Challenges in protein folding simulations: Timescale, representation, and analysis. Nature Physics, 6, 751

Article *in* Nature Physics · October 2010

DOI: 10.1038/nphys1713 · Source: PubMed

Yanxin Liu
University of California, San Francisco

Challenges in protein folding simulations: Timescale, representation, and analysis

Peter L. Freddolino,[1,2] Christopher B. Harrison,[1]
Yanxin Liu,[1,3] and Klaus Schulten[1,3,*]

March 23, 2010

[1] Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[2] Current address: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

[3] Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801

[*] Corresponding author. Email: kschulte@ks.uiuc.edu

Abstract

Experimental studies of protein folding processes are frequently hampered by the fact that only low resolution structural data can be obtained with sufficient temporal resolution. Molecular dynamics simulations offer a complementary approach, providing extremely high resolution spatial and temporal data on folding processes. The effectiveness of such simulations is currently hampered by continuing questions regarding the ability of molecular dynamics force fields to reproduce the true potential energy surfaces of proteins, and ongoing difficulties with obtaining sufficient sampling to meaningfully comment on folding mechanisms. We review recent progress in the simulation of three common model systems for protein folding, and discuss how recent advances in technology and theory are allowing protein folding simulations to address their current shortcomings.

Introduction

In recent years molecular dynamics (MD) simulations, originally developed for numerical simulation of simple model systems in statistical mechanics (1), have developed into a powerful tool for studying the structural and dynamic properties of complex biomolecules (see, *e.g.*, (2)) thanks to advances in computing power and refinements of the underlying models. MD simulations of biomolecules typically treat the molecule of interest and surrounding solvent as classical particles interacting through an empirically derived potential energy function (the "force field"). The system's dynamics propagate through time via numerical integration of Hamilton's equations of motion, typically discretized into steps on the order of femtoseconds in length. The information offered by such simulations is no less than an atomic-resolution model of conformational equilibria and structural transitions in the system of interest, providing a wealth of information to interpret, complement, and design experiments.

One of the most challenging applications of molecular dynamics is the simulation of protein folding processes. Such simulations generally must be very long (on the order of many microseconds) to stand a good chance of observing a single folding event, and the force field being used must correctly describe the relative energies of a wide array of unfolded or misfolded conformations that occur during the folding process. The benefits of such simulations are considerable, as they provide detailed information on the *nature* and *relationships* of structures that occur during protein folding processes, and identify key intermediates and barriers to folding. It should be noted that using molecular dynamics simulations to observe entire folding events from unfolded conformations is only one of a wide variety of ways in which molecular modeling calculations are applied to identify native states of proteins and mechanisms through which they fold. Other examples include predicting the folded structure of a given peptide from its primary sequence (*e.g.*, (3, 4)) or using Monte Carlo simulations to follow an approximation of a dynamically realistic folding pathway (*e.g.*, (5, 6)). While other methods offer more computationally efficient ways to identify the native state of a protein, or even likely intermediate states, only atomistic MD simulations of the folding process provide detailed information about transitions between structures that is key to understanding how the folding of a protein actually proceeds. In the present article we use the phrase "folding simulations" to refer exclusively to atomistic molecular dynamics simulations of all or part of the folding process of a protein, in the absence of biasing potentials targeting the folded state. We begin by providing the reader with a brief overview of the recent progress of folding simulations, focusing on a few well-studied model systems. We then discuss the two linked challenges faced by folding simulations, namely continuing to improve the accuracy of representation of proteins in all-atom MD simulations while at the same time improving sampling, and review recent efforts to overcome them.

Long-timescale molecular dynamics simulations of protein folding

Folding simulations pose harsh challenges for molecular dynamics, due to the computational effort involved and the demands for accuracy placed on the force field. Despite these challenges, folding simulations have an established, and growing, track record not only of successfully folding proteins, but of providing quantitative agreement with experimental data and detailed predictions which can be used to test simulated folding behaviors. In this section we review three frequently targeted model systems which, taken together, illustrate the current

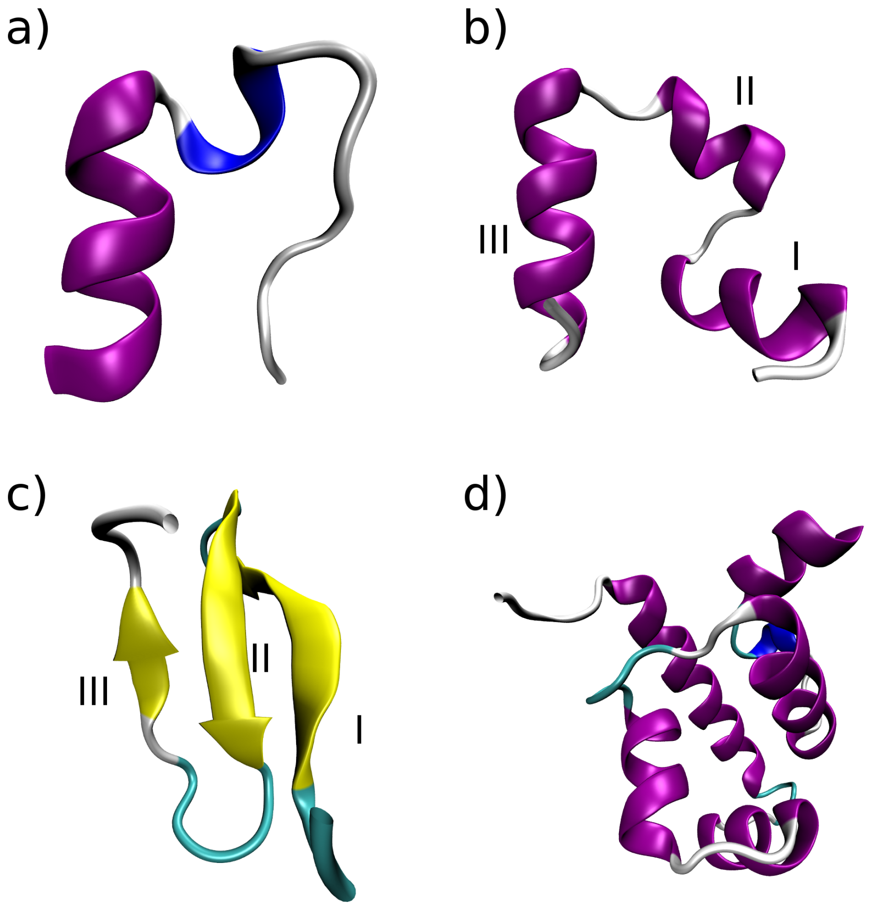

Figure 1: Cartoon representations of proteins discussed in this review. Secondary structures are assigned using STRIDE (97): α helix (purple), β sheet (yellow), turn (cyan), coil (white), or 3_{10} helix (blue). a) Trpcage (PDB code 1L2Y). b) Villin (PDB code 1YRI). c) WW domain (PDB code 2F21). d) λ repressor (PDB code 1LMB). Secondary structure elements for villin and the WW domain are labeled matching discussion in the text.

state of successes and failures encountered in folding simulations: the artificial Trpcage peptide, the chicken villin headpiece subdomain, and the human Pin1 WW domain.

The Trpcage miniprotein (7) (see Fig. 1a) folds in approximately 4 μs, and contains a total of 20 residues. Several early implicit solvent simulations of Trpcage succeeded in folding the protein from a denatured state, and provided realistic estimates of the time required for folding (8–11). Extensive simulations over the following years provided free energy landscapes for folding (using simple order parameters) (12) and even a stability diagram under a variety of thermodynamic conditions (13). Replica exchange simulations revealed an important role for buried water molecules in stabilizing the folded structure (14). Juraszek and Bolhuis employed transition path sampling to study the mechanism of folding/unfolding transitions in Trpcage, finding that the dominant folding pathway involves formation of secondary structure elements only after tertiary contacts are anchored. Their results showed that this pathway coexists with one in which helix formation occurs first (15). Thus, simulations of Trpcage have shown that it is possible to fold a protein from a fully denatured state using unbiased MD simulations. Trpcage simulations highlighted also the importance of water in obtaining a realistic description of Trpcage folding, and provided detailed information on the type of heterogeneous folding mechanism followed by a protein. At the same time, a few challenges still remain: predictions such as the folding pathway partitioning of Juraszek and Bolhuis have not, to our knowledge, been experimentally verified; meanwhile, it has been observed that several thermodynamic inadequacies occur in modern force fields' descriptions of Trpcage. OPLS/AA, for example, incorrectly stabilizes non-native states relative to the native state (16), and AMBER variants have consistently yielded melting temperatures more than 100 K above the experimentally determined value (13).

Computational studies of protein folding often target small portions of natural proteins which have been found to fold rapidly. One example of such a system is the villin headpiece subdomain, a 35-residue three-helix bundle (17) (see Fig. 1b). Wild type villin folds at a rate between $(4.3\ \mu\text{s})^{-1}$ (18) and $(7.4\ \mu\text{s})^{-1}$ (19). The replacement of two lysine residues with norleucine was shown to yield a mutant folding (on average) in less than one microsecond (20). The folding of villin has been subjected to a wide variety of experiments providing data on the kinetics and thermodynamics of folding (18, 21, 22), and contributions from specific contacts to the stability of the transition state (19). Due to its small size and rapid folding, villin was targeted in what was, to our knowledge, the first serious effort to completely fold a protein through atomistic molecular dynamics simulations in explicit solvent (23). While that initial attempt produced only a one microsecond trajectory, and did not reach the native state, a number of subsequent efforts succeeded in reaching the native state from an initially unfolded structure for either or both of the wild type and norleucine mutant proteins, over timescales consistent with experiment (*e.g.*, (6, 24–30)). An early generation of hypotheses regarding villin folding from molecular dynamics simulations were tested through measurement of folding rates of proposed mutants and found to be incorrect (18). More recently, simulations from different groups have lead to several distinct proposals regarding villin folding, (6, 28–30) which now await further testing. One example (from (30)) is shown in Fig. 2: from an initially disordered structure, the protein undergoes hydrophobic collapse and forms a pre-folded conformation with correct secondary structure but incorrect positioning of helix I. The rate limiting step (corresponding to a single long relaxation time observed in experiments) is the partial dissociation of the secondary structure elements from each other, which then re-associate to form the folded structure. Consistently, recent solid state NMR experiments have shown the existance of a long-lived intermediate state with native secondary structure but disordered tertiary structure (31). Validation of the predictions of any of the currently proposed models would provide an atomistically-detailed view of exemplary villin folding pathways (although such a picture would certainly not be complete due to the vast structural heterogeneity expected during folding (29, 32)). At the same time, careful examination of any folding models which do not withstand experimental scrutiny should provide data which can be used to refine protein force fields to aid in future folding attempts.

Where the villin headpiece subdomain serves as an excellent model system for the folding of small α-helical proteins, the WW domain of human Pin1 (henceforth WW domain) has recently become a similar system for simulations of small β-sheet proteins. The WW domain consists of a three-stranded antiparallel β-sheet with the strands connected by tight hydrogen-bonded loops (33) (see Fig. 1c). Analysis of the folding properties of a wide variety of mutants (particularly in the loops) has shown that formation of the first turn (between strands I and II) is the rate limiting step in folding (33, 34), and that stabilizing mutations can shift the WW domain from two-state folding to incipient downhill (*i.e.*, very low barrier) folding. The present experimental evidence provides information on the specific structural change occurring during the rate limiting step, but does not currently reveal other aspects of the pathways followed during WW domain folding. Most crucially, the order of hydrophobic collapse, formation of turn two, and generation of the native β-sheet hydrogen bonding network relative to formation of loop one remains unknown. Initial attempts to study these aspects of WW domain folding generally used coarse grained models due to the slow ($>$50 μs) folding of the wild type protein, and provided a variety of mutually exclusive predictions regarding the order of formation of different structural elements during folding (35–37).

Recently, the discovery of WW domain mutants that fold in less than 15 μs prompted attempts to fold the WW domain through all-atom explicit solvent folding simulations (38). The initial simulations failed to reach the native state and instead became trapped in helical intermediates, which were shown through subsequent free energy calculations to be, in fact, lower in free energy than the native state in the applied force field (39). More recently, a large array of distributed implicit-solvent folding simulations using a different force field provided a small number of folding trajectories; these trajectories suggested the presence of a large amount of kinetic and mechanistic heterogeneity, showing that the questions regarding WW domain folding noted above may in fact be unanswerable (40). On the other hand, the general structural heterogeneity of even the “folded” conformations from that study, and relatively poor agreement with the experimental structure, may indicate the presence of similar (albeit less severe) force field inaccuracies to those noted in (39).

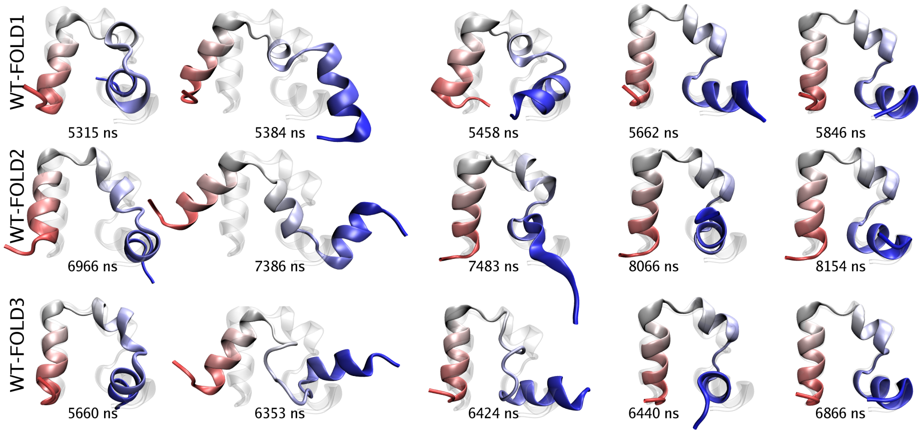

Figure 2: Representative snapshots of the trajectory followed by villin headpiece from the pre-folded intermediate to the native state, with labels corresponding to the discussion in the text. Protein coloring runs blue to red from N terminus to C terminus; the crystal structure is shown as a transparent gray cartoon for comparison. Reprinted from Biophysical Journal 97; Peter L. Freddolino and Klaus Schulten; Common structural transitions in explicit-solvent simulations of villin headpiece folding; 2338–2347; Copyright 2009, with permission from Elsevier.

Challenges in protein folding simulations

As a group, folding simulations (and indeed, MD simulations in general) have throughout their history been faced with two mutually antagonistic challenges. Simulations must be as long as possible in order to obtain reasonable statistics, due to the long correlation times inherent in MD trajectories and the fact that even a single protein folding trajectory requires immense amounts of computing effort. Furthermore, as many such trajectories as possible must be obtained to provide a complete picture of the folding process (40). At the same time, as illustrated by the various points of disagreement still present between simulation and experiment, the accuracy of modern MD force fields in describing long term structural dynamics of proteins remains imperfect, and thus either additional refinements of parameters for force fields, or the use of new developments such as computationally tractable polarizable force fields (*e.g.*, (41, 42)) will be required in many cases for accurate folding simulations.

Timescales and data analysis

In order to address the sampling problem, a number of innovative approaches have been applied to produce recent folding simulations, with varying degrees of generality. At the simplest level, both advances in the processing power available in a given computing node, and the continuing expansion of the avail-

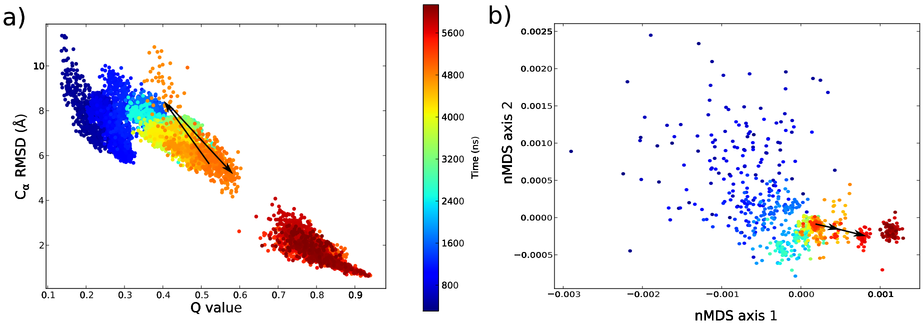

Figure 3: Projections of a villin folding trajectory (corresponding to WT-FOLD1 in Fig. 2) onto two-dimensional surfaces. a) Projection onto Q/C_α-RMSD space; Q represents the fraction of native contacts formed, and is defined as in (98). b) Embedding of the trajectory into a two-dimensional space chosen via nMDS (58) based on the dihedral angles of the protein. In both cases frames prior to the intial hydrophobic collapse are omitted for clarity; the earlier frames are very low Q, high C_α-RMSD, and are scattered randomly in nMDS space. Two arrows are drawn showing the path taken between the 5315 ns, 5384 ns, and 5458 ns time points (*c.f.* Fig. 2); this path corresponds to the crossing of the putative free energy barrier identified in (30).

ability of supercomputing time to researchers, have enabled folding simulations through general purpose computing resources (*e.g.*, (30)). The expansion of such resources is particularly powerful in tandem with recent efforts to improve the performance of MD programs (38, 43, 44), and should continue to provide increasing sampling capabilities to a broad base of researchers.

Several of the most notable simulations of protein folding have instead involved the Folding@Home network (45), a unique distributed computing resource consisting of over 300,000 CPUs donated by users around the world. The Folding@Home architecture, with its massive parallelism, but low density, is particularly well suited to the simultaneous evaluation of large numbers of trajectories, and typical Folding@Home simulations consist of hundreds or thousands of relatively short trajectories (a small fraction of which fold) rather than 1-10 full length trajectories (9, 25, 40).

Another solution to the sampling problem in molecular dynamics folding simulations is the use of special-purpose hardware designed specifically for MD simulations. The most prominant recent example is the Anton platform, a complete special-purpose supercomputer containing sets of application-specific integrated circuits (ASICs) which perform the various tasks required in an MD simulation (46).

While the performance of special-purpose hardware can vastly exceed that provided by general-purpose clusters, such hardware requires substantial resources to develop, and does not benefit from the constant, consumer driven advances that occur with ordinary clusters. The recent development of general-purpose graphics processing units (GPGPUs) offers the possibility of per-node performance orders of magnitude better than that of general purpose computers (47), while at the same time using consumer hardware that will be improved due to market demands for better workstation and gaming graphics. Because GPUs rely on parallel processing of a large array of data using identical procedures to obtain optimal performance, molecular dynamics simulations (involving identical floating-point calculations on a large array of atoms) can be mapped well to the GPU architecture (47). While GPU computing was previously employed in a limited manner for molecular modeling applications (48), the recent advent of a general purpose programming interface for GPUs that does not require extensive low-level effort on the part of the programmer has lead to an explosion of GPU implementations of molecular dynamics simulations (*e.g.*, (49–52)). Such implementations quote accelerations between 10- and 1000-fold over CPU-only implementations, depending on the exact algorithm and target application under consideration, and definition of an “equivalent” CPU-only competitor.

The performance offered by GPGPU-accelerated molecular dynamics simulations does not at present match that of the Anton platform, but as noted previously the performance of GPGPUs is expected to improve over time simply as a function of consumer-driven demand, and thus they may become an increasingly attractive option for long timescale molecular dynamics simulations in the near future. One of the principal challenges associated with applying GPGPUs to molecular dynamics simulations is that network latency between multiple nodes becomes increasingly problematic as the individual nodes become faster (53); these challenges are less relevant in the case of protein folding simulations, where one would be best served by running dozens of simulations of small systems in parallel, each on a single GPU-equipped node.

Whether one obtains a few long folding trajectories or a large array of short folding simulations, eventually it becomes necessary to synthesize the data into as reliable as possible a picture of the folding process of the protein of interest. This, in turn, means that one wishes to understand what general features are present and how they evolve as the protein forms more and more of its native contacts, identify frequently occupied conformations or misfolded traps, and characterize transitions between those conformations. Such analysis is nontrivial given the large amounts of data present in folding trajectories, and requires specialized methods. One of the most common tools for visualization and analysis of protein folding pathways is the projection of the trajectory onto a low (frequently 2) dimensional surface, both to track the progress of trajectories and allow free energy calculations (the latter generally via replica exchange simulations (54)). Such analysis was applied successfully, for example, to *S. aureus* protein A using the C_α root mean squared deviation (C_α-RMSD) and Q (the fraction of native contacts formed) as reaction coordinates (55), and to villin headpiece using the RMSDs of two fragments to the native state as reaction coordinates (26, 27). Inspection of the projected villin free energy landscape (in implicit solvent) revealed a single main pathway to folding with a clearly defined barrier separating the folded and unfolded states, as well as an off-pathway trap conformation with no reasonably accessible direct path to the native state (27). Compatible results were observed by tracking the progress of several folding trajectories through the same projected coordinate space (26).

The utility of the reduced coordinate approach is completely reliant upon the ability of the chosen coordinates to separate the relevant occupied conformations (and their transition states). An example of the failure of such an approach is shown in Fig. 3a. The "opening" transition presumed in (30) to be the rate-limiting step in villin folding (see above) involves backtracking over completely unrelated portions of conformational space in the 2-D projection; furthermore, conformations on either side of the transition state are superimposed on each other. Such difficulties may be circumvented by using trajectory-driven methods to identify the projection space, such as principal component analysis (56, 57) or non-metric data scaling (58). Application of the latter to villin headpiece folding is shown in Fig. 3b, providing improved separation of the transition state ensemble and structures to either side of it.

Another frequently used method in the analysis of protein folding trajectories is conformational clustering (*e.g.*, (59, 60)), in which configurations occurring during a folding trajectory (or set of trajectories) are binned into related groups (clusters) based on a metric such as pairwise RMSDs between them, or the rate of interconversion between conformations (for comparison see (61)). Clustering analysis immediately highlights frequently occupied conformations, and tracking the cluster identity of the protein throughout a trajectory can provide a useful birds-eye view of the path followed during the simulation. Clustering can also be applied in several types of quantitative analysis which aid in the understanding of protein folding trajectories, particularly when information from a large array of simulations must be combined. In such cases it has proven useful to cluster the conformations present and then use the statistics obtained on their interconversion to develop a Markov state model, allowing evaluation of a variety of properties such as mean folding times dependent upon events far longer than the simulations used in constructing the model (25, 62, 63). The primary weakness of such models is, of course, that they are still vulnerable to undersampling in that any transitions or conformations which were not observed in the parameterization simulations, but are actually present, will not be accounted for.

The number of transitions observed between clustered conformations can also be used in the construction of a cut-based free energy profile (64, 65), in which clusters are partitioned into two disjoint sets in a way that minimizes the partition function of the barrier between the sets; such partitions are calculated along a reaction coordinate such as the fraction of the overall sampling weight that is in the same set as some arbitrary node (for example, the native state of a protein). Applied to folding simulations, such a profile allows the identification of the transition state ensemble (66, 67) for transitions of interest noted during the folding process. Crucially, the cut-based approach does not require *a priori* assignment of reaction coordinate(s), but equilibrium sampling of the conformational transitions of interest (which is currently difficult to obtain for most folding model systems) is needed. Once key conformations have been identified (*e.g.*, through clustering analysis), the transitions between them may also be investigated in more detail through application of methods such as transition path sampling (15) and subsequent analysis to optimize the definition of a reaction coordinate and transition state ensemble for a given transition (68).

Force field development

Molecular dynamics simulations utilize force fields to describe the potential energy of atomic systems as a function of their spatial arrangement. The functional form of classical force fields is divided into two sets of terms: bonded, also called internal, and nonbonded contributions. Bonded contributions include bond, angle and dihedral terms that represent interactions between covalently bonded atoms using harmonic potentials. The harmonic potentials are a coarse but rapidly computed approximation of Morse potentials describing bonded interactions. Perhaps the greatest disadvantage of the harmonic approximation is its inability to permit bonds between atoms to change, allowing descriptions of chemical reactions; however, the harmonic potential does permit all-atom simulations three to four orders of magnitude faster than methods allowing changes in electronic structure. Nonbonded terms include pair-wise Coulombic potentials describing electrostatics, and the Lennard-Jones (LJ) 6-12 potential that represents attractive van der Waals dispersion interactions and core-core repulsion between atom pairs.

Of the classical force fields, the most frequently used in all-atom MD simulations of protein folding are AMBER (69) and CHARMM (70). The bonded terms of AMBER and CHARMM are relatively similar (as are the equivalent terms in most other classical force fields); both utilize harmonic approximations for bonded interactions, parameterized through a combination of high-level quantum mechanical calculations and spectroscopic data on model compounds. However, fundamental differences exist in how their nonbonded terms, and particularly their atomic charges, are empirically parameterized. In the CHARMM family of force fields, an atom's charge is determined by fitting the effective interaction of polar groups with a TIP3P water molecule to quantum mechanical data, whereas atomic charges in recent AMBER force fields are determined by optimizing the reproduction of the electrostatic potential around the molecule of interest, subject to restraints to remove the possibility of physically absurd charge distributions (71). Both force fields suffer from a lack of polarizability, relying upon static atomic charges to model electrostatic contributions to protein dynamics which are often intrinsically coupled to a protein's internal and external electrostatic field.

Molecular modeling force fields have been under development for decades, and modern force fields consistently yield values for properties such as free energies of hydration for model compounds within 1-2 kcal/mol of experimental values (72, 73), and provide sub-Å C_α-RMSDs to known structures in simulations of folded proteins (74). Despite the generally excellent agreement between experimental and calculated properties for small model systems and folded proteins, some shortcomings are known to remain, such as the tendency of modern

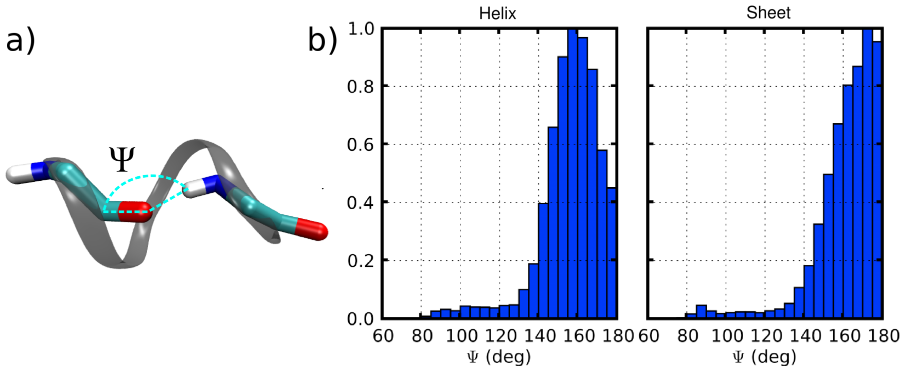

Figure 4: Directionality of hydrogen bonding in folding simulations. a) Illustration of the hydrogen-acceptor-acceptor antecedent angle Ψ in a protein backbone hydrogen bond. b) Normalized histogram of Ψ angles present in MD simulations of a misfolded helical state (Helix) or the native state (Sheet) of the WW domain (39). A survey of the PDB indicated that both should peak between 155 and 160 degrees (80). Part (b) reprinted from supplementary material of Biophysical Journal 96; Peter L. Freddolino, Sanghyun Park, Benoît Roux, and Klaus Schulten; Force field bias in protein folding simulations; 3772–3780; Copyright 2009, with permission from Elsevier.

pairwise additive force fields to overestimate the strength of solute-solute interactions (75). In addition, several recent studies have shown inaccuracies related to the thermodynamic equilibria between different protein secondary structures, including both direct attempts to fold proteins through MD simulations (39) and more general studies of the accuracy with which MD force fields represent proteins (76, 77). As simulations long enough to allow large scale structural transitions such as secondary structure rearrangements only recently became commonplace, for most of their history molecular dynamics force fields have only needed to provide a realistic description of a protein within the neighborhood of a known starting state. With modern computing capabilities, however, another round of modifications and improvements to molecular modeling force fields is clearly required to maintain an accurate description of the simulated systems.

Currently existing classical force fields have undergone many rounds of iterative improvement in which parameters were tuned to provide better agreement with experimental or quantum mechanical data. Over the past few years new sets of corrections for backbone parameters have been applied both to the AMBER (78) and CHARMM (79) families of force fields in order to bring the potential energy surface around protein backbone torsions into better agreement with quantum mechanical data. The changes made to CHARMM were particularly far-reaching: a new cross term (CMAP) was added to the force field, involving addition of a correction based on the ϕ and ψ angles of a given amino acid to bring their energetic contribution into direct agreement with two-dimensional maps of the potential energy surface obtained from high-level quantum mechanical calculations. While the recent backbone corrections would be expected to substantially improve the secondary structure propensities of force fields, problems with the treatment of both small model systems (77) and folding proteins (39) were observed (using AMBER and CHARMM force fields, respectively) even with the corrections in place. In addition, even where further corrections were applied to the backbone dihedrals of AMBER family force fields in order to correct their α helical propensity, both the entropy and enthalpy of helix formation were found to be underestimated (such that the errors canceled out at the temperature at which parameterization was performed) (77).

While a number of recent efforts to improve protein force fields have focused on the parameters for bonded terms, secondary structure elements (particularly β sheets) are inherently non-local, relying in large part on the behavior of hydrogen bonding. The most commonly used force fields in modern molecular dynamics simulations treat hydrogen bonding simply as an interaction between point charges, but hydrogen bonding in fact has a strong directional dependence that is apparent both from quantum mechanical calculations on model compounds and in crystal structures of proteins (80, 81). Molecular modeling force fields incorporating directional hydrogen bonding have frequently shown improved accuracy (80, 82, 83). Analysis of the hydrogen bonding geometries present in recent folding simulations of the WW domain using CHARMM22/CMAP (see Fig. 4) showed that while the (erroneously favored) α helical structures possessed a distribution of hydrogen bonding geometries matching those from quantum mechanical calculations, the simulated crystal-like β sheet structure overpopulated linear hydrogen bonding geometries, reflecting an artificial energetic frustration introduced by the simplistic representation of hydrogen bonding. Likewise, the errors in ΔU and ΔS observed by Best and Hummer during α helix formation are consistent with a lack of proper hydrogen bonding treatment: directional hydrogen bonds would be stronger but lead to a more negative ΔS during helix formation due to the imposed orientation (77). Atomic polarizability, which is neglected in classical force fields, has also been shown to play a significant role in the energetics of α helix formation (84).

Thus, while tuning of bonded parameters continues to be a valuable tool in refining molecular dynamics force fields, more dramatic changes are likely necessary to correct problems currently hampering molecular dynamics simulations of folding. Hydrogen bonding orientation may be included through the addition of explicit hydrogen bonding terms (82) or “lone pair” charge sites maintained at a specific geometry relative to atomic centers (85). Treatment of atomic polarizability is more challenging; several solutions exist in recently developed force fields, including the replacement of point charges with partially polarizable multipole expansions (86), models allowing charge to flow between atoms in response to the electric field (87, 88), and Drude oscillator models in which the charge of specific heavy atoms is partially placed on a very light independent particle coupled to the parent atom by a strong spring (89, 90). In light of the recent simulation results discussed above, it appears likely that the use of some polarizable force field also incorporating explicit hydrogen bonding or off-site lone pairs is essential for protein folding in MD simulations.

While it is easy to become focused on refinements to solute parameters, the protein-protein interactions in folding simulations occur neither in a metaphorical nor a literal vacuum, but instead exist in competition with protein-water and water-water interactions. The treatment of water, either implicit or explicit, and the interactions of the protein with water are thus extremely important to obtaining proper conformational equilibria during such simulations. Despite the added computational expense, we strongly advocate the use of explicit solvent models in protein folding simulations, as implicit solvent models have been shown to be unable to reproduce the relative free energies for folding intermediates obtained using explicit solvent (12, 91), and by their nature cannot capture details such as buried waters which are known to be important even in the case of such simple proteins as Trpcage (14). A recent survey of the thermodynamics of hydration for model compounds related to amino acids suggested that the properties considered (ΔG, ΔH, TΔS, and ΔC_p) are much more dependent on the choice of water force field than on the protein force field (92). Molecular dynamics water models are generally parameterized primarily to reproduce bulk water properties; unfortunately, the accuracy of representation of water properties in such a model is not well correlated with its accuracy in combination with even simple solutes (92). The issue of water model choice is complicated by the fact that protein force fields are generally parameterized and tested using a specific model (most commonly, for the current generation of classical force fields, TIP3P (93)), and thus one cannot simply switch to a new water model even if it has been shown to have superior properties. At present a new generation of water models is under active development for use with polarizable force fields (*e.g.*, (90, 94, 95)); optimal performance of the associated polarizable protein force fields may also require simultaneous refinement of solvent and solute parameters.

Outlook

Molecular dynamics simulations of protein folding can be a tremendously useful tool, providing otherwise inaccessible data that aid the interpretation and testing of protein folding mechanisms. Such simulations face serious challenges, both from the sheer amount of sampling required to adequately model protein folding and the fidelity with which empirical force fields must represent the true free energy surface on which a protein folds. Both challenges can be met, the former through new technologies to improve sampling and improved analysis methods to make more constructive use of the obtained data, and the latter through the use of new force fields explicitly incorporating hydrogen bonding and atomic polarizability. Even for the simple systems reviewed in the present article, much work remains to be done in terms of experimental validation of recent predictions made by MD simulations. In addition, even as new force fields are being developed, it may be possible to expand to the study of slightly larger and more complicated proteins such as the λ-repressor (Fig. 1d), a five-helix bundle with variants folding in 2-15 μs (96), so long as judicious choices are made to target well-studied proteins with secondary structure elements that are expected to be treated as accurately as possible by existing force fields.

Acknowledgements

The field of molecular dynamics simulations of protein folding is so broad that it is impossible here to discuss the work of all major contributors, and thus we apologize to the (many) excellent investigators whose work could not be included due to space constraints.

The authors are supported by grant P41-RR005969 from the National Institutes of Health. The authors gladly acknowledge supercomputer time provided by the National Center for Supercomputing Applications via Large Resources Allocation Committee grant MCA93S028. P.L.F. thanks Dr. Aruna Rajan for assistance in figure preparation.

References

[1] Alder, B. J. & Wainwright, T. E. Studies in molecular dynamics. I. general method. *J. Chem. Phys.* **31**, 459 (1959).

[2] Adcock, S. A. & McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615 (2006).

[3] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. & Tramontano, A. Critical assessment of methods of protein structure prediction - round VIII. *Proteins* **77 Suppl 9**, 1–4 (2009).

[4] Das, R. *et al.* Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins* **69 Suppl 8**, 118–128 (2007).

[5] Hubner, I. A., Deeds, E. J. & Shakhnovich, E. I. Understanding ensemble protein folding at atomic detail. *Proc. Natl. Acad. Sci. USA* **103**, 17747–17752 (2006).

[6] Yang, J. S., Wallin, S. & Shakhnovich, E. I. Universality and diversity of folding mechanics for three-helix bundle proteins. *Proc. Natl. Acad. Sci. USA* **105**, 895–900 (2008).

[7] Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Biol.* **9**, 425–430 (2002).

[8] Simmerling, C., Strockbine, B. & Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **124**, 11258–11259 (2002).

[9] Snow, C. D., Zagrovic, B. & Pande, V. S. The trp cage: folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **124**, 14548–14549 (2002).

[10] Chowdhury, S., Lee, M. C., Xiong, G. & Duan, Y. Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* **327**, 711–717 (2003).

[11] Pitera, J. W. & Swope, W. Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. *Proc. Natl. Acad. Sci. USA* **100**, 7587–7592 (2003).

[12] Zhou, R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* **53**, 148–161 (2003).

[13] Paschek, D., Hempel, S. & Garca, A. E. Computing the stability diagram of the Trp-cage miniprotein. *Proc. Natl. Acad. Sci. USA* **105**, 17754–17759 (2008).

[14] Paschek, D., Nymeyer, H. & Garca, A. E. Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol.* **157**, 524–533 (2007).

[15] Juraszek, J. & Bolhuis, P. G. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. USA* **103**, 15859–15864 (2006).

[16] Juraszek, J. & Bolhuis, P. G. Rate constant and reaction coordinate of Trp-cage folding in explicit water. *Biophys. J.* **95**, 4246–4257 (2008).

[17] McKnight, C. J., Doering, D. S., Matsudaira, P. T. & Kim, P. S. A thermostable 35-residue subdomain within villin headpiece. *J. Mol. Biol.* **260**, 126–134 (1996).

[18] Kubelka, J., Eaton, W. A. & Hofrichter, J. Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.* **329**, 625–630 (2003).

[19] Bunagan, M. R., Gao, J., Kelly, J. W. & Gai, F. Probing the folding transition state structure of the villin headpiece subdomain via side chain and backbone mutagenesis. *J. Am. Chem. Soc.* **131**, 7470–7476 (2009).

[20] Kubelka, J., Chiu, T. K., Davies, D. R., Eaton, W. A. & Hofrichter, J. Sub-microsecond protein folding. *J. Mol. Biol.* **359**, 546–553 (2006).

[21] Buscaglia, M., Kubelka, J., Eaton, W. A. & Hofrichter, J. Determination of ultrafast protein folding rates from loop formation dynamics. *J. Mol. Biol.* **347**, 657–664 (2005).

[22] Wang, M. *et al.* Dynamic NMR line-shape analysis demonstrates that the villin headpiece subdomain folds on the microsecond time scale. *J. Am. Chem. Soc.* **125**, 6032–6033 (2003).

[23] Duan, Y. & Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744 (1998).

[24] Zagrovic, B., Snow, C. D., Shirts, M. R. & Pande, V. S. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323**, 927–937 (2002).

[25] Jayachandran, G., Vishal, V. & Pande, V. S. Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J. Chem. Phys.* **124**, 164902 (2006).

[26] Lei, H. & Duan, Y. Two-stage folding of HP-35 from ab initio simulations. *J. Mol. Biol.* **370**, 196–206 (2007).

[27] Lei, H., Wu, C., Liu, H. & Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **104**, 4925–4930 (2007).

[28] Lei, H., Deng, X., Wang, Z. & Duan, Y. The fast-folding HP35 double mutant has a substantially reduced primary folding free energy barrier. *J. Chem. Phys.* **129**, 155104–7 (2008).

[29] Ensign, D. L., Kasson, P. M. & Pande, V. S. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* **374**, 806–816 (2007).

[30] Freddolino, P. L. & Schulten, K. Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.* **97**, 2338–2347 (2009).

[31] Hu, K.-N., Yau, W.-M. & Tycko, R. Detection of a transient intermediate in a rapid protein folding process by solid-state nuclear magnetic resonance. *J. Am. Chem. Soc.* **132**, 24–25 (2010).

[32] Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75 (2004).

[33] Jäger, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The folding mechanism of a beta-sheet: the ww domain. *J. Mol. Biol.* **311**, 373–393 (2001).

[34] Jäger, M. *et al.* Structure-function-folding relationship in a WW domain. *Proc. Natl. Acad. Sci. U S A* **103**, 10648–10653 (2006).

[35] Cecconi, F., Guardiani, C. & Livi, R. Testing simplified proteins models of the hPin1 WW domain. *Biophys. J.* **91**, 694–704 (2006).

[36] Karanicolas, J. & Brooks, C. L. Improved Go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* **334**, 309–325 (2003).

[37] Luo, Z., Ding, J. & Zhou, Y. Temperature-Dependent Folding Pathways of Pin1 WW Domain: An All-Atom Molecular Dynamics Simulation of a Go Model. *Biophys. J.* **93**, 2152–2161 (2007).

[38] Freddolino, P. L., Liu, F., Gruebele, M. & Schulten, K. Ten-microsecond MD simulation of a fast-folding WW domain. *Biophys. J.* **94**, L75–L77 (2008).

[39] Freddolino, P. L., Park, S., Roux, B. & Schulten, K. Force field bias in protein folding simulations. *Biophys. J.* **96**, 3772–3780 (2009).

[40] Ensign, D. L. & Pande, V. S. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. *Biophys. J.* **96**, L53–L55 (2009).

[41] Patel, S. & Brooks, C. L. Fluctuating charge force fields: recent developments and applications from small molecules to macromolecular biological systems. *Molecular Simulation* **32**, 231–249 (2006).

[42] Harder, E. *et al.* Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theor. Comput.* **2**, 1587–1597 (2006).

[43] Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comput.* **4**, 435–447 (2008).

[44] Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)* (Tampa, Florida, 2006).

[45] Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S. & Pande, V. S. Folding@home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel&Distributed Processing* (Rome, Italy, 2009).

[46] Shaw, D. E. *et al.* Millisecond-scale molecular dynamics simulations on Anton. In *SC '09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, 1–11 (New York, NY, USA, 2009).

[47] Stone, J. E. *et al.* Accelerating molecular modeling applications with graphics processors. *J. Comp. Chem.* **28**, 2618–2640 (2007).

[48] Elsen, E. *et al.* N-body simulation on GPUs. In *SC06 Proceedings* (IEEE Computer Society, 2006).

[49] Ufimtsev, I. S. & Martinez, T. J. Quantum chemistry on graphical processing units. 3. analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theor. Comput.* **5**, 2619–2628 (2009).

[50] Harvey, M. J., Giupponi, G. & Fabritiis, G. D. Acemd: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theor. Comput.* **5**, 1632–1639 (2009).

[51] Friedrichs, M. S. *et al.* Accelerating molecular dynamic simulation on graphics processing units. *J. Comp. Chem.* **30**, 864–872 (2009).

[52] Anderson, J. A., Lorenz, C. D. & Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Chem. Phys.* **227**, 5342–5359 (2008).

[53] Phillips, J. C., Stone, J. E. & Schulten, K. Adapting a message-driven parallel application to GPU-accelerated clusters. In *SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing* (IEEE Press, Piscataway, NJ, USA, 2008).

[54] Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).

[55] García, A. E. & Onuchic, J. N. Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. *Proc. Natl. Acad. Sci. USA* **100**, 13898–13903 (2003).

[56] Ichiye, T. & Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Struct., Func., Gen.* **11**, 205–217 (1991).

[57] García, A. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68**, 2696–2699 (1992).

[58] Rajan, A., Freddolino, P. L. & Schulten, K. Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PLoS One* (2010). In press.

[59] Karpen, M. E., Tobias, D. J. & Brooks, C. L. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* **32**, 412–420 (1993).

[60] Daura, X. *et al.* Peptide folding: When simulation meets experiment. *Angewandte Chemie International Edition* **38**, 236–240 (1999).

[61] Keller, B., Daura, X. & van Gunsteren, W. F. Comparing geometric and kinetic cluster algorithms for molecular simulation data. *J. Chem. Phys.* **132**, 074110–16 (2010).

[62] Singhal, N., Snow, C. D. & Pande, V. S. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **121**, 415–425 (2004).

[63] Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A. & Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007).

[64] Krivov, S. V. & Karplus, M. One-dimensional free-energy profiles of complex systems: progress variables that preserve the barriers. *J. Phys. Chem. B* **110**, 12689–12698 (2006).

[65] Krivov, S. V. & Karplus, M. Diffusive reaction dynamics on invariant free energy profiles. *Proc. Natl. Acad. Sci. USA* **105**, 13841–13846 (2008).

[66] Krivov, S. V., Muff, S., Caflisch, A. & Karplus, M. One-dimensional barrier-preserving free-energy projections of a beta-sheet miniprotein: new insights into the folding process. *J. Phys. Chem. B* **112**, 8701–8714 (2008).

[67] Muff, S. & Caflisch, A. Identification of the protein folding transition state from molecular dynamics trajectories. *J. Chem. Phys.* **130**, 125104 (2009).

[68] Best, R. B. & Hummer, G. Reaction coordinates and rates from transition paths. *Proc. Natl. Acad. Sci. USA* **102**, 6732–6737 (2005).

[69] Case, D. A. *et al. AMBER 10.* University of California, San Francisco (2008).

[70] MacKerell, A. D., Jr. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[71] Baily, C., Cieplak, P., Cornell, W. & Kollman, P. A well-behaved electrostatic potential-based method using charge restraints for deriving atomic charges: The RESP model. *J. Phys. Chem.* **100**, 10269–10280 (1993).

[72] Shirts, M. R. & Pande, V. S. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **122**, 134508 (2005).

[73] Deng, Y. & Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **113**, 2234–2246 (2009).

[74] Buck, M., Bouguet-Bonnet, S., Pastor, R. W. & MacKerell, A. D. Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme. *Biophys. J.* **90**, L36–38 (2006).

[75] Chen, J., Brooks, C. L. & Scheraga, H. A. Revisiting the carboxylic acid dimers in aqueous solution: interplay of hydrogen bonding, hydrophobic interactions, and entropy. *J. Phys. Chem. B* **112**, 242–249 (2008).

[76] Best, R. B., Buchete, N.-V. & Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* L07–L09 (2008).

[77] Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113**, 9004–9015 (2009).

[78] Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).

[79] MacKerell Jr., A. D., Feig, M. & Brooks III, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comp. Chem.* **25**, 1400–1415 (2004).

[80] Kortemme, T., Morozov, A. V. & Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **326**, 1239–1259 (2003).

[81] Morozov, A. V., Kortemme, T., Tsemekhman, K. & Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA* **101**, 6946–6951 (2004).

[82] Lii, J.-H. & Allinger, N. L. Directional hydrogen bonding in the MM3 force field: II. *J. Comp. Chem.* **19**, 1001–1016 (1998).

[83] Fabiola, F., Bertram, R., Korostelev, A. & Chapman, M. S. An improved hydrogen bond potential: impact on medium resolution protein structures. *Protein Sci.* **11**, 1415–1423 (2002).

[84] Morozov, A. V., Tsemekhman, K. & Baker, D. Electron density redistribution accounts for half the cooperativity of alpha helix formation. *J. Phys. Chem. B* **110**, 4503–4505 (2006).

[85] Cieplak, P., Caldwell, J. & Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comp. Chem.* **22**, 1048–1057 (2001).

[86] Ponder, J. W. *et al.* Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114**, 2549–2564 (2010).

[87] Patel, S. & III, C. L. B. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comp. Chem.* **25**, 1–16 (2004).

[88] Archambault, F. *et al.* Polarizable intermolecular potentials for water and benzene interacting with halide and metal ions. *J. Chem. Theor. Comp.* **5**, 3022–3031 (2009).

[89] Yu, H. & van Gunsteren, W. F. Charge-on-spring polarizable water models revisited: from water clusters to liquid water to ice. *J. Chem. Phys.* **121**, 9549–9564 (2004).

[90] Lamoureux, G. & Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.* **119**, 3025–3039 (2003).

[91] Zhou, R. & Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a beta -hairpin folding in water? *Proc. Natl. Acad. Sci. USA* **99**, 12777–12782 (2002).

[92] Hess, B. & van der Vegt, N. F. A. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **110**, 17616–17626 (2006).

[93] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

[94] Lamoureux, G., Harder, E., Vorobyov, I. V., Roux, B. & MacKerell Jr., A. D. A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* **418**, 245–249 (2006).

[95] Ren, P. & Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **107**, 5933–5947 (2003).

[96] Liu, F., Gao, Y. G. & Gruebele, M. A survey of lambda repressor fragments from two-state to downhill folding. *J. Mol. Biol.* **397**, 789-798 (2010).

[97] Frishman, D. & Argos, P. Knowledge-based secondary structure assignment. *Proteins* **23**, 566–579 (1995).

[98] Eastwood, M. P., Hardin, C., Luthey-Schulten, Z. & Wolynes, P. G. Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J. Res. Dev.* **45**, 475–497 (2001).