Blind Test of Physics-Based Prediction of Protein Structures

M. Scott Shell,[†*] S. Banu Ozkan,[§] Vincent Voelz,[‡] Guohong Albert Wu,[†] and Ken A. Dill[†]
[†]Department of Pharmaceutical Chemistry, and [‡]Graduate Group in Biophysics, University of California, San Francisco, California; and [§]Department of Physics, Arizona State University, Tempe, Arizona

ABSTRACT We report here a multiprotein blind test of a computer method to predict native protein structures based solely on an all-atom physics-based force field. We use the AMBER 96 potential function with an implicit (GB/SA) model of solvation, combined with replica-exchange molecular-dynamics simulations. Coarse conformational sampling is performed using the zipping and assembly method (ZAM), an approach that is designed to mimic the putative physical routes of protein folding. ZAM was applied to the folding of six proteins, from 76 to 112 monomers in length, in CASP7, a community-wide blind test of protein structure prediction. Because these predictions have about the same level of accuracy as typical bioinformatics methods, and do not utilize information from databases of known native structures, this work opens up the possibility of predicting the structures of membrane proteins, synthetic peptides, or other foldable polymers, for which there is little prior knowledge of native structures. This approach may also be useful for predicting physical protein folding routes, non-native conformations, and other physical properties from amino acid sequences.

INTRODUCTION

In the past 15 years, investigators have made major advances in computer-based predictions of the native structures of small proteins (1,2). This enterprise is important for efficiently converting genome information to knowledge of protein structures and mechanisms. However, current methods are largely bioinformatics-based; their inference engines draw heavily on the Protein Data Bank (PDB), a large collection of known native protein structures. Indeed, a key motivation for structural genomics initiatives in recent years is to grow such databases for the purpose of structure prediction.

Although bioinformatics methods have demonstrated great success in protein structure prediction in recent years, it would ultimately be advantageous to use purely physical principles to predict structures and folding routes, given only a protein's amino acid sequence, without knowledge of its native structure. In contrast to bioinformatics methods, physics-based approaches draw their inferences largely from the physicochemical properties of atoms and small molecules, and not from prior knowledge of native structures. Physics-based methods could be used to study molecules for which there are as yet no structural databases, such as membrane proteins, which are important pharmaceutical targets; synthetic polypeptides with D-amino acids; or foldamers with nonbiological backbones. Physical approaches would also offer the potential to explore how folding processes depend on denaturants and stabilizers, pH, salts, temperature, and mutations. And, they could explore the dynamics important to biological function, including folding mechanisms, misfolding, aggregation, conformational transitions, and induced-fit binding.

However, up to now, purely physical approaches have not been practical. An important component of physical modeling has been an ongoing, decades-old effort to develop accurate classical atomic force fields for polypeptides, such as AMBER (3), CHARMM (4), OPLS (5), and GROMOS (6). Most current protein structure prediction methods involve some level of hybridization, incorporating force-field components into bioinformatics predictions either by training potential energy functions on known structures or by adding physics-inspired terms to informatics scoring functions. However, such hybrid methods can miss some of the advantages of purely physical methods. They are not intended to capture the true molecular-thermodynamic entropies and free energies of folding, or to predict how a protein structure will change under different solution conditions. Moreover, database-derived models are potentially less transferable to cases that are far outside the training sets upon which they are parameterized, as might occur with membrane proteins or synthetic amino acids.

There have been two barriers to applying all-atom, physics-based force fields to the prediction of protein native structures. First, to satisfy the Newtonian equations of motions, femtosecond time steps are required. To reach typical millisecond folding times requires computational resources that are substantially beyond all but the largest supercomputing resources. At present, key projects using dedicated supercomputers such as IBM's Blue Gene [http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html] or distributed-grid computing methods such as Folding@Home [http://folding.stanford.edu/] can invest thousands to

Submitted September 12, 2008, and accepted for publication November 5, 2008.

*Correspondence: shell@engineering.ucsb.edu

M. Scott Shell's present address is Department of Chemical Engineering, University of California, Santa Barbara, California.

Vincent Voelz's present address is Department of Chemistry, Stanford University, Palo Alto, California.

G. Albert Wu's present address is Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California.

Editor: Nathan Andrew Baker.

0006-3495/09/02/0917/8 $2.00

doi: 10.1016/j.bpj.2008.11.009

tens of thousands of CPU years in folding a single protein that is smaller than ~70 monomers long. Second, because of these limitations in conformational sampling, it has been difficult to carry out sufficient testing of the force fields to learn whether they are accurate enough for protein folding.

Over the past 15 years, a standard has emerged that protein structure predictions should be blind-tested in a biennial community-wide event called Critical Assessment of Techniques for Protein Structure Prediction (CASP) (2). In this context, purely physics-based approaches have traditionally been regarded as not competitive (7,8), and it has been noted that in the attempt to refine poor structures, ''energy minimization or molecular dynamics generally leads to a model that is less like the experimental structure'' (7). In early CASPs, several physics-based folding algorithms were used, but more recently the physical methods ''largely have been displaced'' (2). In addition, various studies have suggested that there are some problems, such as imbalances between helices and sheets, and inaccurate ion-pairing interactions, with common molecular mechanics force fields (9,10) and/or their companion implicit solvation models (11,12).

On the other hand, there is also some evidence that purely physical force fields, when properly paired with solvation models, may now be adequate for reaching native structures, given sufficient conformational sampling. Duan and Kollman (13) performed a milestone microsecond molecular dynamics (MD) simulation of the 36-residue villin headpiece in explicit solvent starting from an unfolded conformation, and the most stable configurational cluster reached 5.7 Å from the NMR structure. Scheraga et al. (14) folded a 46-residue protein A fragment to 3.5 Å root mean-square deviation (RMSD) using a modified Monte Carlo sampling algorithm with an implicit solvation model. High-resolution structures of villin were recently reached by Pande et al. (15) and Duan et al. (16,17). In addition, three groups (Simmerling et al. (18), the IBM Blue Gene group of Pitera and Swope (19), and Duan et al. (20)) have folded the 20-residue Trp-cage peptide to ~1 Å. Recently, Lei and Duan (21) folded the albumin binding domain, a 47-residue, three-helix bundle, to 2.0 Å. These successes indicate that current state-of-the-art force fields may be useful for protein folding, but so far no such method has been tested using a single protocol on multiple molecules or proteins much larger than 50 mers, or in blind tests.

In this study, we performed such a test. We attempted the folding of six proteins in CASP7, of chain lengths up to 112 mers, with an off-the-shelf all-atom physical model, AMBER 96, combined with the generalized Born/surface area (GBSA) implicit solvation model of Onufriev, Bashford, and Case (OBC; AMBER option ''igb=5'') (22,23). To surmount the tremendous computational sampling barriers, we used a technique that accelerates folding according to a putative folding mechanism, called zipping and assembly (ZA). In a previous in-house test on proteins whose structures we knew in advance, the ZA-based approach successfully folded eight small, single domain proteins to better than 3 Å accuracy (24). Here we tested that method in a different setting that offers new perspectives. In this work, native structures were not known a priori, the predictions were time-limited to roughly 1 month, and the proteins studied are typical of the kinds of targets currently produced by structural genomics efforts and of practical interest for structure prediction.

We also believe the work presented here provides new insight into the role that physics-based methods might play in structure prediction. Of the more than 250 research groups that participated in CASP7, we are not aware of any, besides ours, that used purely physics-based methods, according to the strict criteria that scoring is based entirely on an all-atom force field with equilibrium sampling without relying on templates, database-derived potentials, or secondary structure predictions. Close in spirit to our work is that of Scheraga et al. (25,26), who have pioneered physical methods for protein structure prediction. However, in CASP7, Scheraga et al. (27) derived their coarse-grained potentials from a combination of an all-atom force field and a training database. Our goal here, instead, was to test a stricter physical strategy.

In this work, we describe both the methods used in our folding simulations and the results of our predictions in CASP7. Although our predictions are not as good as the best bioinformatics methods, we find that purely physics-based methods do surprisingly well on three counts. First, our predictions are roughly on par with the average performance of bioinformatics methods in CASP7. Second, the predictions exceed expectations for modern force fields in predicting correct secondary structures and basic topologies, which to our knowledge have never been tested on proteins of this size before. Third, although our simulations require significant computational overhead relative to bioinformatics methods, the ZA-based method folds ~100 mer proteins on commodity compute clusters and in a fraction of the time required by current physics-based supercomputer efforts, expanding the potential for all-atom physics-based methods to contribute to structure prediction.

MATERIALS AND METHODS

We use the AMBER ff96 force field with the GBSA implicit solvation model of Onufriev, Bashford, and Case (OBC) (22,23). We previously tested the combined ff96/OBC force field and solvent model to ascertain whether it predicts stability for known native structures in short peptides with extensive sampling (28). We sampled conformations using the replica exchange molecular dynamics (REMD) method that was pioneered by Sugita and Okamoto (29) and is now a standard protocol for peptide simulations. In our simulations, no input from databases of native structures, such as PDB templates, secondary structure webservers, or statistical potentials were used.

In addition to the ''fine-grained'' sampling afforded by REMD, this work was made possible by a new and very fast ''coarse-grained'' conformational sampling method called the ZA method (ZAM) (24). We refer to ZAM as a ''mechanism-based'' conformational sampling method because it is based on a model of how we believe proteins reach their native states so quickly and avoid so much conformational searching as they physically fold up (24).

In the ZA mechanism, it is postulated that a protein explores only a small fraction of its full conformational space to find its native state (30). On the earliest timescales (nanoseconds), small peptide segments within the chain independently adopt tentative locally metastable structures; the chain cannot sample more broadly on those timescales. These peptide structures can then nucleate additional structure locally by reeling in nearby sections of chain (zipping), or join together with other neighboring partially structured peptide pieces to form larger units (assembly). In this way, different degrees of freedom are active on different timescales, so the large global optimization problem of folding is accomplished by a hierarchical series of smaller local optimization problems. Evidence that ZA is a physical mechanism of folding includes consistency with experimental ϕ-values (24,31–33), the experimentally observed correlation between folding rate and native topology contact order (34,35), experiments on circularly permuted proteins (31), experiments on Φ values in small proteins in nonblinded all-atom modeling (24), and the folding speeds of small proteins that indicate their high search efficiencies (33). However, irrespective of whether ZAM actually mimics true physical folding routes, a key point of the work presented here is to show that it is nevertheless a highly efficient computer search algorithm for protein folding.

ZAM implements this putative folding mechanism within a computational search method as follows: First, ZAM breaks the full protein sequence into small, overlapping fragments. Each fragment is simulated separately. Then each fragment that is found to have an ensemble of metastable structures (see below) is grown by accreting additional residues, or two fragments nearby in sequence are simulated together and assembled, leading to structures that are larger and more stable. In growth, two new residues are added to each end of clustered structures from the parent fragment simulation. In assembly, fragments are put into a diverse set of rigid body orientations that have a compact hydrophobic core. To approach equilibrium conformational populations (conditional on the previously applied restraints), each step of growth or assembly is followed by REMD sampling.

A crucial part of this process is the identification of fragments that form stable hydrophobic contacts, assessed by computing metrics such as the potential of mean force and contact free energies. Fragments that do not form new hydrophobic contacts after growth or assembly, or that result in loss of contacts, are not pursued further. ZA routes are enforced by imposing a harmonic spring restraint to contacts that have been found stable in the previous time step, limiting the further sampling of those same degrees of freedom in later steps. In essence, these restraints guide a protein to zip or assemble along a particular pathway, driven by the physical interactions, by focusing the sampling mainly on the few new degrees of freedom that are added at each step.

ZAM pursues various folding routes in parallel and many different possible combinations of structured pieces generated along the way. When multiple full-chain predictions have been generated, ZAM allows these structures to ''compete'' with each other by seeding a single REMD simulation and examining which ones dominate at the target temperature. During these final steps, restraints are removed. However, when we were time-limited in CASP7, we skipped this final simulation by simply selecting structures with the most compact hydrophobic cores. Post-CASP, we performed such REMD simulations that confirmed that the selected models were indeed the most stable.

The specific procedure followed by ZAM for each target in CASP7 is as follows: The full protein is first parsed into overlapping 8-mer fragments spaced every three residues apart. Each such fragment is capped and begins in the extended state; it is then energy-minimized followed by 5 ns of REMD, in the absence of the rest of the chain. The REMD simulations span a temperature range of 270–600 K and consist of swaps attempted every 10 ps, for efficiency (later reduced to every 1 ps for larger fragments). Fifteen replicas are used for 8 mers.

After REMD, we retain those 8 mers that satisfy one of two criteria: either they have persistent backbone structure, or they exhibit cooperative behavior in that multiple locally stable residue-residue contacts tend to form simultaneously rather than independently. Backbone structure is quantified in two ways: 1), using a modified k-means clustering algorithm on the last 2 ns of the lowest temperature trajectory, and assessing the population of the dominant cluster; and 2), determining a coarse-grained backbone ''mesoentropy'' by assigning each residue to a helical, β, or loop configuration (36). Two residues are considered in contact when the distance between residue centroids is less than 8 A; contact cooperativity is assessed by the probability that a pair of contacts is formed simultaneously, averaged over the last 2 ns of the trajectory. The overall cooperativity of a particular 8 mer is computed from the average contact cooperativity over all residue contact pairs. For final selection of the 8 mers to retain, we collect all fragments that are within the top third scoring percentile in any of the three metrics: high cluster population, low mesoentropy, and high overall cooperativity. Our goal here is to cast a wide umbrella for fragments that exhibit possible zipping behavior, as evidenced by structural stability and cooperativity.

To those 8 mers that were retained, new chain is then added. They are grown into 12 mers, followed by REMD for another 5 ns. The process is repeated to reach partially structured 16 mers. In each case, the new chain is added in extended form to each of the clustered structures for a fragment. These structures are then minimized and used as initial starting configurations in each of the replicas of the new REMD simulation (structures are repeated as necessary to reach the target number of replicas). For both 12 mers and 16 mers, 20 replicas are used per fragment, with the same temperature range.

At the 16 mer level, stable contacts are identified within each fragment using the potential of mean force (PMF) versus distance for all possible residue pairs, computed by weighted histogram analysis from the last 2 ns of the simulation and at 270 K. We take the residue pairs for which the PMFs show a pronounced minimum in free energy at a distance less than 8.0 Å as favorable and stable (sampled at least 50% of the time). Any fragment that has mutually exclusive (i.e., ''competing'') stable contacts is split into separate ensembles in which that fragment has either of the two possible contacts. Our approach aims to capture a wide variety of potential zipping nucleation sites. Although this PMF-based heuristic for identifying residue-residue contacts does not originate from any molecular theory, it is a fairly inclusive criterion designed to maintain a broad sampling of potential folding pathways, rather than at this point to directly identify the native folding route.

To enforce any particular emerging folding route, a stable contact is locked into place by imposing a harmonic restraint between residue centroids with a force constant of 0.5 kcal/(mol Å^2). Fragments are then grown in the same manner as before by adding new residues in extended conformation at each terminus, followed by 5 ns REMD simulations. Thus, most of the new sampling focuses on the newly added residues, largely avoiding resampling the existing structure. This procedure is iterated until fragments cannot be grown further, that is, until no new stable contacts are found using the PMF heuristic.

When fragments cannot zip further, assembly of existing fragments that are neighboring in sequence is attempted, consistent with the ZA idea of local formation of superstructures from small prefolded sections of the chain. To do this, we generate a diverse distribution of rigid-body arrangements of two structured fragments in a way that promotes formation of new hydrophobic contacts. This is accomplished through Monte Carlo sampling that is used to identify fragment arrangements that minimize the hydrophobic radius of gyration (i.e., average radius of gyration of α-carbons in hydrophobic residues). The Monte Carlo procedure works as follows: The two fragments to be assembled are connected together with the missing loop residues. Simulated annealing is then performed using a potential energy function that includes steric repulsion and a term proportional to the number of hydrophobic residue-residue contacts. During the annealing, random perturbations are made to the ϕ-ψ=angles of each loop residue. The annealing is performed 100 times and we keep the top 10 structures (most hydrophobic contacts) with mutual RMSDs greater than 2 Å. This entire procedure is performed for every possible pair of clustered structures from the two fragment simulations. The final ensemble of conformations generated is then clustered and ranked by lowest hydrophobic radius of gyration. The

top-ranked structures are used as initial conformations, spread among the replicas, for another round of REMD simulations. This provides a fast way to sample possible topological assemblies while still using replica exchange to approach proper Boltzmann weights. Thus, the Monte Carlo rigid-body alignment is not intended to generate quantitatively accurate structures, but rather to merely generate a distribution of topologies with new hydrophobic contacts that serve as initial guesses for replica exchange simulations.

We continue the processes of growth and assembly until the full protein chain is completed. At that point, several different topologies may exist as the result of pursuing different mechanistic routes. One option then would be to evaluate a final REMD simulation in which these structures are used as initial conformations in the different replicas, with all the restraints removed. Alternatively, we have used a faster approach in time-limited cases whereby we rank the final structures by coarse metrics, such as hydrophobic radius of gyration, as a method to choose the best structure. For CASP, we chose the five topologies with the smallest hydrophobic radius of gyration as our submitted models. After the CASP competition was concluded, we performed long REMD simulations that showed that our submitted structures were indeed the most stable of those we generated.

RESULTS

In CASP7, we made predictions for six targets (T0283, T0309, T0311, T0335, T0358, and T0363) that were chosen because they are short (less than 120 amino acids) and hence within the capabilities of our computing resources. The proteins were also selected, when possible, to have little sequence similarity to existing proteins in the PDB, and hence where physical methods should ultimately be the most useful. Each target required an average of ~240 aggregate CPU months on a 2.4 GHz Xeon cluster, or slightly less than 1 month of real time on a 256-processor cluster. Although this time is long compared to the requirements of many of the bioinformatics-based algorithms in CASP, it is significantly shorter than that required for direct folding simulations of smaller proteins using supercomputing resources, such as the Folding@Home and Blue Gene projects, owing to the efficiency of the ZA mechanism. In fact, purely physics-based methods may have been uncommon in recent CASP competitions because of their computational demands; with a directed-sampling approach like ZAM, physical models can be more readily applied.

Fig. 1 provides a schematic of the ZA sequence of events for the CASP7 target T0358, an α/β protein of 87 residues. For clarity, this diagram of events shows only a subset of all of the pathways pursued in the ZAM simulation. It is evident that at intermediate stages in the procedure, the number of possible folding pathways (i.e., the number of ways grown fragments can assemble together) increases significantly. Moreover, though each fragment in the figure is represented by a single static structure, in reality ZAM

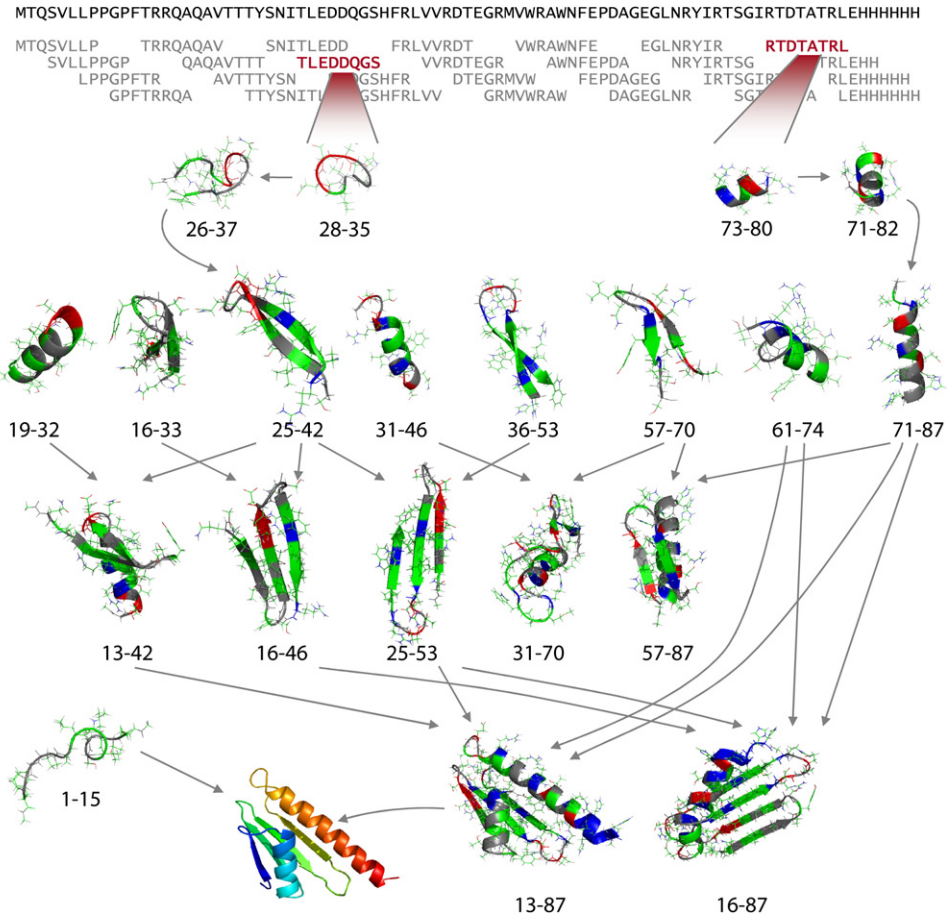

FIGURE 1 Folding routes found in the ZAM conformational search process for the CASP7 target T0358. Residue number ranges are shown below fragments. Only a subset of all steps and pathways is shown. ZAM begins by dividing the full chain into overlapping 8-mer fragments, spaced every three residues. These 8-mer fragments are grown over several stages to 16–20 mers; each stage involves adding new terminal residues to the structures followed by REMD sampling. Subsequently, secondary structure pieces neighboring in sequence are assembled together in various combinations using rigid-body alignment, followed by additional REMD sampling. The process continues along all possible pathways until a full fold is reached. Some assembly steps fail, resulting in loss of previous developed structure (31–70), whereas others cannot be assembled or grown so as to form new secondary structure later on (57–87)—those pathways are not pursued further. Along any one pathway, harmonic restraints are used to reinforce stable hydrophobic-hydrophobic contacts in the fragments and to avoid resampling existing structure. Colors in the fragments are as follows: green, hydrophobic; gray, polar; red, acidic; blue, basic.

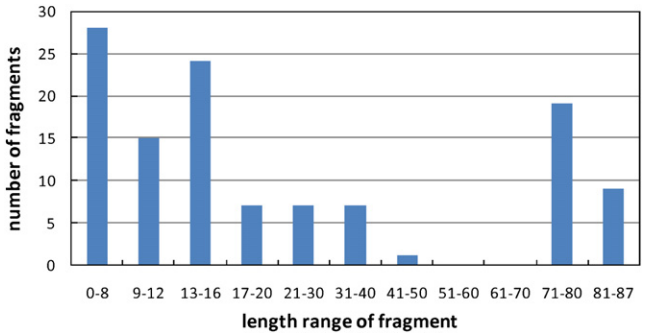

FIGURE 2 Number of fragments simulated during ZAM as a function of fragment length, for the target T0358. The high number of fragments at lengths of 70+ residues corresponds to the generation of different topologies from the assembly of two fragments of shorter lengths.

maintains an ensemble of conformations at each stage, extracted from runs using a clustering analysis. For this target, the pathway corresponding to the final structure produced is as follows: In the initial fragment stages, the N- and C-terminal helices and the two mid-chain hairpins form independently. Subsequently, the two hairpins are assembled into a single three-stranded sheet, to which the helices are then packed in various topologies. The most favorable topology survives the final stages of REMD simulation and our final selection criteria.

Fig. 2 provides a measure of the number of fragments simulated at each stage during the ZAM run for T0358. Many sub-20-residue fragments are simulated in the initial growth phases of the simulation, and a distribution of fragments of mid-range length are simulated during assembly. A large number of near-whole-chain fragments for this protein are explored at the final stages of the run, corresponding to the pursuit of different topologies assembled from the component helices and β-strands. These later stages require the bulk of the computational effort for this protein, as the size of the fragments greatly increases simulation cost relative to the initial short-fragment growth stages.

Fig. 3 shows the predictions for four of the six proteins that we attempted in CASP7. Each panel in Fig. 3 shows 1), an experimental native structure; 2), the best corresponding ZAM prediction; and 3, a CASP global distance test (GDT) plot comparing our five allowed predictions with the many predictions from the best current bioinformatics methods (253 teams participated in CASP7). Of the six structures that we attempted, two (targets T0309 and T0363)

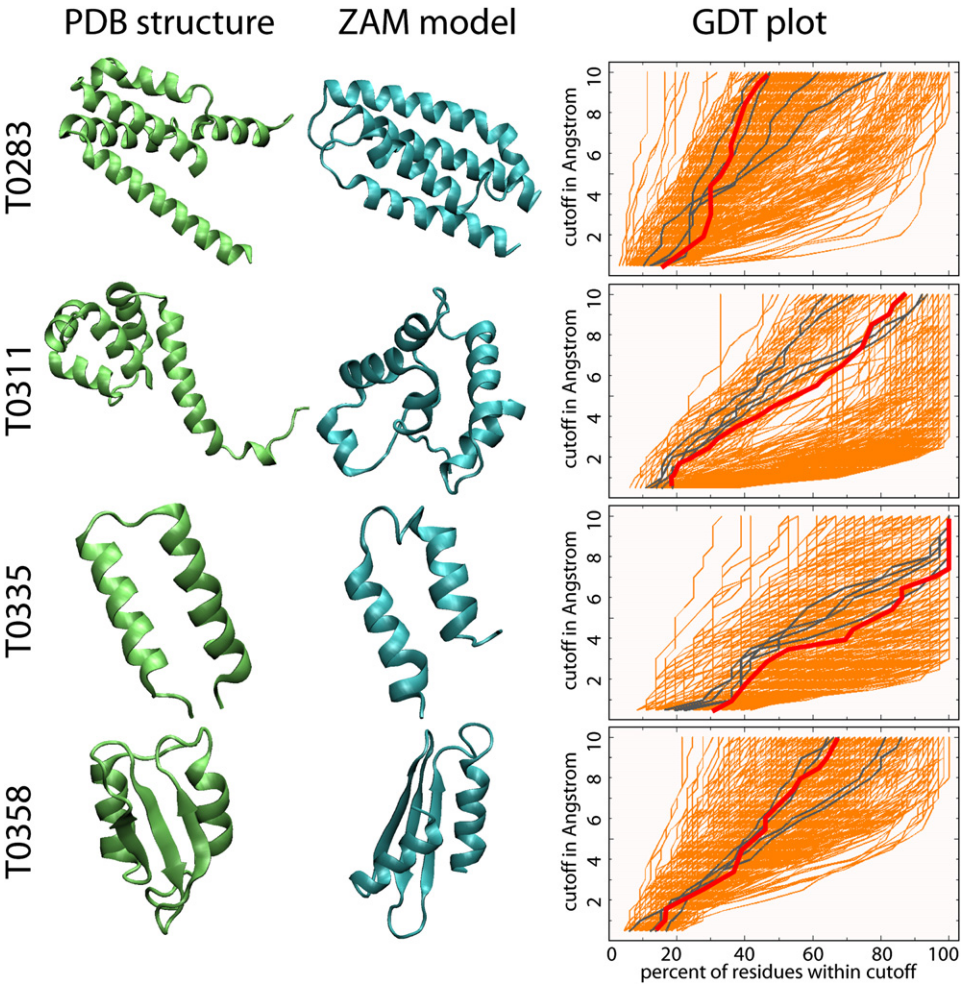

FIGURE 3 ZAM predictions in CASP7 compared with experimental PDB structures. The CASP GDT gives the percentage of residues (x axis) whose Cα coordinates lie within a given cutoff distance (y axis) from the native structure, for predictions by all participants in CASP7 (*orange colors*). The best predictions correspond to lines in the lower-right quadrant of the graph. The five ZAM models are shown for each target in gray, with the structural model shown highlighted in red.

turned out to be domain-swapped and hence inappropriate tests for our method, since the folds in these two cases are partially stabilized by interactions with parts of the chain of additional copies of the same protein. Because we did not explore multichain interactions in our simulations, these targets are prone to errors that reflect more the single-chain focus of our method than real deficiencies of the ZA strategy or the force field. Therefore, since those proteins are likely to be uninformative indicators of sampling or force field errors in ZAM, they are not shown here.

For the four single-domain targets shown in the figure, ZAM predicts roughly correct tertiary structures and segments of 40 residues with an average RMSD of 5.9 Å. Of interest, ZAM predicts secondary structures with 73% accuracy, a value that is close to that of the best bioinformatics-based secondary structure prediction techniques. Fig. 4 shows a comparison of the predicted versus native secondary structures. Though the targets studied are predominantly α-helix dominated, ZAM successfully predicts the three-stranded sheet in T0358. Moreover, the early fragment stages for all of the targets generate a number of candidate β-hairpin structures; in the helical targets, these β hairpins do not survive as stable structures throughout the zipping process.

The secondary structure predictions using ZAM are somewhat surprising given that there has been much concern about imbalances between helix and sheet in force fields (9,10), and about weaknesses in implicit solvent models (11,12). Our results suggest that these problems may be less pronounced for the longer chains we tested here than for the shorter peptides typically studied. We find that the later assembly stages of ZAM help to correct for flaws at earlier stages because wrong early secondary structures are not then later stabilized by the emerging tertiary fold. The implication is that protein folding may have a redundancy of contacts that allows for some robustness against flaws in energy models.

Although ZAM's complete tertiary structure predictions are not as accurate as the best bioinformatics methods, the structures produced fall roughly in line with the average performance of groups at CASP7 and therefore suggest that purely physics-based methods might offer more potential for structure prediction than previously thought. In all four of the targets, we were able to identify sources of error that could be readily addressed in future methods. Target T0283 is a helix bundle protein of 112 residues. For this target, the initial growth stages of ZAM quickly located a large number of potential helices; however, because this was the longest target studied, CASP time constraints greatly limited the number of potential assemblies and topologies that could be pursued at the later stages. Thus, although the secondary structure accuracy was 79% for this protein, our predicted tertiary structure (overall RMSD: 12.1 Å) would have been improved by more sampling at later assembly stages.

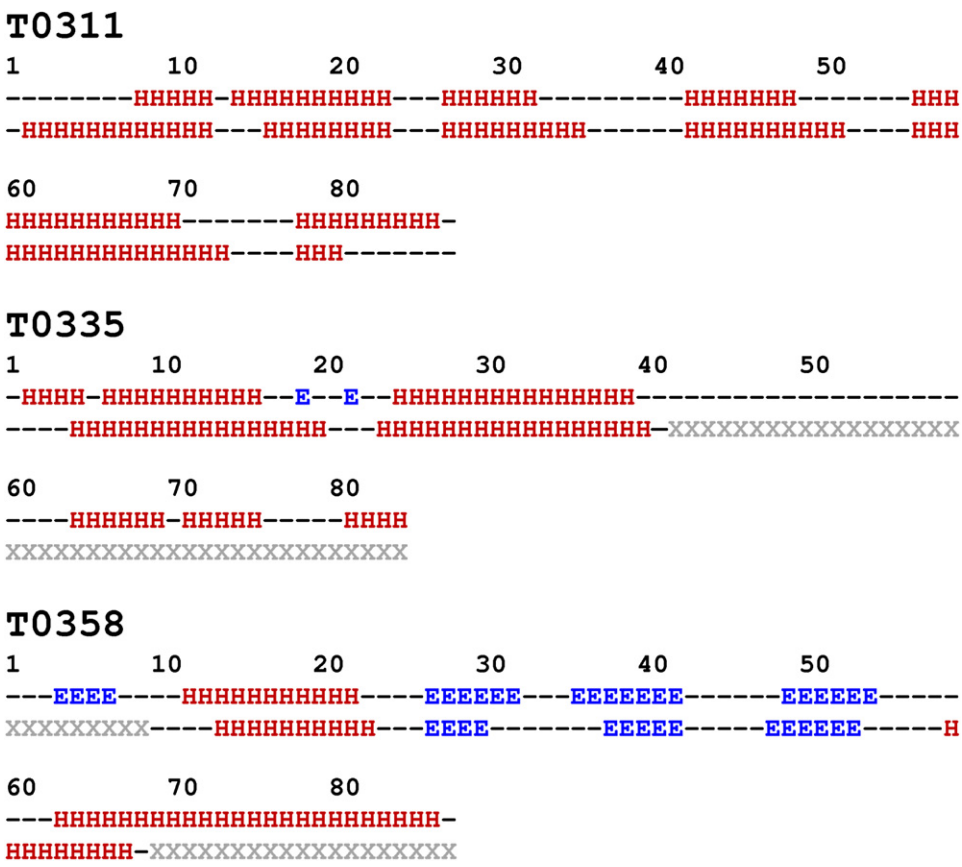

FIGURE 4 Secondary structure analysis of ZAM predictions. The first line for each target gives the DSSP-computed (39) secondary structure; the bottom line gives the same for the native structure. Residues not solved in the native structure are indicated by X.

The target T0311 shows a limitation in the force field we used: salt bridges are overpopulated in the computed structures. This DNA-binding protein consists of roughly 33% charged residues, and in the ZAM folding routes pursued, a dense network of ion-pairing interactions emerged and overstabilized structures with salt bridges in the protein core, resulting in an overall RMSD of 10.6 Å. These interactions were less prominent in the initial growth stages, where there were fewer potential ion-pairing residues in any one fragment, and thus the secondary structure accuracy is still quite good at 72%. Such salt-bridge problems are being studied extensively by force-field developers, which will lead to improved solvation models (11,12,37,38).

ZAM's tertiary predictions for targets T0335 and T0358 were the best of those that we attempted, at overall RMSDs of 5.0 and 7.8 Å. Roughly the second half of the sequence for T0335 appeared as unstructured in the final NMR-determined structure for this protein, leaving only two helices for comparison. Of interest, however, the two solved helices emerged very quickly in the ZAM folding routes as dominant stable structures relative to the remainder of the sequence, which manifested a larger number of possible structural conformers. For T0358, the topology is essentially correct, except for an incorrect symmetry: the three-stranded β-sheet in the prediction is flipped ''inside out'' such that each strand is rotated toward the opposite side of the sheet. This event occurred early on in the ZAM simulation and resulted in the two terminal helices being packed on the opposite side of the sheet, rotating along with the strands.

We also compared the force-field energies (intramolecular potential energy + GBSA solvent free energy) between the native and submitted structures of each protein, and generally found that the native structures differ by no more than 8% in energy, although typically higher than our submissions. However, we caution that energies alone are not definitive metrics in this context, for two reasons 1), in a physical force field with Boltzmann-averaged sampling, it is the free energy (an ensemble quantity) that is minimized, not the force-field energy; and 2), overpopulated salt bridges tend to decrease the energies of our structures somewhat artificially.

In general, we believe that most errors are the result of limiting the numbers of structures retained and ZA pathways explored; that is, given our finite computational resources and the roughly 1-month CASP deadlines, ZAM was not able to fully explore every topological combination of grown fragments and use all clustered conformations from each. This fact is particularly evident from the increased errors found at the tertiary assembly stages. Future studies are under way to evaluate the effects of increasing conformational diversity in the ZAM search mechanism.

CONCLUSIONS

This study gives insight into the capabilities of current purely physics-based methods and modern physical force fields for predicting native protein structures, given only an amino acid sequence. There are certainly some weaknesses to these approaches, however. We find that the current force-field/solvation model tends to overstabilize ion pairs, and that physics-based simulations require, on average, much more computational time than equivalent bioinformatics methods. On the other hand, we find that the pure-physics methods are better at predicting protein structure than expected by many participants at CASP, and perform at a level that is comparable to current average bioinformatics approaches. Furthermore, it is possible that many remaining errors can be managed with ongoing force-field refinements and improvements to the sampling protocol.

Moreover, the present work demonstrates that the ZA-mechanism-based conformational search method is an efficient way to reach native-like structures of proteins in physical models. We have no direct timing comparisons, but folding a 100-mer protein using current brute force Monte Carlo or MD methods is estimated to take tens of thousands of CPU-years, whereas with ZAM it takes about 20 CPU-years. Although this timescale is still long compared to bioinformatics approaches, it represents an acceleration for those systems for which physics-based methods are essential, such as synthetic polypeptides and proteins in nonstandard solution conditions. A clear limitation, however, is that ZAM does not give rate quantities—in part because REMD washes out rate information through temperature swaps, in part because we treated water using an implicit solvent model, and in part because there is as yet no quantitative model for transforming the ZAM restraints and sequence of events into kinetic information.

We appreciate the use of the computing resources provided by the NCSA Supercomputing Center in Illinois, and the UCSF QB3 computing cluster provided by Andrej Sali and his group. This study was supported by a grant from the National Institutes of Health (NIH; GM34993), a UCSF Sandler Opportunities award, and an NIH National Research Service Award fellowship.

REFERENCES

1. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–96.
2. Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15:285–289.
3. Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham Iii, et al. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* 91:1–41.
4. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, et al. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.
5. Jorgensen, W. L., and J. Tirado-Rives. 1988. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110:1657–1666.
6. Hermans, J., H. J. C. Berendsen, W. F. Van Gunsteren, and J. P. M. Postma. 1984. A consistent empirical potential for water-protein interactions. *Biopolymers*. 23:1513–1518.
7. Koehl, P., and M. Levitt. 1999. A brighter future for protein structure prediction. *Nat. Struct. Biol.* 6:108–111.

8. Fiser, A., M. Feig, C. L. Brooks Iii, and A. Sali. 2002. Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* 35:413–421.

9. Gnanakaran, S., and A. E. Garcia. 2003. Validation of an all-atom protein force field: from dipeptides to larger peptides. *J. Phys. Chem. B*. 107:12555–12557.

10. Yoda, T., Y. Sugita, and Y. Okamoto. 2004. Comparisons of force fields for proteins by generalized-ensemble simulations. *Chem. Phys. Lett.* 386:460–467.

11. Felts, A. K., Y. Harano, E. Gallicchio, and R. M. Levy. 2004. Free energy surfaces of β-hairpin and α-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins*. 56:310–321.

12. Zhou, R., and B. J. Berne. 2002. Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water? *Proc. Natl. Acad. Sci. USA*. 99:12777–12782.

13. Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. 282:740–744.

14. Vila, J. A., D. R. Ripoll, and H. A. Scheraga. 2003. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc. Natl. Acad. Sci. USA*. 100:14812–14816.

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

16. Lei, H., C. Wu, H. Liu, and Y. Duan. 2007. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*. 104:4925–4930.

17. Lei, H., and Y. Duan. 2007. Two-stage folding of HP-35 from ab initio simulations. *J. Mol. Biol*. 370:196–206.

18. Simmerling, C., B. Strockbine, and A. E. Roitberg. 2002. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc*. 124:11258–11259.

19. Pitera, J. W., and W. Swope. 2003. Understanding folding and design: replica-exchange simulations of ''Trp-cage'' fly miniproteins. *Proc. Natl. Acad. Sci. USA*. 100:7587–7592.

20. Chowdhury, S., M. C. Lee, G. Xiong, and Y. Duan. 2003. Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.* 327:711–717.

21. Lei, H., and Y. Duan. 2007. Ab initio folding of albumin binding domain from all-atom molecular dynamics simulation. *J. Phys. Chem. B*. 111:5458–5463.

22. Kollman, P. A., R. Dixon, W. Cornell, T. Fox, C. Chipot, et al. 1997. The development/application of a ''minimalist'' organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. *In* Computer Simulation of Biomolecular Systems, Vol. 3, A. Wilkinson, P. Weiner, and W. F. van Gunsteren, editors. Elsevier.

23. Onufriev, A., D. Bashford, and D. A. Case. 2000. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B*. 104:3712–3720.

24. Ozkan, S. B., G. H. A. Wu, J. D. Chodera, and K. A. Dill. 2007. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA*. 104:11987–11992.

25. Simon, I., G. Némethy, and H. A. Scheraga. 1978. Conformational energy calculations of the effects of sequence variations on the conformations of two tetrapeptides. *Macromolecules*. 11:797–804.

26. Liwo, A., P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy, et al. 2002. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application to the UNRES force field. *Proc. Natl. Acad. Sci. USA*. 99:1937–1942.

27. Liwo, A., M. Khalili, and H. A. Scheraga. 2005. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. USA*. 102:2362–2367.

28. Shell, M. S., R. Ritterson, and K. A. Dill. 2008. A test on peptide stability of AMBER force fields with implicit solvation. *J. Phys. Chem. B*. 112:6878–6886.

29. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

30. Fiebig, K. M., and K. A. Dill. 1993. Protein core assembly processes. *J. Chem. Phys*. 98:3475–3487.

31. Weikl, T. R., and K. A. Dill. 2003. Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* 332:953–963.

32. Weikl, T. R., and K. A. Dill. 2003. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol*. 329:585–598.

33. Voelz, V. A., and K. A. Dill. 2006. Exploring zipping and assembly as a protein folding principle. *Proteins*. 66:877–888.

34. Hockenmaier, J., A. K. Joshi, and K. A. Dill. 2006. Routes are trees: the parsing perspective on protein folding. *Proteins*. 66:1–15.

35. Dill, K. A., A. Lucas, J. Hockenmaier, L. Huang, D. Chiang, et al. 2007. Computational linguistics: a new tool for exploring biopolymer structures and statistical mechanics. *Polymer (Guildf.)*. 48:4289–4300.

36. Ho, B. K., and K. A. Dill. 2006. Folding very short peptides using molecular dynamics. *PLoS Comput. Biol*. 2:e27.

37. Feig, M., A. Onufriev, M. S. Lee, W. Im, D. A. Case, et al. 2004. Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem*. 25:265–284.

38. Tan, C., L. Yang, and R. Luo. 2006. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *J. Phys. Chem. B*. 110:18680–18687.

39. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22:2577–2637.