Union of Geometric Constraint-Based Simulations with Molecular Dynamics for Protein Structure Prediction

Tyler J. Glembo and S. Banu Ozkan*
Center for Biological Physics, Department of Physics, Arizona State University, Tempe, Arizona

ABSTRACT Although proteins are a fundamental unit in biology, the mechanism by which proteins fold into their native state is not well understood. In this work, we explore the assembly of secondary structure units via geometric constraint-based simulations and the effect of refinement of assembled structures using reservoir replica exchange molecular dynamics. Our approach uses two crucial features of these methods: i), geometric simulations speed up the search for nativelike topologies as there are no energy barriers to overcome; and ii), molecular dynamics identifies the low free energy structures and further refines these structures toward the actual native conformation. We use eight α-, β-, and α/β-proteins to test our method. The geometric simulations of our test set result in an average RMSD from native of 3.7 Å and this further reduces to 2.7 Å after refinement. We also explore the question of robustness of assembly for inaccurate (shifted and shortened) secondary structure. We find that the RMSD from native is highly dependent on the accuracy of secondary structure input, and even slightly shifting the location of secondary structure along the amino acid sequence can lead to a rapid decrease in RMSD to native due to incorrect packing.

INTRODUCTION

The question of how the amino acid sequence (i.e., the primary structure) of a protein folds into a unique 3-D structure is considered as one of the biggest challenges in science (1,2). Therefore, developing tools for protein structure prediction is one of the most pursued goals by scientists from different disciplines. As the massive amount of data from the genome sequencing effort builds, the demand for these useful tools increases. One of the biggest contributions to structure prediction methods comes from the competition called the CASP, which is an international competition that assesses the current state of the art in protein structure prediction. The results of CASP have shown that the synergy of interdisciplinary effort has allowed for many advances to be made in surmounting this challenge (3–5).

There are two foremost approaches for protein structure prediction: i), comparative modeling: template-based homology modeling based on sequence similarity of experimentally known 3-D protein structures; and ii), ab initio or de novo modeling: free modeling without knowledge of a 3-D template model. The first goal of comparative modeling methods is to associate the target protein with at least one or more structurally related proteins with known experimental structure. This is usually achieved by sequence alignment of the target (unknown) protein with a database of known proteins and if a high sequence similarity exists, the 3-D structure of the target protein is assumed to be the same fold. Then, the target protein is modeled by threading the target protein into the template structure along with some energy minimization sampling methods. Significant progress has been made in this area especially in threading methods (6,7). However, as the sequence similarity decreases between the target protein and the database proteins (with a sequence similarity of $<\sim$15%), the errors in prediction increase (3,7) due to incorrect template or reduced structural similarity between target and the template proteins.

Free modeling, on the other hand, aims to predict the 3-D structure from scratch, without using any 3-D structure of known proteins as a template; thus, the success in free modeling is certainly the ''Holy Grail'' of protein folding. The advancements in free modeling help lead to a better understanding of protein folding mechanisms and to better methods of designing new enzymes (8). One of the leading structure prediction techniques, called Rosetta, is based on the free modeling assembly of 3–10 residue templated fragments using a Monte Carlo-based sampling method (9,10). Besides Rosetta, there are other useful methods that use the same approach of ''fragment assembly'' for protein prediction (6,11–15). Typically the methods differ in the way they extract fragments and the sampling methods used in fragment assembly. A method developed recently called TASSER (6) predicts the structure of low homology sequences successfully (i.e., the difficult case by comparative modeling) by dividing the target sequence into two regions after threading: regions that aligned well with the template and gapped regions that need to be treated with free modeling.

Although many protein structure prediction techniques are based on the bioinformatics method of statistical inference techniques, physics-based methods of structure prediction,

Submitted September 9, 2009, and accepted for publication November 17, 2009.

*Correspondence: Banu.Ozkan@asu.edu

Abbreviations used: CASP, Critical Assessment of Techniques for Protein Structure Prediction; FRODA, framework rigidity optimized dynamics algorithm; MD, molecular dynamic; REMD, replica exchange molecular dynamics; RMSD, root mean-square deviation; r-REMD, reservoir replica exchange molecular dynamics; ZAM, zipping and assembly method; ZAMF, ZAM with FRODA; 3-D, three-dimensional; 1-D, one-dimensional.

Editor: Ruth Nussinov.

0006-3495/10/03/1046/9 $2.00

doi: 10.1016/j.bpj.2009.11.031

which use molecular mechanics force fields to reproduce the true intramolecular and solvent interactions governing protein structure, are still being worked on rigorously. These physics-based methods are not as fast or accurate in prediction when compared to bioinformatics-based methods, but they are working to catch up, if not without difficulty (5,16,17). A large part of this difficulty arises due to the large and rugged the conformational space of proteins. This difficulty can be overcome, in part, with the smart sampling methods, and although there are studies indicating that work remains necessary to improve current force fields (16,18–20), in many cases it is possible to reach the native structure using physics-based methods (17,21–33).

Ozkan et al. (24) have recently developed a purely physics-based structure prediction method called ZAM. ZAM uses a search strategy on top of conventional molecular dynamics to explore putative folding routes that lead most directly and efficiently to the native structure. ZAM has been tested through the folding of eight small proteins from the PDB to within 2.5 Å, giving good agreement with the experimental ϕ-values known for four of them from experimental transition state studies. In a more stringent test, ZAM was applied in CASP7, to the folding of six small proteins with from 76 to 112 residues (34).

CASP7 results have shown that ZAM found secondary structural elements relatively efficiently, however the assembly of these secondary structures to 3-D structure takes longer. In this study, we develop what we believe to be a novel approach to speed up the assembly the secondary structural elements into tertiary nativelike structures in the assembly stage of ZAM. Our approach has two unique features. First, it uses a geometric-based conformational sampling technique called framework rigidity optimized dynamics algorithm (FRODA) to generate a variety of different topological structures given the secondary structures (35). FRODA is a Monte Carlo-based geometric simulation that explores the motion of proteins through random motion of rigid clusters within the protein. It can explore the large-amplitude motions of larger systems (i.e., longer timescale motions) up to 160 times faster than MD (M. F. Thorpe, unpublished results). These considerable computational savings allow us to speed up the conformational search for assembly. Moreover, it has been shown that geometry-based type of approaches in folding studies can shed light into principles of protein folding (36–38). Second, our approach couples the FRODA generated assemblies with REMD (39) using a reservoir (40) to select the energetically favorable structures because there is no real energy function in geometry-based simulations. Thus, coupling FRODA generated structures with r-REMD helps evaluate the best nativelike topologies among all of the FRODA assembled structures, and due to the metropolis nature of REMD runs, the low energy structures dominates the lowest replica. In addition, this step also helps to refine the best model, and, after a couple iterations, allows exploration of possible dihedral and distance constraints. With this approach we folded small, globular α-, β-, and α/β-proteins with an average RMSD of 2.7 Å when we use the correct secondary structures. We also explored how the accuracy in prediction changes as accuracy in the secondary structure predictions decreases.

Although this current work focuses on the assembly and refinement stages of assembling secondary structures into a tertiary structure where much work has already been done successfully (11,14,41), our assembly stage is different in that: i), it is not database driven; ii), it does not build loops between already packed helices; and iii), it explores all possible assembly pathways. In addition, our method can begin with only the 1-D sequence and move to tertiary structure. However, exploring all possible assembly paths of all the secondary structure found by zipping would be considerably more computationally expensive than the assembly of correct secondary structures, thus our next goal is to incorporate the accuracy of secondary structure prediction to our assembly by using new features of FRODA.

In this study, we first provide the details of the method: how we choose the assembled units, what type of FRODA parameters are used, how we differentiate the nativelike structures from nonnativelike structures. Then we discuss the results of our method as to eight proteins and show how accuracy of the tertiary predictions are related to accuracy of secondary structure predictions. Finally we provide some concluding remarks.

METHODS

To generate secondary structure from the 1-D amino acid sequence we make use of the zipping method (24) in which the amino acid sequence is chopped into overlapping 8mers that are then grown into stable secondary structure through multiple REMD runs. However, as this portion of has already been tested, for our initial test we will be using exact secondary structure from experimentally determined structures. These are then assembled with a combinatorial approach to generate many possible topologies for the final, complete protein, using FRODA, as discussed below in the Assembly section. Finally, these many structures are ranked and refined to determine the most likely native candidate via r-REMD as discussed in Refinement. These steps are outlined in Fig. 1.

Assembly

Once the initial fragments are determined via zipping we can then pass them to FRODA for assembly. FRODA is a constrained geometric Monte Carlo routine that is based on identifying rigid clusters within a protein using the pebble game algorithm from graph theory (42–44). These rigid clusters are then used to create inequality constraints (35,45) that provide an artificial energy landscape and hence, a function to be minimized. After a random perturbation of rigid clusters, by minimizing this energy function, all the constraints that do not allow steric collision, etc., are met and the protein is now in a new stereochemically acceptable conformation. With an iterative random perturbation of rigid clusters, FRODA effectively samples the conformational space available to a protein while preserving the local stereochemistry. With this tool, the naïve method might be to generate secondary structures via zipping, generate a loop between fragments in extended conformation, and simply let FRODA run, exploring unique conformations. Considering the sheer number of possible conformations, this is not a tenable solution. To intelligently use FRODA we introduce additional steps. First, the assembly is done both iteratively and all at once. We could put all the

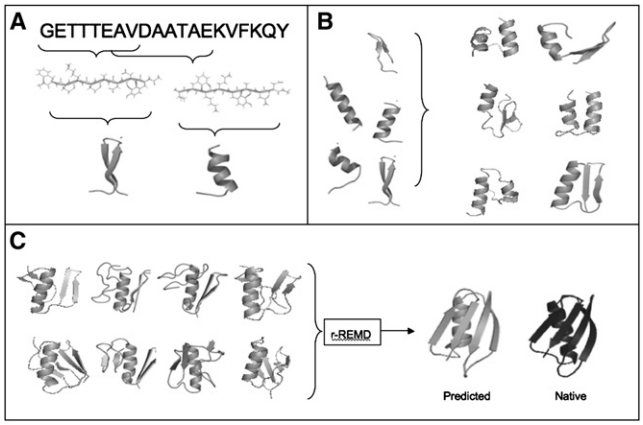

FIGURE 1 (*A*) The 1-D sequence is chopped into overlapping 8mers, which are then run in MD simulations. If stable bonds are formed those bonds are stabilized and additional residues are added to the 8mer, allowing secondary structure to form over multiple iterations. (*B*) The many secondary structure fragments are then combinatorially assembled to form partial proteins with FRODA. This is done iteratively until the entire protein is assembled. (*C*) All possible structures are ran in multiple r-REMD simulations to discriminate which is the likely native structure. The structures that dominate most heavily the lowest temperature replica are assume to be the most nativelike.

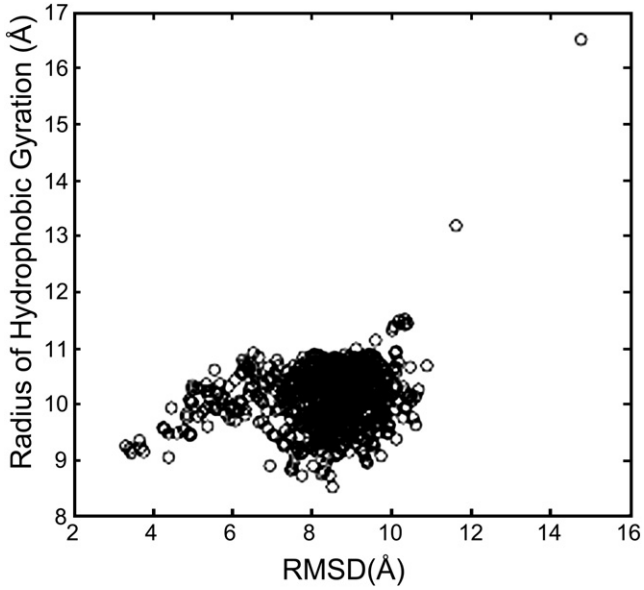

FIGURE 2 1PRB radius of hydrophobic gyration versus RMSD to native. The lowest radius of hydrophobic gyration states are populated by a near native configuration and its mirror image.

fragments together, and build all the necessary loops in one step, thus giving us the full sequence of the protein. Unfortunately, this does not always lead to good results so we use both this and the iterative, or two-by-two method, where we start assembling the secondary structural motifs (i.e., folded fragments) with the shortest loop in between (11). We use only two fragments initially, and build a loop between these two. After running this mini-structure through FRODA, we then take other mini-structures that have already been ran, and build loops between them. Thus, we will now have a structure with four secondary structure units, built up two-by-two. This process is continued until the entire protein has been assembled from fragments and sets of fragments. The missing residues between these fragments are built in as a loop in extended conformation, and thus, the initial conformation has units of secondary structure that are very distant from each other. The radius of gyration is actually maximized to help prevent steric collision during this loop building phase, with the added benefit that this will allow for the greatest amount of freedom of movement once the actual geometric simulation begins. Once this is done the fragments are biased toward each other in two unique approaches. First, there is a simulated annealing step in which we minimize the radius of gyration of hydrophobic residues (RgPh) using the list of {ALA, VAL, LEU, ILE, MET, PHE, TRP} residues. There is a correlation between nativeness of a conformation and its RgPh for structures with enforced secondary structure, as Fig. 2 shows, thus this step will help discriminate between possible nativelike conformations and unfolded conformations (11). However, this alone does not ensure multiple unique conformations being generated. Therefore, as the second approach, we introduce what we believe to be a novel scheme of doing many serial runs in parallel, each of which has a unique set of hydrophobic residues that have been paired together. The pairing does not introduce a simulated annealing, but rather introduces a perturbation. After each random perturbation in FRODA, an additional perturbation of 10% of the random throw is then added to each chosen pair of hydrophobic residues that pushes them closer together. If they are already within 7.0 Å of each other no perturbation is added. We choose 7.0 Å as a representative distance as we wish to mimic a hydrophobic tether of proper distance (44) while still allowing for motion at the same time.

Once these initial conditions are set, the actual run begins. After each random move, up to 500 steps of conjugant gradient minimization are carried out. After the minimization, the direction of the move from initial to final state is noted and the next random throw is biased to be in the same direction as the previous, which we describe as a momentum run on. This allows large conformational changes to be made quickly and efficiently. We find speed increases from 5- to 40-fold by using momentum run on.

Running multiple FRODA runs in parallel quickly generates many unique conformations, however, each successive snapshot is likely to be very similar to the previous, with large scale motions happening only after many steps. To cut down on the raw data being analyzed, we use a *k*-means clustering algorithm-based on a 2.0 Å RMSD between atomic positions to get representative structures of the many thousands that are generated (34). The final clustered structures are then scored by the radius of gyration computed over all the hydrophobic residues, with the structures having the lower radius of gyration assumed to be more nativelike, as shown in Fig. 2. During the intermediate assembly stages(i.e., assembly of three or more secondary structural motifs) the number of clustered structures far exceeds the number of structures it would be reasonable to continue assembling, thus we must include a filtering step to discriminate among the clustered structures. Therefore, we choose only a set of few structures with two criteria: i), the selected structures must have low RgPh; and ii), the selected structures must have all different topologies. We achieve this through selecting the structures with high RMSD between them among the set of low RgPh structures. However, during intermediate assembly stages, we exclude any structures that scored in the top 10% for lowest RgPh as these structures typically are packed too tightly to continue assembly, or loop regions have migrated to the centroid, which then push any structured regions out to the surface. Once these representative structures are chosen we continue on to the next stage of assembly.

Refinement

After assembling a protein with FRODA there are two issues. The first issue is one of dealing with the large numbers of clustered structures generated, and the second is one of refinement. To deal with the first issue, we take into account scoring mechanisms such as the RgPh and the number of hydrophobic contacts. We combine these into a single scoring function and, as this has a correlation with nativelike topologies, we can then rank each of our structures from FRODA and decide which structures to use as seeds for each replica. The rest of the structures are coupled to REMD as a reservoir.

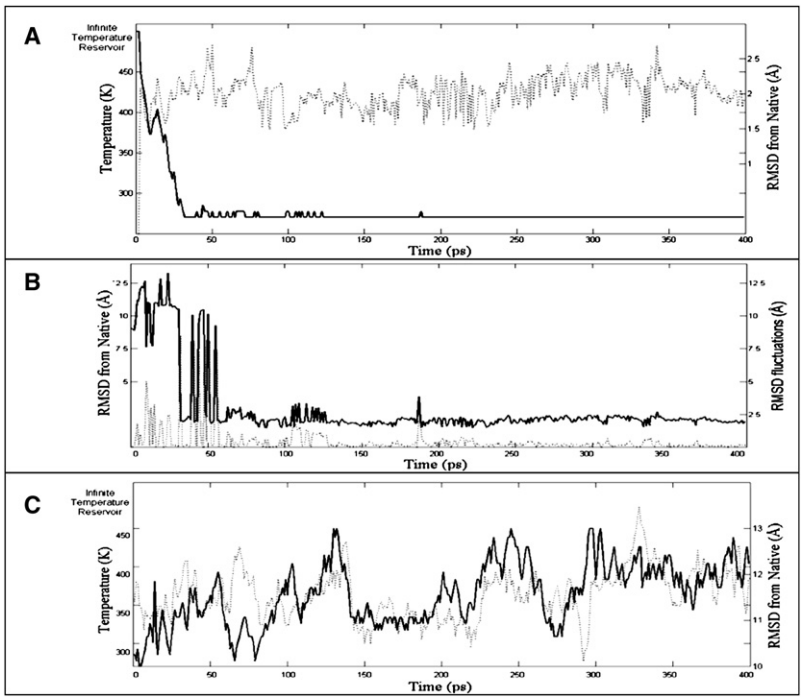

FIGURE 3 (*A*) Plot showing which replica the most nativelike structure exists in (*solid line*) throughout a short r-REMD simulation, along with the RMSD from native of that structure (*dotted line*). After a few ps, the nativelike structure is swapped from the reservoir into the highest temperature replica and quickly falls down into the lowest temperature replica. The RMSD of this nativelike structure from the experimentally determined structure remains close to 2 Å throughout the simulation despite passing through high temperature replicas. (*B*) The RMSD from the experimental structure of the lowest temperature replica (*solid line*) for the same simulation in *A* and the RMSD fluctuations of the lowest temperature replica (*dotted line*). The RMSD decreases to 2 Å around 40 ps and the RMSD fluctuations converge to 0 Å, showing that a single nativelike structure will dominate the lowest temperature replica over time. (*C*) Following a random structure assembled by FRODA as it travels through the replicas (*solid line*) using the same r-REMD run as *A*, along with the RMSD from native of that structure (*dotted line*). Initially this structure makes a random walk through temperature space, but as the lowest temperature replicas become dominated by nativelike structures, it is relegated to higher temperatures. Note that this structure was one of the few that was not swapped out into the reservoir during this run.

This is achieved by introducing an infinite temperature reservoir full of structures that can be swapped into the highest temperature replica during the REMD run (40). The REMD has replicas from 270 K to 450 K to achieve maximum efficiency, using the AMBER ff96 force field (46) with the generalized Born/surface area (GBSA) implicit solvation model of Onufriev et al. (47) (AMBER option ''igb'' = 5) and a surface tension of 0.5 kcal/mol. Swaps between replicas are attempted every 1 ps and a molecular dynamics timestep of 2 fs is used. Each swap is carried out five times per cycle. The swap likelihood is set at 50% by distributing the replica temperature exponentially (39) (e.g., 70 residue protein 2K53 uses 24 temperature replicas spaced exponentially and achieves an overall swap ratio of 0.534 during a 3-ns run).

Once the seeds and the reservoir structures are chosen and minimized they are each run for 3 ns. Of interest are two phenomena that occur during an r-REMD run, both of which aid us greatly in our search for the native protein structure. The first is that proteins in nonnative topologies often unfold spontaneously during a run due to poor energetics. The second is that proteins that are near native are already ''falling down'' into the native well. Fig. 3 *A* tracks a nativelike conformation that is swapped from the reservoir into an REMD run, showing that nativelike conformations will dominate the lowest temperature replica. This means that if there is a near-native topology somewhere within our seeds or reservoir structures, it will quickly move as close to the native structure as our force fields will allow, whereas at the same time quickly becoming the dominant structure by moving into and staying in the lowest temperature replica. These two phenomena lead us to the second phenomena of refinement. The first r-REMD run is analyzed for likely native contacts between residues and these are used as restraints in the second iteration of the REMD run. We cluster the lowest temperature replica and use these clustered structures as seeds for the second iteration run. This process is continued until a single dominant cluster emerges, and this cluster is taken as the native structure. It typically requires 3–4 iterations until a single dominant cluster emerges, although there have been cases where a dominant cluster emerges after only the first r-REMD run as shown in Fig. 3 *B*, the time evolution of RMSD of the snapshots sampled at the lowest replica and corresponding RMSD fluctuations. We observe that RMSD drops to ~2 Å in a very short time frame and the fluctuations at the lowest temperature replica quickly converge to zero. This indicates that when a nativelike conformation exists in the infinite temperature reservoir it can travel along the replicas quickly and dominate the lowest replica. Additionally, Fig. 3 *C* shows time evolution of a random nonnativelike structure as it moves through the replicas during the same simulation as Fig. 3, *A* and *B*, along with its RMSD from the experimentally determined structure. This structure is naturally moved into the higher temperature replicas as the run progresses, due to more nativelike structures with lower energy dominating the lower temperature replicas, which is exactly what is aimed for. However, this structure still samples the available temperature space effectively, showing the efficiency of the r-REMD method.

Of note is the introduction of the reservoir during our refinement stage to discriminate between the large numbers of structures in each REMD run (40). Our entire philosophy is to try and generate a great number of conformations that fairly well represent the possible conformational space, and therefore the introduction of a reservoir is necessary, otherwise it would not be possible to analyze all of the data generated. Although the typical swap likelihood we choose between replicas is 48% (set by appropriate exponential spacing between replica temperatures for an appropriate Boltzmann factor) (39), the reservoir is essentially at infinite temperature so therefore the likelihood of a swap between the highest temperature replica and the minimized structures in the reservoir nears 100%. This, of course, violates detailed balance, but reversibility is not the end goal here, only reaching a desired end state. Additionally, it could be argued that you can deconstruct the REMD run into many individual runs between each swap with the reservoir, and each of these would obey detailed balance.

RESULTS AND DISCUSSION

Using our method, we attempted to assemble a number of small, globular proteins, including all α-proteins, β-proteins, and α/β-proteins using their native secondary structures as initial fragments. As explained in the details of the Methods section, we use experimentally determined secondary structure during this initial test and generate an ensemble of structures using FRODA that are then later refined with r-REMD, in the same way as in ZAM, to determine the most nativelike structure. FRODA is a geometric, constraint-based algorithm

that uses Monte Carlo moves to explore possible conformational space, and as such it is up to 160 times faster than Amber MD with generalized Born implicit water and a cutoff of 12 Å. With the momentum run on consideration described earlier, FRODA is potentially orders of magnitude faster, not at exploring conformational space, but at moving in a directed motion, such as undergoing a hydrophobic collapse. In the test case for 1AIL (73 residues), it took <30 s to move from a conformation in which all the loops were in extended conformation to a hydrophobic collapsed state for the full protein. Naturally, FRODA does not have an energy function that can distinguish between nativelike or nonnative conformations, and thus, the addition of our clustering algorithm with multiple steps of r-REMD merges geometric constraint-based algorithms with all-atom MD in a unique manner. To verify that coupling r-REMD with FRODA delivers good results, we tracked a series of structures through the replicas they populated during an r-REMD run. Fig. 3 *A* shows the results of tracking a nativelike conformation through a short r-REMD run. As shown in Fig. 3 *A*, we have had success with this method due to the fact that if a structure exists in the REMD simulations that is very nativelike, it will move to and dominate the lowest temperature replica. Additionally, Fig. 3 *A* shows that introducing a reservoir from which structures can be swapped allows for a greater number of initial conformations to be sampled without losing accuracy, as a nativelike structure swapped into the highest temperature replica will still move down through the replicas to dominate the lowest temperature replica, thus greatly increasing efficiency. Due to this, we do not decide which is the most nativelike structure by computing the RMSD of sampled snapshots, but rather, it is assumed that the structure that dominates the lowest temperature replica is most nativelike.

Table 1 presents the RMSD from the experimentally determined structures for our FRODA assembled structures both before and after refinement with an average RMSD of 3.7 Å before refinement and 2.7 Å after refinement. A few of the proteins on Table 1 improved their RMSD by almost 2 Å by undergoing the refinement stage, with an average improvement of a full angstrom. Additionally, although RMSD can be a good measure of our method, it should be noted that many of the FRODA structures are hundreds of kJ/mol higher than the minimized REMD structures in AMBER potential, and thus provide a very useful first step but must be refined. Fig. 4 presents ribbon diagrams of seven of those structures. The structures in the bottom row in Fig. 4 are the predicted structures from ZAM with FRODA (ZAMF) whereas the top row of structures are determined experimentally. The average RMSD of the predicted structures from the experimentally determined structures is 2.7 Å for the top-ranked structures in this test set. All of the structures determined in Fig. 4 were done so by completely enumerating all possible combinations of secondary structure, however, once a nativelike structure emerges is it possible to trace the assembly path back to the secondary structure units. For the following assembly pathway discussion, the secondary structure will be identified by proximity to the N-terminus (e.g., helix 1 will be the N-terminal helix, whereas helix 2 will be the helix adjacent to helix 1, and likewise for β-hairpins).

TABLE 1 RMSD from native both before and after refinement for each of the proteins of our test set

			RMSD Å	
Protein name	Type	Length (AA)	FRODA assembly only (Å)	With refinement (Å)
1AIL	α	73	4.5	3.2
1PRB	α	53	3.3	1.9
1BDD	α	60	3.9	3.1
1EOL	δ	37	1.9	1.7
1EON	δ	27	1.8	1.7
1GB1	α δ	56	3.8	2.1
2ICP	α	94	5.8	5.0
2K53	α	76	4.4	2.8

Assembly pathway that gives the lowest RMSD structures

For 1PRB (1.88 Å) and 1AIL (3.2 Å), the native structure assembly path was found to be first assembling helices 2 and 3, followed by assembling helix 1 onto helices 2 and 3. For 1BDD (3.1 Å), the native structure assembly path was found to be first assembling helices 1 and 2, followed by assembling helix 3 onto helices 1 and 2. Interestingly, MD and experimental studies have both found the folding pathway of protein A to be similar to the folding pathway found using our method (24,48). 1E0L and 1E0N (1.7 Å each) were both found to have been assembled by assembling the β-strands 1 and 2 with the extended strand 3. Again, using our simple approach, the folding pathway found agrees with previous MD studies (16,24). For 1GB1 (2.1 Å) the assembly pathway to the nativelike structure was found to be first assembling the helix with N-terminal β-hairpin, followed by assembling the C-terminal β-hairpin on to the helix-N-terminal β-hairpin. Finally, for 2K53 (2.8 Å), the assembly pathway that led to the nativelike structure was found to be first assembling the N-terminal and C-terminal helices with the adjacent helices (i.e., helix 1 with helix 2 and helix 3 with helix 4), and then assembling helices 3 and 4 onto helices 1 and 2. Our assembly method compares favorably in terms of final results with other methods that have been used to assemble these proteins (16,23,24). The proteins assembled in this test case are currently too large for all atom explicit water MD simulations on all but the larger computing clusters, and current force fields do not always generate the experimentally determined conformation (16), thus ZAMF is able to generate accurate results for certain notoriously difficult to fold proteins. Other methods (14) have found that combinations of secondary structure and hydrogen bonding can effectively

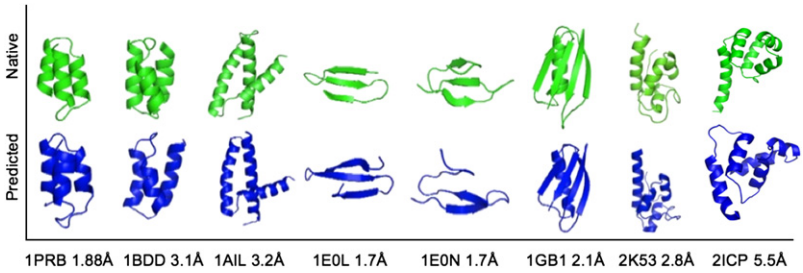

FIGURE 4 Proteins successfully folded by ZAMF. The top row is the experimentally determined structure. The bottom row is the predicted structure. The RMSD of predicted structures are, from left to right: 1.88 Å (1PRB), 3.1 Å (1BDD), 3.2 Å (1AIL), 1.7 Å (1E0L), 1.7 Å (1E0N), 2.1 Å (1GB1), 2.8 Å (2K53), 5.5 Å (2ICP) with an average of 2.74 Å.

sample nativelike states in a Monte Carlo simulation, although actually determining which of the states sampled is the native state is not yet possible in such methods. Our method uses secondary structure along with hydrophobic pairing/collapse, although in the future dynamically searching hydrogen bonding is something to be learned from these methods. CPU time used is highly dependent on the size of the protein assembled and the number of possible hydrophobic contacts, so it is hard to determine an average. However, for this test set, on the average we use ~500 CPU h/protein, a number that should continue to reduce with future refinements. Other methods, such as UNRES/MD (23) were able to fold multiple proteins using minimal computer time to within 5 Å RMSD from native, with a required CPU time of 2–10 h per trajectory and 10 trajectories each. Our CPU time is certainly longer, ~6 CPU h for the FRODA assembly stage of protein A and ~190 CPU h for the refinement stage, however, the end result is a more refined protein that is 3.1 Å RMSD from native. Although there is a tradeoff between accuracy and CPU time/power in all methods, each of these methods all work toward solving the protein folding problem in their own way, and the synergy of all our methods combined helps to bring us all closer to a robust physics-based protein folding solution.

Accuracy when the secondary structures are not correct

During our initial test case, as discussed previously, we used only experimentally determined secondary structure. ZAM predicts secondary structure with ~73% accuracy (34) that leads to concern over the robustness of ZAMF when secondary structure is not exact. To test this, the secondary structure units of protein 2K53 were both shortened and shifted to determine the effect each would have on the robustness of ZAMF, as shown in Fig. 5, i.e., the helix adjacent to the C-terminal runs from residues 41–49 and during our runs it was shortened to residues 42–48, again to residues 42–47, and shifted to residues 39–47. When 2K53 was folded with exact experimentally determined secondary structure, the RMSD from the experimentally determined structure was 2.8 Å after refinement. When the secondary structure was only slightly shortened, the RMSD was 5.3 Å, whereas when the secondary structure was shortened

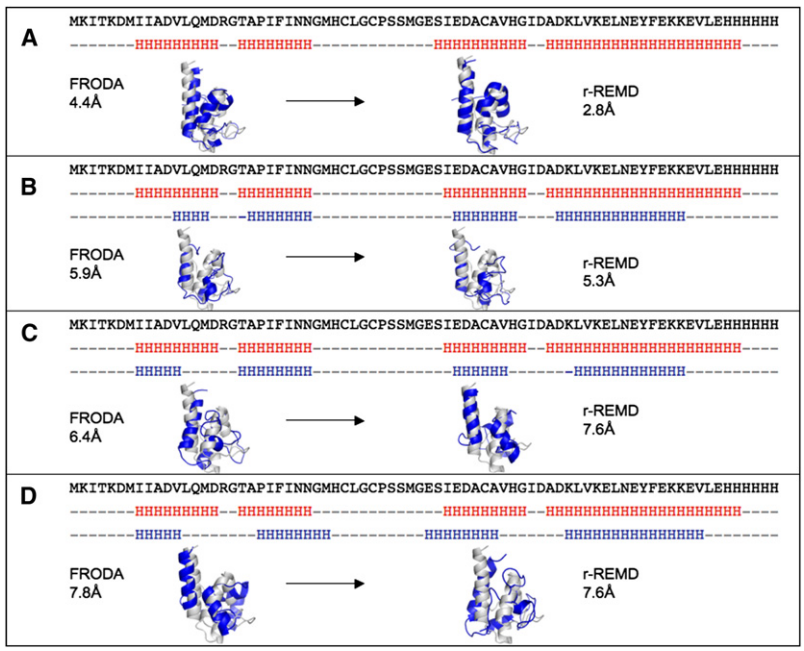

FIGURE 5 Native secondary structure (*red/first row*) and the altered secondary structure (*blue/second row*) are presented along with the superimposed ribbon diagrams of 2K53. Experimental structure is in gray/light, predicted 3-D structure is blue/dark. (*A*) When using perfectly accurate secondary structure, good results are achieved at both intermediate (after FRODA assembly) and final steps (after r-REMD). (*B*) Using shortened secondary structures (i.e., shorter helices) that mostly preserve loop lengths closer to correct native secondary structure allows for good packing and correct topology, and after a short r-REMD run the secondary structure is beginning to reform. (*C*) Significantly lengthening loop length by shortening helices causes a larger conformational search space, thus causing the topology to be slightly incorrect and lead to high RMSD structures. (*D*) Shifting the secondary structure leads to significantly perturbed topologies, and secondary structure begins to unfold during r-REMD simulations.

significantly near the loop regions, the RMSD dropped to 7.6 Å after refinement. The assemblies of shortened secondary structure found the interesting result that the robustness of our assembly method decreased as the secondary structures continued to shorten, or, to look at it from another perspective, as the loop length between secondary structures increases the possible conformational space to be searched (assuming rigid secondary structural units) increases and thus the ability of our search algorithm to sample this entire space decreases. The secondary structure did begin to reform during refinement when the overall topology was near nativelike, showing that even with shortened secondary structures it remains possible to sample and eventually refine to near native conformations. Had we continued further refinement, it is likely that these structures would have become more nativelike and the native secondary structure would have fully reformed. When the secondary structure units were shifted the robustness of ZAMF greatly decreased. The RMSD of the predicted structure with shifted secondary structure was 7.6 Å after refinement. There is likely a twofold reason for this. The first is that the correct topology cannot be sampled with a simple Monte Carlo geometric constraint-based algorithm if the loops are in the wrong places. The second is that once these shifted structures were refined, the incorrect secondary structure unfolds at a greater rate than correct secondary structure can begin to fold in REMD, which we see during refinement.

There are a few ways in which to address the issue of low resolution structures (i.e., RMSDs >4 Å) as the accuracy of secondary structures decreases. The first is to increase the accuracy of secondary structures with the use of secondary structure prediction servers (49–53) and incorporate their predictions along with our own predictions. ZAM predicts secondary structure at ~73% (34) and servers predict secondary structure with an accuracy of up to 80% (49–53) so coupling ZAM predictions with server prediction at the early zipping stage can increase the accuracy of secondary structures. Another possibility along those lines is to expend further computer time on predicting the secondary structures, however, the secondary structure prediction will ultimately be limited by the force field used (19). To combat the issue of increased CPU usage while still running longer simulations, coarse grained iterative fixing models such as the ItFix algorithm (17) can also be explored for more accurate secondary structure prediction. Finally, the shortened secondary structures, although they do not produce results as well as those with exact secondary structure, still are able to accurately sample the correct topology, whereas shifted structures do not. It would be conceivable to then simply shorten all our predicted secondary structure to avoid any shifted type structures, and use the shortened structures during all future assembly. Thus we can increase the accuracy of consistently sampling the right topology by using shortened structures and combat this on the backend by increasing the simulation length of the REMD during the refinement stage with more strong dihedral restraints to ensure the right secondary structural motifs emerge. In truth, it will likely be a combination of the above methods that will ensure the most accurate prediction of secondary structure in future predictions.

Overall, our analysis has shown that folding purely α-helical proteins has so far been very successful, with most targets giving good results. On the other hand, β-sheet proteins prove more of a challenge. Two issues that arise in our approach are in keeping β-strands extended in a Monte Carlo simulation and forming nonlocal β-sheets.

The first of these issues can be addressed in multiple ways. The first is by ensuring excellent hydrogen bonding along local beta sheets that will help to rigidify them. One future improvement that we are working on is the ability to analyze hydrogen bonds dynamically during a FRODA run. The issue of ensuring excellent hydrogen bonding to further rigidify β-sheets would alleviate itself to some degree if, instead of analyzing β-sheets before the run we could use the software to push toward better hydrogen bonding during a run. This would also greatly help to identify possible nonlocal β-sheets by analyzing the contact order of hydrogen bonds made and broken during a run. We believe this will also help to improve the accuracy of the prediction when we assemble the secondary structures with low accuracy.

The second of these issues is forming nonlocal β-sheets. This is an issue that we hope to solve by sampling. Our philosophy of sampling many unique topologies should theoretically allow us to sample nonlocal β-structures. If indeed we do sample this, then it should come through during the r-REMD run. However, given the huge number of possible conformations, this is far from a perfect solution. To this end we hope to make future improvements, especially in dynamically forming hydrogen bonds during a FRODA run, and so further stabilize such nonlocal structures during a run.

CONCLUSION

This initial test of ZAMF gave very promising results. Of the initial test set of eight proteins, FRODA sampled to within an average RMSD of 3.7 Å of the experimentally determined structure and after refinement the predicted structures were within an average RMSD of 2.7 Å to the experimentally determined structure. Although these results are promising, we wish to continue exploring our method and expand it to full tests where we include fragment generation via the zipping method or from secondary structural prediction servers. This will help examine how robust our method is without having exact secondary structure. We are currently pursuing other paths for improving ZAMF to further automate and improve the accuracy produced.

The authors thank the Fulton High Performance Computing Initiative for computer time, and Mike Thorpe, Dan Farrell, Scott Menor, Kirill Speransky, M. Scott Shell, and Stephen Wells for help with FRODA, ZAM, and other insightful input.

REFERENCES

1. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–96.
2. Dill, K. A., S. B. Ozkan, …, V. A. Voelz. 2007. The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* 17:342–346.
3. Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15:285–289.
4. Kryshtafovych, A., C. Venclovas, …, J. Moult. 2005. Progress over the first decade of CASP experiments. *Proteins*. 61 (Suppl):225–236.
5. Dill, K. A., S. B. Ozkan, …, T. R. Weikl. 2008. The protein folding problem. *Annu. Rev. Biophys.* 37:289–316.
6. Zhang, Y., A. K. Arakaki, and J. Skolnick. 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*. 61 (Suppl):91–98.
7. Zhang, Y. 2008. Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* 18:342–348.
8. Baker, D. 2006. Prediction and design of macromolecular structures and interactions. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 361:459–463.
9. Rohl, C., C. Strauss, …, D. Baker. 2004. Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins*. 55:656–677.
10. Das, R., and D. Baker. 2008. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* 77:363–382.
11. Wu, G. A., E. A. Coutsias, and K. A. Dill. 2008. Iterative assembly of helical proteins by optimal hydrophobic packing. *Structure*. 16:1257–1266.
12. Inbar, Y., H. Benyamini, …, H. Wolfson. 2003. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics*. 19:158–168.
13. Kifer, I., R. Nussinov, and H. Wolfson. 2008. Constructing templates for protein structure prediction by simulation of protein folding pathway. *Proteins*. 73:380–395.
14. Fleming, P. J., H. Gong, and G. D. Rose. 2006. Secondary structure determines protein topology. *Protein Sci.* 15:1829–1834.
15. Kolodny, R., and M. Levitt. 2003. Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*. 68:278–285.
16. Schulten, K., P. L. Freddolino, …, M. Gruebele. 2008. Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.* 94:L75–L77.
17. DeBartolo, J., A. Colubri, …, T. Sosnick. 2008. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc. Natl. Acad. Sci. USA*. 106:3734–3739.
18. Nymeyer, H. 2009. Energy landscape of the trpzip2 peptide. *J. Phys. Chem. B*. 113:8288–8295.
19. Yoda, T., Y. Sugita, and Y. Okamoto. 2004. Comparisons of force fields for proteins by generalized-ensemble simulations. *Chem. Phys. Lett.* 386:460–467.
20. Zhou, R., B. J. Berne, and R. Germain. 2001. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA*. 98:14931–14936.
21. 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.
22. Rhee, Y. M., and V. S. Pande. 2003. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* 84:775–786.
23. Liwo, A., M. Khalili, and H. 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.
24. Ozkan, S. B., G. A. Wu, …, K. A. Dill. 2007. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. USA*. 104:11987–11992.
25. Huang, X., G. R. Bowman, and V. S. Pande. 2008. Convergence of folding free energy landscapes via application of enhanced sampling methods in a distributed computing environment. *J. Chem. Phys.* 128:205106–205126.
26. Periole, X., and A. E. Mark. 2007. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* 126:041903.
27. Paschek, D., H. Nymeyer, and A. E. García. 2007. 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.
28. Chen, J., W. Im, and C. L. Brooks, 3rd. 2006. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *J. Am. Chem. Soc.* 128:3728–3736.
29. Seibert, M. M., A. Patriksson, …, D. van der Spoel. 2005. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.* 354:173–183.
30. Nguyen, P. H., G. Stock, …, M. S. Li. 2005. Free energy landscape and folding mechanism of a beta-hairpin in explicit water: a replica exchange molecular dynamics study. *Proteins*. 61:795–808.
31. Felts, A. K., Y. Harano, …, R. M. Levy. 2004. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins*. 56:310–321.
32. Rao, F., and A. Caflisch. 2003. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* 119:4035–4042.
33. Sanbonmatsu, K. Y., and A. E. Garcia. 2002. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins*. 46:225–234.
34. Shell, M. S., S. B. Ozkan, …, K. A. Dill. 2009. Blind test of physics-based prediction of protein structures. *Biophys. J.* 96:917–924.
35. Wells, S., S. Menor, …, M. F. Thorpe. 2005. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* 2:S127–S136.
36. Lee, A., I. Streinu, and O. Brock. 2005. A methodology for efficiently sampling the conformation space of molecular structures. *Phys. Biol.* 2:S108–S115.
37. Amato, N. M., K. A. Dill, and G. Song. 2003. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.* 10:239–255.
38. Shehu, A., C. Clementi, and L. Kavraki. 2006. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins*. 65:164–179.
39. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.
40. Roitberg, A. E., A. Okur, and C. Simmerling. 2007. Coupling of replica exchange simulations to a non-Boltzmann structure reservoir. *J. Phys. Chem. B*. 111:2415–2418.
41. Fain, B., and M. Levitt. 2003. Funnel sculpting for in silico assembly of secondary structure elements of proteins. *Proc. Natl. Acad. Sci. USA*. 100:10700–10705.
42. Jacobs, D. J., A. J. Rader, …, M. F. Thorpe. 2001. Protein flexibility predictions using graph theory. *Proteins*. 44:150–165.
43. Thorpe, M. F., M. Lei, …, L. A. Kuhn. 2001. Protein flexibility and dynamics using constraint theory. *J. Mol. Graph. Model.* 19:60–69.
44. Hespenheide, B. M., A. J. Rader, …, L. A. Kuhn. 2002. Identifying protein folding cores from the evolution of flexible regions during unfolding. *J. Mol. Graph. Model.* 21:195–207.
45. Mamonova, T., B. Hespenheide, …, M. Kurnikova. 2005. Protein flexibility using constraints from molecular dynamics simulations. *Phys. Biol.* 2:S137–S147.
46. Pearlman, D. A., D. Case, …, P. Kollman. 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.

47. 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.

48. Clemente, C., A. E. Garcia, and J. Onuchic. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein L. *J. Mol. Biol.* 326:933–954.

49. Boden, M., Z. Yuan, and T. L. Bailey. 2006. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. *BMC Bioinformatics*. 7:68.

50. Cole, C., J. D. Barber, and G. J. Barton. 2008. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* 36(Web Server issue):W197–W201.

51. Raghava, G. P. S. 2002. APSSP2: a combination method for protein secondary structure prediction based on neural network and example based learning. CASP5 A-132.

52. Rost, B., and C. Sander. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584–599.

53. Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195–202.