumerably large number of conformations available to a polypeptide chain,[55] the smallest rmsd can be used to estimate the force field ability to provide the necessary bias toward the

dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX–XXX

Table 1. Summary of Equilibrium Folding Simulations Results[a]

name	length	min C_α rmsd	P-value	sampling, ζ	GDT-TS
villin headpiece	33	1.2 Å	10^{-215}	10	55.5 ± 1.7
WW domain	23	0.6 Å	10^{-308}	6	61.4 ± 3.7
ACBP	76	4.1 Å	10^{-58}	0.0019	34.3 ± 1.3
ubiquitin	76	6.6 Å	10^{-16}	0.0055	33.1 ± 1.8
SH3	56	6.1 Å	10^{-15}	0.0096	27.6 ± 1.2
λ-repressor	79	5.5 Å	10^{-29}	0.0078	34.8 ± 1.2
villin 14T	121	9.9 Å	10^{-6}	0.00064	18.1 ± 0.7

[a]The length shown for the segment actually used to compute rmsd. The GDT-TS score was computed for the lowest potential energy conformations. The significance of the observed rmsd values was estimated using the distribution of rmsd values of random protein alignments.[13]

native state. In order to evaluate the force field bias toward a native state, we calculate the probability of observing a smallest rmsd structure by chance. A recent analysis[13] finds that rmsd for alignments of pairs of random proteins of M residues can be well-described by the Gumbel distribution function $f(x) = (1/\sigma)e^{-(x-\mu)/\sigma}e^{e^{-(x-\mu)/\sigma}}$ with a peak at $\mu = 3.37M^{0.32}$ and scale of $\sigma = 0.48M^{0.32}$. The selectivity of DMD force field to the native state basin can be characterized by the ratio of the fraction of near-native conformations with given rmsd to the P-value computed from the rmsd for random structures (Table 1).

Given the a priori insufficient sampling to observe the complete folding of the larger proteins, we estimate the predictive capability of DMD using the GDT score.[54] This score takes into account both local and global protein structure, which makes it less sensitive to the presence of outlier fragments as compared to rmsd.

RESULTS AND DISCUSSION

Folding Kinetics of Small Fast-Folding Proteins. To evaluate the performance of our method, we study the folding equilibrium and kinetics of small, fast-folding proteins. Fast-folding peptides such as WW domain or villin headpiece are the popular benchmarks for computational folding methods. WW domain is an all-beta domain of 39 residues found in many proteins and capable of binding proline-rich sequences. The folding rate of engineered fast-folding mutants[27] of WW domain is of the order of 10^5 s^{-1}. For this study, we have utilized a 34 residue WW domain (residues 6−39) of the hPIN1 FIP mutant[56] (PDB ID 2F21). Villin headpiece[57] (PDB ID 1YRF) is the all-alpha fragment of 35 residue of an actin-binding protein villin. Villin headpiece is an ultrafast folding protein, with folding times of certain sequences reaching 0.2 μs.[26]

We have shown previously[15,58] that the DMD method is capable of sampling folded protein states within 2 Å of root-mean-square deviation of backbone atoms for several small proteins using replica-exchange simulations. With the updated DMD force field combined with the enhanced sampling enabled by parallel computing (see the Methods section), we are able to sample multiple folding-unfolding transitions within a single DMD trajectory at constant temperature (Figure 1A). For villin headpiece, we observe structures that feature rmsd as low as ∼1 Å from the crystal structure, while simulations of the WW domain feature structures with rmsd ∼2 Å from the crystal structure (Figure 1B, Supporting Information Figures S6 and S7). Given that reference crystal structures themselves have finite resolution (1 Å for villin headpiece and 1.5 Å for WW domain), we can infer that DMD simulations have accurately reproduced the experimental crystal structures. Folding of all-β proteins constitute a significant challenge[59] as β-strands are stabilized by tertiary contacts and their formation requires cooperativity between residues located far from each other along the backbone. Further, we observe both the proteins to spend tens of nanoseconds in their near-native folded states, suggesting that these states are not transient conformations, but are associated with energy minima.

Since the simulations of the Fip35 WW-domain and villin headpiece feature multiple folding−unfolding transitions within ∼30 ns, we expect at least one folding event in every independent trajectory of 0.5 μs each. To estimate the average folding time $\langle\tau_1\rangle$, we perform multiple independent folding simulations to compute the probability $P_f(\tau_1)$ that a fully stretched polypeptide chain will fold to a near-native state after a given period of time τ_1 in our simulations. Since our initial configuration is always a stretched chain, $\langle\tau_1\rangle$ is not a true average protein folding time τ_f but only an approximation of folding time. However, given that the initial collapse time $\langle\tau_0\rangle < 0.1\langle\tau_1\rangle$ is small compared to folding time, and initial velocities are randomized at every run, we consider $\langle\tau_1\rangle$ as a good approximation of τ_f.

Since our DMD simulations are based on the implicit solvent model, we expect that our estimates of $\langle\tau_1\rangle$ are significantly smaller than experimentally observed protein folding times. This acceleration is due to the larger self-diffusion constant of protein chain and faster segmental dynamics in the absence of collisions with solvent molecules. It is possible to reproduce experimental diffusion rates using a method for correction of protein dynamics proposed by Javidpour et al.[60] However, for simplicity we assume that diffusion acceleration is independent of protein sequence and is of the same magnitude for all our test proteins. Thus, protein folding time computed by our DMD method primarily characterizes accuracy of the potential energy function (force field).

We define $P_f(\tau_1)$ as the fraction of trajectories that have reached rmsd < 2.2 Å from the native state at least once within time τ_1 (see the Methods section; Figure 1C), where a threshold of 2.2 Å was selected as a separation between the folded and unfolded state (Supporting Information Figures S6, S7). The exponential decay of $P_f(\tau_1)$ indicates the presence of a single rate controlling barrier, in line with experimental observations.[26,27] Single exponential two-parameter fitting $P_f(\tau_1) = e^{(\tau_1-\tau_0)/\langle\tau_1\rangle}$ produced average folding times of 35 ns for villin headpiece and 68 ns for WW-domain. The second parameter $\tau_0 \sim 3$ ns takes into account the initial collapse time from the stretched conformation. It is interesting to note that the folding time of WW-domain is about two times that of villin headpiece. The absolute values of folding times in our simulation are about 2 orders of magnitude smaller than the times observed experimentally, as expected for our model. However, the approximately 2-fold difference of folding times agrees with experimental observations.

Equilibrium Protein Folding: Sampling the near-Native States of Larger Proteins. The successful application of our new force field to short proteins has motivated us to perform folding simulations of longer proteins with the extended all-atom force field. Even though there have been studies where small proteins have been folded successfully

dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX−XXX

(both villin headpiece and WW-domain have been folded computationally within 1 Å deviation from crystal structure[7]), folding of proteins beyond 50 amino acids is still a challenging task. In order to evaluate the ability of the modified DMD force field to predict native protein conformations, we chose a group of larger proteins whose folding mechanism has been studied both theoretically and experimentally. These proteins feature different ratios of secondary-structure elements and a relatively short experimental folding time. The selected proteins are known to fold on the millisecond scale:[61] all-β SH3 domain (1 ms, 56 amino acids (aa)), $\alpha-\beta$ ubiquitin (2 ms, 76 aa), all-α λ-repressor (2 ms, 80 aa), all-α ACBP (~6 ms, 86 aa), and villin 14T (15 ms, 126 aa). Unlike the case of villin headpiece and WW-domain (described in the previous section), estimated folding times of other test proteins are much longer than individual simulation trajectories. Studying folding kinetics for these long proteins requires application of special approaches[62,63] that are beyond the scope of the current work. Thus, we focus only on studying the ability of DMD to sample native-like conformations in multiple independent equilibrium folding simulations (see the Methods section).

With the exception of villin 14T, the sampling quality achieved in our DMD simulations is sufficient to observe strong correlation between low backbone rmsd, high Q-value, and low potential energy (Figure 2, Supporting Information Figures S6–S11). The conformations sampled in the DMD trajectories recapitulate many features of the native folds, such as hydrophobic cores or characteristic fragments. Below we briefly discuss the specific behavior of each protein.

Acyl-Coenzyme A Binding Protein (ACBP). This is a small four-helix bundle consisting of 86 amino acids which folds in ~6 ms[64] in an apparent two-state process.[64,65] Formation of contacts between 8 residues of helix α1 and α4 (Figure 2D) was determined to be the rate-limiting step.[64] We use bovine ACBP[66] (PDB ID 2ABD) as the reference structure. In the lower-rmsd state, the core is well-packed and the rate limiting structure consisting of residues F5, A9, L15, Y73, I74, and L80 is formed. The per-residue fraction of native contacts (Figure 2E) for this α-helical protein is mostly contributed by intrahelical contacts. We also observe the early formation of the secondary structure during the simulation, which is in line with experimental data on ACBP unfolding.

Ubiquitin. This is a 76 amino-acid highly conserved β-grasp protein that folds in ~1 ms.[67] We use the human ubiquitin[68] (PDB ID 1UBQ) as the reference structure. In most of the simulation trajectories, formation of native contacts occurs first in the β1- and β2-strands and the α1 helix at the N-terminal fragment (Supporting Information Figure S8D,E). This order is consistent with the ubiquitin folding pathway suggested by Sosnick et al.[69]

Src Homology Domain (SH3) . This is a conserved, independently folding, protein binding domain arranged in a characteristic β-barrel consisting of five, sometimes six β-strands packed into two orthogonal β-sheets with a long unstructured loop between β-strand 1 and 2 (RT loop). We use the fastest known folding variant of FYN SH3[70] (PDB ID 1FYN, 56 residues) with two mutations (A39G and V55F)[71] as the reference structure in our simulations.

In our simulations, we are able to observe the experimentally detected formation of the hydrophobic core by I28, A39, and I50 at the early stages of SH3 folding[71] (Supporting Information Figure S9F). However, the primary difficulty in sampling structures close to the native state as defined by the crystal structure is due to the improper packing of the RT-loop. Contrary to the expected packing of the unstructured RT-loop on top of the β-barrel, we observe an RT-loop in an open conformation with a tendency to form either α-helical or β-strand secondary structures (Supporting Information Figure S9D). However, the RT-loop itself and core β-barrel are formed, and several trajectories have sampled near-native structures with rmsd ≈ 6 Å.

λ-Repressor. This protein consists of five helices, with folded state stabilized by the hydrophobic core formed by L36, L40, and I47.[72] We use the structure[70] (PDB ID 1LMB) of 80-residue segment (residues 6–85) of the fast-folding λ-repressor mutant[73] as the reference. In our simulations, we did not observe the formation of the native hydrophobic core, although structures very similar to the native state (Supporting Information Figure S10D, rmsd ≈ 5.5 Å) can be stabilized by an alternate hydrophobic core, such as the core formed by I46, I68, and F76 in the structure. Nevertheless, similar to other proteins, sampling from 32 trajectories shows significant correlation between small rmsd, high Q-value, and low potential energy (Supporting Information Figure S10A–C).

Villin 14T. This protein features two hydrophobic cores on sides of the central β-sheet.[74] Core 1 is formed by predominantly aliphatic residues of the long helix (α2, amino acids (aa) 80–90), and core 2 is formed by short helix α3 (aa 103–110), and β-strands β6 (aa 36–40) and β7 (aa 114–118) with a high fraction of aromatic residues.[74] We use the chicken villin[75] (PDB ID 2VIK) as the reference structure. In most of the trajectories, we observe rapid formation of the central β-sheet and presence of many hydrophobic contacts of core 2 (Supporting Information Figure S11E). However, the conformations in most of our trajectories feature helical content lower than that of the native state. In particular, the longest helix α2 is often replaced by one or two β-strands. Nevertheless, the DMD force field correctly captures many important structural features of villin 14T such as the central β-sheet and short helix α3.

CONCLUSIONS

For short proteins, we generate 0.5 μs long trajectories using DMD, which are sufficient to observe at least one folding transition event, with most trajectories featuring multiple folding–unfolding transitions. In the ensemble of trajectories, we compute the average folding time from the exponential decay of fraction of unfolded conformations, the characteristic feature for two-state folding proteins. From our folding simulations, we study the ability of DMD to sample near-native conformations of proteins up to 126 residues long, including an all-β WW-domain and mixed α/β-proteins. In all simulations, except for villin 14T, we have sampled structures much closer to the native structure than could be achieved by random sampling, with the *P*-value of the rmsd many orders of magnitude below the fraction of near-native structures observed in the trajectories (Table 1).

We estimate the structure predictive capability of DMD with the commonly used GDT score.[54] Here we limit ourselves to structure prediction for the subset of 1–10 ms folding proteins; however, there is evidence[61] that folding rates of the significant number of studied proteins falls in this range. In our simulations, the average GDT total scores for the lowest energy states in case of long proteins range from 18 for villin 14T to 35 for the λ-repressor. For comparison, most ab initio predictions made in CASP9 in the free modeling category using

dx.doi.org/10.1021/jp2114576 | J. Phys. Chem. B XXXX, XXX, XXX–XXX

common MD algorithms and force fields fall into the range of 15−30,[76,77] indicating that native state prediction with the DMD force field is on par with commonly used structure prediction methods.

Enhanced molecular dynamics simulation methods have been instrumental in routine tasks, such as estimation of protein stability and structure rigidity, correlation analysis, and structure fitting to electron density maps.[78] Application of implicit solvation models enhanced the performance by several orders of magnitude compared to methods utilizing an explicit solvent. However, concerns about the force field applicability range limited the use of implicit solvent models. We demonstrate that the implicit solvent force field of DMD adequately represents the potential energy function of many common proteins and, thus, can be instrumental in studies of many interesting phenomena, such as protein dynamics,[79,80] active site function,[81] and ligand binding,[34] as well as for protein structure optimization.[82]

We demonstrate that the DMD force field in its present state can predict protein core structure at the level of the standard explicit-solvent MD methods, while the DMD algorithm allows for significantly smaller computational efforts than explicit-solvent MD. We show that DMD can be parallelized at a very high collision rate, which opens a new avenue for more computationally intensive modeling of proteins.

■ ASSOCIATED CONTENT

Supporting Information

Detailed description of the DMD force field modifications, parallel DMD implementation, and force field parameter tables, as well as additional state diagrams for all simulated proteins. These materials are available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dokh@unc.edu.

Notes

The authors declare the following competing financial interest(s): Parallel DMD (pDMD) code was developed by Molecules in Action, LLC (Chapel Hill).

■ ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health grant numbers R01GM080742. Parallel DMD (πDMD) code was developed by Molecules in Action, LLC (Chapel Hill). πDMD executables are available at the Molecules in Action, LLC, website (http://moleculesinaction.com).

■ REFERENCES

(1) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903−1904.
(2) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(3) Schulten, K.; Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(4) Kevin, J. B.; Edmond, C.; Huafeng, X.; Ron, O. D.; Michael, P. E.; Brent, A. G.; John, L. K.; Istvan, K.; Mark, A. M.; Federico, D. S.;et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; ACM: Tampa, FL, 2006.
(5) Brooks, B. R.; Brooks, C. L. 3rd; Mackerell, A. D. Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(6) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27−85.
(7) 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.; et al. *Science* **2011**, *330*, 341−346.
(8) Vendruscolo, M.; Dobson, C. M. *Curr. Biol.* **2011**, *21*, R68−70.
(9) Dokholyan, N. V. *Curr. Opin. Struct. Biol.* **2006**, *16*, 79−85.
(10) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751−758.
(11) Zwier, M. C.; Chong, L. T. *Curr. Opin. Pharmacol.* **2010**, *10*, 745−752.
(12) Reva, B. A.; Finkelstein, A. V.; Skolnick, J. *Fold Des.* **1998**, *3*, 141−147.
(13) Jia, Y.; Dewey, T. G. *J. Comput. Biol.* **2005**, *12*, 298−313.
(14) Lei, H.; Wu, C.; Liu, H.; Duan, Y. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4925−4930.
(15) Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010−1018.
(16) Pande, V. S.; Ensign, D. L. *Biophys. J.* **2009**, *96*, L53−L55.
(17) Bolhuis, P. G.; Juraszek, J. *Biophys. J.* **2010**, *98*, 646−656.
(18) Pande, V. S.; Snow, C. D.; Zagrovic, B. *J. Am. Chem. Soc.* **2002**, *124*, 14548−14549.
(19) Simmerling, C.; Strockbine, B.; Roitberg, A. E. *J. Am. Chem. Soc.* **2002**, *124*, 11258−11259.
(20) Ota, M.; Ikeguchi, M.; Kidera, A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17658−17663.
(21) Day, R.; Paschek, D.; Garcia, A. E. *Proteins* **2010**, *78*, 1889−1899.
(22) Irback, A.; Mohanty, S. *Biophys. J.* **2005**, *88*, 1560−1569.
(23) Germain, R. S.; Fitch, B. G.; Mendell, M.; Pitera, J.; Pitman, M.; Rayshubskiy, A.; Sham, Y.; Suits, F.; Swope, W.; Ward, T. J. C.; et al. *J. Parallel Distribut. Comput.* **2003**, *63*, 759−773.
(24) IBM. http://www.research.ibm.com/bluegene/.
(25) Shaw, D. E. *Abstr. Pap. Am. Chem. Soc.* **2009**, 238.
(26) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546−553.
(27) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2369−2374.
(28) Qiu, L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952−12953.
(29) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Fold Des* **1998**, *3*, 577−587.
(30) Proctor, E. A.; Ding, F.; Dokholyan, N. V. *Wiley Interdisc. Rev.: Comput. Molec. Sci.* **2011**, *1*, 80−92.
(31) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *J. Mol. Biol.* **2000**, *296*, 1183−1188.
(32) Khare, S. D.; Ding, F.; Dokholyan, N. V. *J. Mol. Biol.* **2003**, *334*, 515−525.
(33) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133−152.
(34) Tsao, D.; Liu, S.; Dokholyan, N. V. *Chem. Phys. Lett.* **2011**, *506*, 135−138.
(35) Okamoto, Y. *J. Mol. Graph Model.* **2004**, *22*, 425−439.
(36) Rapaport, D. C. *J. Comput. Phys.* **1980**, *34*, 184−201.
(37) Smith, S. W.; Hall, C. K.; Freeman, B. D. *J. Comput. Phys.* **1997**, *134*, 16−30.
(38) Emperador, A.; Meyer, T.; Orozco, M. *Proteins* **2010**, *78*, 83−94.
(39) Franklin, J.; Doniach, S. *J. Chem. Phys.* **2005**, *123*, 124909.
(40) Rakowski, F.; Grochowski, P.; Lesyng, B.; Liwo, A.; Scheraga, H. A. *J. Chem. Phys.* **2006**, *125*, 204107.
(41) Faccioli, P. *J. Chem. Phys.* **2010**, 133.
(42) Izaguirre, J. A.; Reich, S.; Skeel, R. D. *J. Chem. Phys.* **1999**, *110*, 9853−9864.
(43) Miller, S.; Luding, S. *J. Comput. Phys.* **2004**, *193*, 306−316.
(44) Paul, G. *J. Comput. Phys.* **2007**, *221*, 615−625.
(45) Berrouk, A. S.; Wu, C. L. *Powder Technol.* **2010**, *198*, 435−438.
(46) Isobe, M. *Int. J. Modern Phys. C* **1999**, *10*, 1281−1293.

(47) Marin, M.; Risso, D.; Cordero, P. *J. Comput. Phys.* **1993**, *109*, 306−317.
(48) Khan, M. A.; Herbordt, M. C. *J. Comput. Phys.* **2011**, *230*, 6563−6582.
(49) Ripoll, D. R.; Vila, J. A.; Scheraga, H. A. *J. Mol. Biol.* **2004**, *339*, 915−925.
(50) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459−474.
(51) Ibragimova, G. T.; Wade, R. C. *Biophys. J.* **1999**, *77*, 2191−2198.
(52) Rapaport, D. C. *The art of molecular dynamics simulation*, 2nd ed.; Cambridge University Press: Cambridge, 2003.
(53) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 3626−3630.
(54) Zemla, A. *Nucleic Acids Res.* **2003**, *31*, 3370−3374.
(55) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 20−22.
(56) Jager, M.; Zhang, Y.; Bieschke, J.; Nguyen, H.; Dendle, M.; Bowman, M. E.; Noel, J. P.; Gruebele, M.; Kelly, J. W. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10648−10653.
(57) Chiu, T. K.; Kubelka, J.; Herbst-Irmer, R.; Eaton, W. A.; Hofrichter, J.; Davies, D. R. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 7517−7522.
(58) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147−155.
(59) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−77.
(60) Javidpour, L.; Tabar, M. R.; Sahimi, M. *J. Chem. Phys.* **2009**, *130*, 085105.
(61) Gromiha, M. M.; Thangakani, A. M.; Selvaraj, S. *Nucleic Acids Res.* **2006**, *34*, W70−74.
(62) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(63) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; et al. *J. Phys. Chem. B* **2004**, *108*, 6582−6594.
(64) Kragelund, B. B.; Osmark, P.; Neergaard, T. B.; Schiodt, J.; Kristiansen, K.; Knudsen, J.; Poulsen, F. M. *Nat. Struct. Biol.* **1999**, *6*, 594−601.
(65) Thomsen, J. K.; Kragelund, B. B.; Teilum, K.; Knudsen, J.; Poulsen, F. M. *J. Mol. Biol.* **2002**, *318*, 805−814.
(66) Andersen, K. V.; Poulsen, F. M. *J. Biomol. NMR* **1993**, *3*, 271−284.
(67) Roberts, A.; Jackson, S. E. *Biophys Chem* **2007**, *128*, 140−149.
(68) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531−544.
(69) Sosnick, T. R.; Krantz, B. A.; Dothager, R. S.; Baxa, M. *Chem. Rev.* **2006**, *106*, 1862−1876.
(70) Musacchio, A.; Saraste, M.; Wilmanns, M. *Nat. Struct. Biol.* **1994**, *1*, 546−551.
(71) Northey, J. G.; Di Nardo, A. A.; Davidson, A. R. *Nat. Struct. Biol.* **2002**, *9*, 126−130.
(72) Lim, W. A.; Hodel, A.; Sauer, R. T.; Richards, F. M. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 423−427.
(73) Liu, F.; Gao, Y. G.; Gruebele, M. *J. Mol. Biol.* **2010**, *397*, 789−798.
(74) Choe, S. E.; Matsudaira, P. T.; Osterhout, J.; Wagner, G.; Shakhnovich, E. I. *Biochemistry* **1998**, *37*, 14508−14518.
(75) Markus, M. A.; Matsudaira, P.; Wagner, G. *Protein Sci.* **1997**, *6*, 1197−1209.
(76) Shi, S.; Pei, J.; Sadreyev, R. I.; Kinch, L. N.; Majumdar, I.; Tong, J.; Cheng, H.; Kim, B. H.; Grishin, N. V. *Database (Oxford)* **2009**, *2009*, bap003.
(77) CASP9. http://prodata.swmed.edu/CASP9/evaluation/Categories.htm, 2010.
(78) Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. *Structure* **2008**, *16*, 673−683.
(79) Gyimesi, G.; Ramachandran, S.; Kota, P.; Dokholyan, N. V.; Sarkadi, B.; Hegedus, T. *Biochim. Biophys. Acta—Biomembr.* **2011**, *1808*, 2954−2964.
(80) Karginov, A. V.; Ding, F.; Kota, P.; Dokholyan, N. V.; Hahn, K. M. *Nat. Biotechnol.* **2010**, *28*, 743−747.
(81) Kiss, G.; Rothlisberger, D.; Baker, D.; Houk, K. N. *Protein Sci.* **2010**, *19*, 1760−1773.
(82) Ding, F.; Dokholyan, N. V. *PLoS Comput. Biol.* **2006**, *2*, e85.