Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Protein Structure Refinement of CASP Target Proteins Using GNEIMO Torsional Dynamics Method

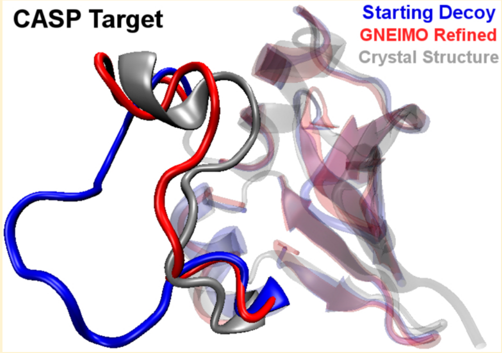
Adrien B. Larsen,[†] Jeffrey R. Wagner,[†] Abhinandan Jain,[‡] and Nagarajan Vaidehi*,[†]

[†]Division of Immunology, Beckman Research Institute of the City of Hope, 1500, E. Duarte Road, Duarte, California 91010, United States

[‡]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, United States

Supporting Information

ABSTRACT: A longstanding challenge in using computational methods for protein structure prediction is the refinement of low-resolution structural models derived from comparative modeling methods into highly accurate atomistic models useful for detailed structural studies. Previously, we have developed and demonstrated the utility of the internal coordinate molecular dynamics (MD) technique, generalized Newton−Euler inverse mass operator (GNEIMO), for refinement of small proteins. Using GNEIMO, the high-frequency degrees of freedom are frozen and the protein is modeled as a collection of rigid clusters connected by torsional hinges. This physical model allows larger integration time steps and focuses the conformational search in the low frequency torsional degrees of freedom. Here, we have applied GNEIMO with temperature replica exchange to refine low-resolution protein models of 30 proteins taken from the continuous assessment of structure prediction (CASP) competition. We have shown that GNEIMO torsional MD method leads to refinement of up to 1.3 Å in the root-mean-square deviation in coordinates for 30 CASP target proteins without using any experimental data as restraints in performing the GNEIMO simulations. This is in contrast with the unconstrained all-atom Cartesian MD method performed under the same conditions, where refinement requires the use of restraints during the simulations.



■ INTRODUCTION

Comparative modeling methods, also known as homology modeling or template-based modeling methods, are used widely to model protein structures. Significant improvement in the accuracy of homology models stemming from various advancements has been described in other publications.[1−8] However, even when using multiple templates, the resulting homology models can show significant deviation from the crystal structures, especially in certain local areas of the protein structure, depending on the sequence alignments. In order to be useful for functional analysis and drug design, these homology models must be refined further to higher accuracy. One of the outstanding problems in protein structure prediction is the lack of a consistent and reliable method for refinement of low resolution protein structural models to atomic level accuracy.[9]

Torsional Monte Carlo methods have been quite successful for protein structure refinement. However, limitations in the conformational search occur in the energy driven torsional Monte Carlo. These limitations can possibly be overcome by using the force driven molecular dynamics (MD) methods that enable going over energy barriers.[10] Therefore, an MD simulation offers an attractive force driven conformational search method that overcomes the pitfalls of Monte Carlo-based methods.[11] All-atom MD simulations, also known as Cartesian MD simulations, have shown limited success in protein structure refinement.[12−17] However, in combination with knowledge based potentials and/or restraints using experimental data, structural refinement has been achieved using all-atom MD simulations.[18−20] Mirjalili and Feig have shown that the use of restraints to the starting structure during MD gives better refinement than without. This is in agreement with the findings from the work of Shaw and co-workers.[13] Feig et al also showed that the MD ensemble averaged structures show better refinement than selecting one conformation from the ensemble.[20] In this paper, we focus on the use of MD methods for structure refinement without restraints.

The generalized Newton−Euler inverse mass operator (GNEIMO) method is an MD simulation method based on the use of internal coordinates. Torsional dynamics is one of many possible applications of the GNEIMO dynamics method.[21−23] The use of the GNEIMO torsional dynamics method combined with the temperature replica exchange (REXMD) method[25] has been demonstrated for refinement of protein homology models without the use of knowledge based restraints.[26] Treating the high frequency degrees of freedom as rigid using hard holonomic constraints, along with temperature based replica exchange, leads to efficient conformational sampling in the low frequency torsional degrees of freedom.

Received: August 16, 2013
Published: January 7, 2014

Table I. Extent of Refinement in the GDT and TM Scores from GNEIMO Compared with the Best Structure Submitted in CASP for Each Target[a]

	TR Decoys, All C-alpha								
	GDT_TS			TM-Score			RMSD		
target	start	best GNEIMO	best CASP	start	best GNEIMO	best CASP	start	best GNEIMO	best CASP
429	31.5	45.7	39.8	0.46	0.59	0.53	6.82	5.76	6.62
435	80.2	87.9	83.4	0.86	0.91	0.89	2.14	1.65	1.88
453	86.6	91.5	86.6	0.87	0.92	0.88	1.51	1.10	1.48
454	58.5	71.0	60.2	0.79	0.87	0.81	3.26	2.47	3.09
461	89.4	91.2	90.4	0.93	0.94	0.94	1.63	1.55	1.60
462	63.8	67.1	69.1	0.80	0.81	0.83	2.55	2.55	2.28
464	75.4	83.3	81.2	0.76	0.81	0.82	2.77	2.45	2.28
469	76.6	80.3	89.3	0.74	0.79	0.85	2.13	1.89	1.68
476	36.5	45.8	42.5	0.42	0.50	0.47	6.92	6.31	5.42
488	85.3	86.8	90.5	0.88	0.89	0.92	2.13	1.91	1.57
517	68.5	72.8	69.4	0.77	0.80	0.78	4.60	3.59	3.95
530	82.4	90.7	88.5	0.84	0.90	0.88	2.00	1.33	1.63
557	63.4	68.0	66.6	0.73	0.76	0.78	4.10	3.37	3.30
568	50.8	53.9	56.2	0.55	0.57	0.60	6.26	5.60	4.26
569	68.4	72.2	77.8	0.71	0.73	0.81	3.05	2.94	1.98
574	57.3	66.4	58.6	0.64	0.72	0.65	3.52	2.90	3.37
576	61.3	61.3	66.4	0.72	0.72	0.76	6.67	6.67	3.86
592	89.8	93.5	93.4	0.92	0.94	0.95	1.22	1.09	0.95
594	85.3	85.5	85.8	0.90	0.91	0.91	1.83	1.62	1.64
606	67.1	67.1	75.9	0.73	0.73	0.81	4.87	3.95	2.91
614	71.9	71.9	80.2	0.76	0.76	0.84	5.36	4.41	2.78
622	66.7	66.7	73.5	0.74	0.74	0.78	6.54	6.17	3.25
624	50.0	59.3	63.4	0.49	0.58	0.63	5.21	3.95	3.86
avg score	68.1	73.0	73.4	0.74	0.78	0.79	3.79	3.27	2.85
avg improvement		4.9	5.3		0.04	0.05		0.52	0.93

[a] The third column in each block shows the scores for the best structure submitted to CASP. Note that the best CASP structure came from different groups. The RMSD deviations have been calculated for the Cα atoms in angstroms.

In this paper, we have applied the GNEIMO-REXMD method for the refinement of 30 proteins from the list of target proteins released by the critical assessment of techniques for protein structure refinement and prediction (CASP) CASP8[27] and CASP9.[28] These 30 target proteins consist of both structure prediction and structure refinement categories from CASP8 and CASP9. Since the focus of this work was to examine the performance of GNEIMO-REXMD method for homology model refinement, we chose CASP targets for which the crystal structures were available at the time of our study. Hence we did not include CASP10 targets.[29] We have studied the extent of structure refinement that GNEIMO provides without using experimental data as restraints. Our ultimate goal is to examine if the GNEIMO torsional MD, in combination with torsional Monte Carlo method with and without experimental data is capable of protein structure refinement.

■ COMPUTATIONAL METHODS

Generalized Newton−Euler Inverse Mass Operator (GNEIMO)—Constrained Dynamics Method. Details of the GNEIMO method can be found in multiple publications.[21,30,31] Briefly, GNEIMO is a constrained MD method using internal coordinates, where the high frequency degrees of freedom are held rigid using holonomic constraints and the protein is modeled as a collection of user-defined rigid bodies known as "clusters" connected by flexible hinges. The hinges can be modeled with one to six degrees of freedom. Clusters can range in scale from single atoms to helices to whole domains of proteins, as chosen by the user. The equations of motion in internal coordinates are coupled, and the computational cost of solving the coupled equations of motion in internal coordinates scales as cubic power of the number of degrees of freedom.[31,32] Using the GNEIMO algorithm, however, the computational cost scales linearly with the number of degrees of freedom that enables the use of torsional MD computationally feasible for protein simulations.[21,22] Other groups have used the GNEIMO algorithm for torsional MD simulations and for NMR structure refinement.[14,33−35] We have incorporated various advanced internal coordinate dynamics techniques to make the current implementation of GNEIMO a robust MD technique for long time scale simulations.[23,30] We have also demonstrated the use of GNEIMO for structure refinement of small proteins[26] and for mapping the domain motion in proteins.[36] In this paper we have used the GNEIMO torsional MD that restricts internal motion to torsional angles for refining protein homology models.

All-Torsion GNEIMO MD Protocol for Homology Model Refinement. The GNEIMO-based protocol for protein structure refinement combines the GNEIMO torsional MD method with the REXMD method for extended conformational search in torsional space. The GNEIMO-REXMD protocol used for refinement is an adaptation of the protocol previously derived[26] for protein structure refinement. Briefly, the GNEIMO constrained MD simulations were carried out using the GNEIMO code[23,24] using the AMBER99SB force field[37] with the generalized Born/surface area (GB/SA) OBC implicit solvation model,[38] an interior dielectric value of 1.5 for the solute, and exterior dielectric constant of 78.3 for the solvent. We used a solvent probe radius of 1.4 Å for the nonpolar solvation energy component of GB/SA. The nonbonded forces were switched off at a cutoff radius of 20 Å. GNEIMO MD simulations were performed using all torsional degrees of freedom and at constant temperature using the Nose-Hoover method,[22] a Lobatto integrator,[23] and an integration time step size of 5 fs. We added the temperature replica exchange MD (REXMD)[25] algorithm to the GNEIMO MD method to enhance conformational sampling. GNEIMO all-torsion REXMD was then performed using 32 replicas over the 310−415 K range of temperatures, and the temperature sorting was done based on the Metropolis algorithm every 5 ps.

We have studied refinement of CASP targets from two categories: (1) the "refinement category" where a given decoy is to be refined and (2) the structure prediction category where only the sequence of amino acids in the target protein is provided. In this paper, we have studied 23 proteins of various sizes from the CASP8 and CASP9 refinement category and 7 proteins from the structure prediction category. The decoy structures for these 23 proteins were downloaded from the CASP Web site: www.predictioncenter.org. These structures were first subjected to all-atom conjugate gradient minimization using the "sander" program and the AMBER99SB force field.[37] A total simulation time of 15−100 ns (for each replica) for 32-replicas GNEIMO-REXMD simulations were performed for each target.

Structure Preparation of the Protein Targets. The list of 23 target proteins from the structure refinement category (TR) and seven target proteins from structure prediction category (T0) selected from the CASP8 and CASP9 is shown in Table I. The CASP10 targets were not released when we started this work. In addition, we wanted to validate the use of GNEIMO method for known targets and derive a protocol before we applied it to unknown targets. The starting decoys for the structure refinement targets were taken from the CASP Web site. We performed GNEIMO-REXMD simulations with 32 replicas for each target. For the targets from the structure prediction category, we derived homology models using the MODELER[39] method. The template structures for the MODELER program were chosen by using the PDB sequence query search[40] for sequences of 30−80% identity to the target. We removed the structure of the target protein and its close homologues (that were published after the respective CASP competitions) from the template structure search, to avoid any bias. One hundred models for each target were generated using MODELER, and these models were then clustered into five groups. The best representative structure out of the five groups, as scored by procheck G-factor,[41] was chosen to be the starting decoy for refinement for that target. Minimization on each decoy was run using the "sander" utility in the AMBER suite and with the Amber FF99SB force field, followed by GNEIMO-REXMD simulations.

Calculation of RMSD, Percentage Native Contacts, and GDT Scores. We have calculated several metrics to assess the native likeness of the conformations sampled by GNEIMO-REXMD. These are the standard metrics used in the CASP assessment. We have calculated the root-mean-square deviation in coordinates (RMSD) of the backbone atoms to the X-ray and NMR structures. The RMSD was calculated using snapshots from the combined REXMD trajectory of all replicas. To determine whether the refinement in the model structure came from the secondary structure regions or the loop regions, we further calculated the RMSD of the backbone atoms in the secondary structure region (as defined by the set of residues which are in helix or sheet in the native conformation) and the RMSD of the backbone atoms for the whole structure including the loops and termini, both with respect to native structures. We used the package known as "MDAnalysis" for the RMSD calculations.[42]

To measure the extent of refinement in the overall packing and fold of a protein, we compared the percent of native contacts made in the GNEIMO simulation trajectories with those in the native crystal and NMR structures. We calculated the $N \times N$ matrices consisting of pairwise $C\alpha(i)-C\alpha(j)$ atom distances for the native structures, and for the whole trajectories of GNEIMO-REXMD, where N is the number of residues in the protein, and i and j are residue indices. A pairwise $C\alpha(i)-C\alpha(j)$ distance smaller than 8 Å was considered a contact and given an index value of 1. Hence, the calculation of the contact map results in the construction of an $N \times N$ contact matrix consisting of 0 s (atom pairs farther than the 8 Å distance cutoff or within the 4-residue neighbor cutoff in sequence) and 1's (an atom pair within the 8 Å cutoff and more than 4 residues apart in sequence). We then considered each $C\alpha$ pair in the simulation snapshots to be a contact, if the distance between both the atoms were within 0.5 Å of the same distance in the contact map of the native structure.[43] We calculated the percentage native contacts as (number of identical contacts between native and decoy/total number of contacts seen in the native structure). MDAnalysis was used to calculate the pairwise distances needed for determination of percent native contacts.

Calculations of Measures Used for Structure Prediction and Refinement Assessment in CASP. The GDT and TM scores are commonly used metrics in the CASP assessments since they are more stringent than RMSD. GDT is defined as the average number of aligned $C\alpha$ atoms that fit under a distance-to-native cutoff for four different cutoff distances. The most often used set of cutoff values is {8, 4, 2, 1} (Å) and is referred to as GDT_TS. The TM-score was developed to correlate well with human-expert assessment of protein model quality and to address the limitations of other scores such as RMSD and GDT.[44] The program "MaxCluster" (www.sbg.bio.ic.ac.uk/~maxcluster) was used to calculate the GDT_TS and TM-Scores.

Reference Structures. The crystal and/or NMR structures used as reference structures for the calculation of the above-described metrics were downloaded from the PDB Web site (www.rcsb.org) corresponding to each target as indicated by CASP. If the native structure was deposited as an NMR ensemble, we used the top ranked NMR structure as the reference. The missing residues in the native structure were not considered during scoring calculations.

■ RESULTS AND DISCUSSION

Protein Structure Refinement Category. Table I lists the GDT, the TM score, and the RMSD for the $C\alpha$ atoms of the best structures from the GNEIMO-REXMD trajectories for 23 different CASP targets in the refinement category. These target names start with the letters "TR". The size of the refinement targets range from 63 to 192 amino acids. The GDT and the TM scores for the best structure from GNEIMO improved in comparison to the starting decoy for 19 out of 23 proteins. The GDT scores showed an increase of up to 14.0 points while the increase in TM score is up to 0.13. The extent of refinement by GNEIMO is comparable to the best structure submitted to

Table II. Extent of Refinement in the GDT and the TM Score of the Best Structure from the GNEIMO-REXMD Trajectories for the CASP Structure Prediction Targets[a]

TO MODELER Decoys, All C-alpha									
	GDT_TS			TM-Score			RMSD		
Target	start	best GNEIMO	best CASP	start	best GNEIMO	best CASP	start	best GNEIMO	best CASP
T0387	85.7	89.3	95.5	0.88	0.90	0.94	1.95	1.43	1.01
T0453	80.3	83.0	87.1	0.82	0.84	0.89	3.75	1.85	1.47
T0469	82.0	86.5	73.4	0.79	0.83	0.74	1.93	1.88	2.47
T0472	89.5	89.7	61.8	0.93	0.93	0.76	1.21	1.06	2.68
T0488	71.8	75.3	86.0	0.75	0.79	0.87	4.60	3.54	1.97
T0492	82.3	92.7	85.8	0.82	0.91	0.87	1.67	1.16	1.70
T0554	67.1	73.7	32.3	0.80	0.84	0.44	3.40	2.65	8.31

[a] The RMSD deviations have been calculated for the Cα atoms in angstroms.

CASP for each target. The average increase in GDT score over all the 23 targets is 4.9, 0.04 in TM score, and 0.52 Å in RMSD. This was a modest improvement for 19 out of 23 targets and in 4 cases there was no refinement from the starting model. To understand if there is a correlation between the extent of structure refinement and the secondary structure content, we calculated the secondary structure content of all the TR targets. We observed more than a 5.0 point improvement in the GDT scores for the TR429, TR435, TR453, TR454, TR464, TR476, TR530, TR557, TR574, and TR624 targets, and these proteins showed at least 55% secondary structure content. Some of the refinement targets for which there was little to no refinement by GNEIMO had less than 40% secondary structure content.

TR462, a target that showed little improvement with GNEIMO, is a two-domain protein connected by a linker region. Although the overall refinement is 3.3 in GDT score, we observed refinement of 5.7 and 6.2 in the GDT scores for the individual domains. The lack of refinement in the overall structure came from the linker region. Other targets such as TR576, TR594, TR606, TR614, and TR622 showed no improvement in GDT scores. TR614 has a loop that is 25 residues long, and TR606 has two loops that are 10 and 15 residues long that showed no improvement for GNEIMO simulations. Figure S1 of the Supporting Information shows the extent of the refinement in the secondary structure regions in the target protein structures.

Refinement for the Targets in the Protein Structure Prediction Category. Apart from the refinement category, we also predicted the structures for the CASP8 and CASP9 targets in the structure prediction category denoted as the "T0" targets. We predicted structures of the seven proteins listed in Table II. As described in the Computational Methods section, we first derived a homology model for each target using the MODELER program starting from the amino acid sequence. We avoided using crystal structures that were published after the corresponding CASP8 or CASP9 assessment dates as templates for the homology models. Starting from the homology models, we performed the GNEIMO-REXMD simulations for refinement of the model. Table II shows the extent of refinement (GDT and TM scores) yielded by just the refinement cycle using the GNEIMO simulations. All the structures predicted were within 4 Å RMSD from the respective crystal structures except for T0488. The average improvements of 4.5, 0.04, and 0.7 Å in GDT, TM scores, and RMSD, respectively, were obtained using GNEIMO-REXMD refinement simulations. The RMSD of the structural models generated by MODELER ranges from 1 to 9 Å. In every case, we observed substantial refinement in the structures compared to the starting decoy from MODELER.

Analysis of the Extent of Refinement. Figure 1 shows the contact map for targets with substantial refinement (T0453, TR429, TR454, TR530) and one target that did not show any refinement (TR576). These figures show the absolute distance between each residue in the GNEIMO refined model and the same residue in the corresponding crystal structure for various targets. The deeper the red color, the farther it is from the crystal structure. The cartoon picture of the backbone of the initial decoy, the GNEIMO refined, and the experimental structures are also shown for these targets. The white regions in the contact map are closer to the native structure, and the dark red regions are farther from the corresponding native structures. Figure 1A for target T0453 shows substantial improvement in the packing of the loop structure with the rest of the protein structure upon GNEIMO refinement. The long-range inter-residue contacts for residues in the loop region between 30 and 40 with residues that are in the core of the protein between 40 and 60 (shown in dark red in contact map for the starting decoy) improve from 14 to 16 Å to 2−4 Å as seen in the contact map for the refined structure. These regions are marked with dashed line rectangles in the figure. This is an example where most of the refinement is in the packing of the loop with the secondary structure core of the rest of the protein. Figure 1B for TR429 shows that residue ranges 25−31 and 36−42 form a β-sheet that is missing in the starting decoy structure (see dashed line rectangles in Figure 1B). The three-strand β-sheet motif that ranges from residues 20−60 improves its packing to the rest of the protein, upon GNEIMO refinement. Figure 1C shows that the two helices in the target TR454 are already formed in the initial decoy. Refinement in this situation requires improved packing of these two helices shown in dashed rectangles in Figure 1C. As seen in Figure 1C, TR454 shows the long-range helix packing into a more nativelike structure, resulting in an RMSD of 2.47 Å in coordinates of the backbone atoms to the crystal structure. GNEIMO-REXMD refinement of the TR530 target leads to the proper folding of the N-terminal region, as shown in Figure 1D (dashed line rectangles). Refinement is also seen in other long-range contacts throughout the molecule, due to slight improvements in local packing.

The protein TR576 showed no refinement. TR576 has a substantial β-sheet content as well as long loop regions, and the carboxy terminus region of TR576 needed substantial refinement. This structure was deemed to be of high difficulty by the CASP assessment team, due to crystal contacts.[9] The starting decoy is misfolded into a partial antiparallel β-sheet while it is a

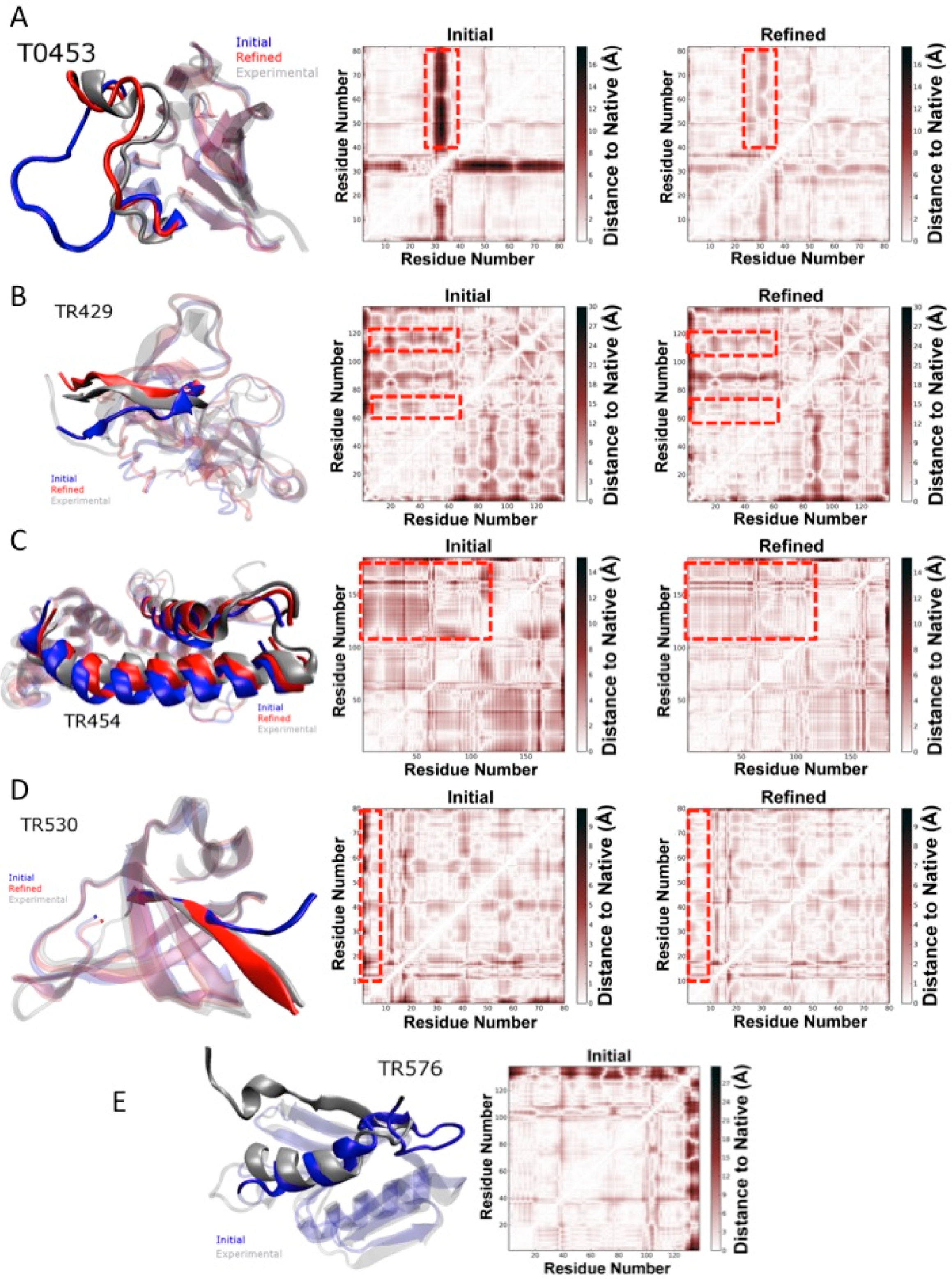


Figure 1. (A−D) Refinement of CASP targets with different types of secondary structure as shown in the figures and corresponding distance-to-native map. The distance to native map shows how far each residue is corresponding to the native structure. (1A) Example of refinement of loop structure; (1B) refinement of β-sheet packing; (1C) packing of α-helices; (1D) β-sheet growth; (1E) example of a structure that was not refined by GNEIMO.

parallel β-sheet in the native structure. This could not be refined by GNEIMO possibly due to a high energy barrier in refolding this region. There are waters stabilizing this region observed in the crystal structure. We did not use explicit water in these GNEIMO simulations, which may play an important role in stabilizing the loop structures.

Many of the targets studied here show better refinement with GNEIMO simulations than the best structure submitted for the

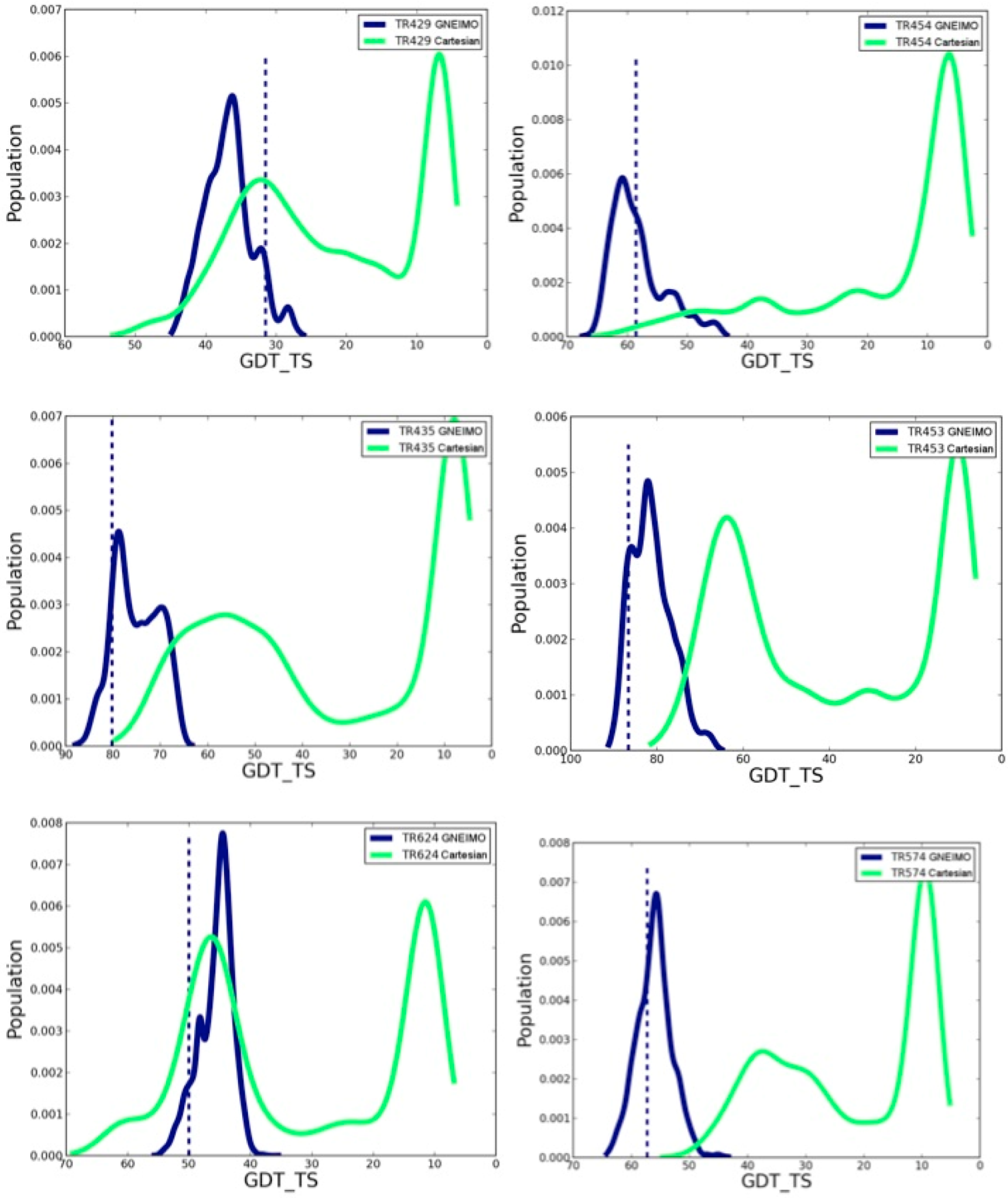


Figure 2. Population distribution of the ensemble generated from the GNEIMO simulations compared to the Cartesian simulations for various refinement CASP targets. The dotted line is the GDT_TS score for the starting decoy.

CASP assessment. It should be emphasized that we have compared the best structure by RMSD in coordinates generated by GNEIMO-REXMD simulations to the native structure and not utilized energy or scoring functions to pick the best structure. However, the GNEIMO method can be combined with any energy function or scoring function for picking the best structure.[11,18,19,45]

Enrichment of Nativelike Structures in the GNEIMO-REXMD Trajectories. The chances of identifying the closest to native structure as the best scoring structure are improved if there is a substantial population of near-native conformations compared to the starting decoy generated during the GNEIMO-REXMD simulations. Therefore, we calculated the fraction of the population from GNEIMO-REXMD simulations that are closer to the native structure compared to the starting decoy to assess the enrichment of nativelike structures in the GNEIMO-REXMD trajectories. Figure 2 shows the population histogram with respect to GDT score and TM score for some

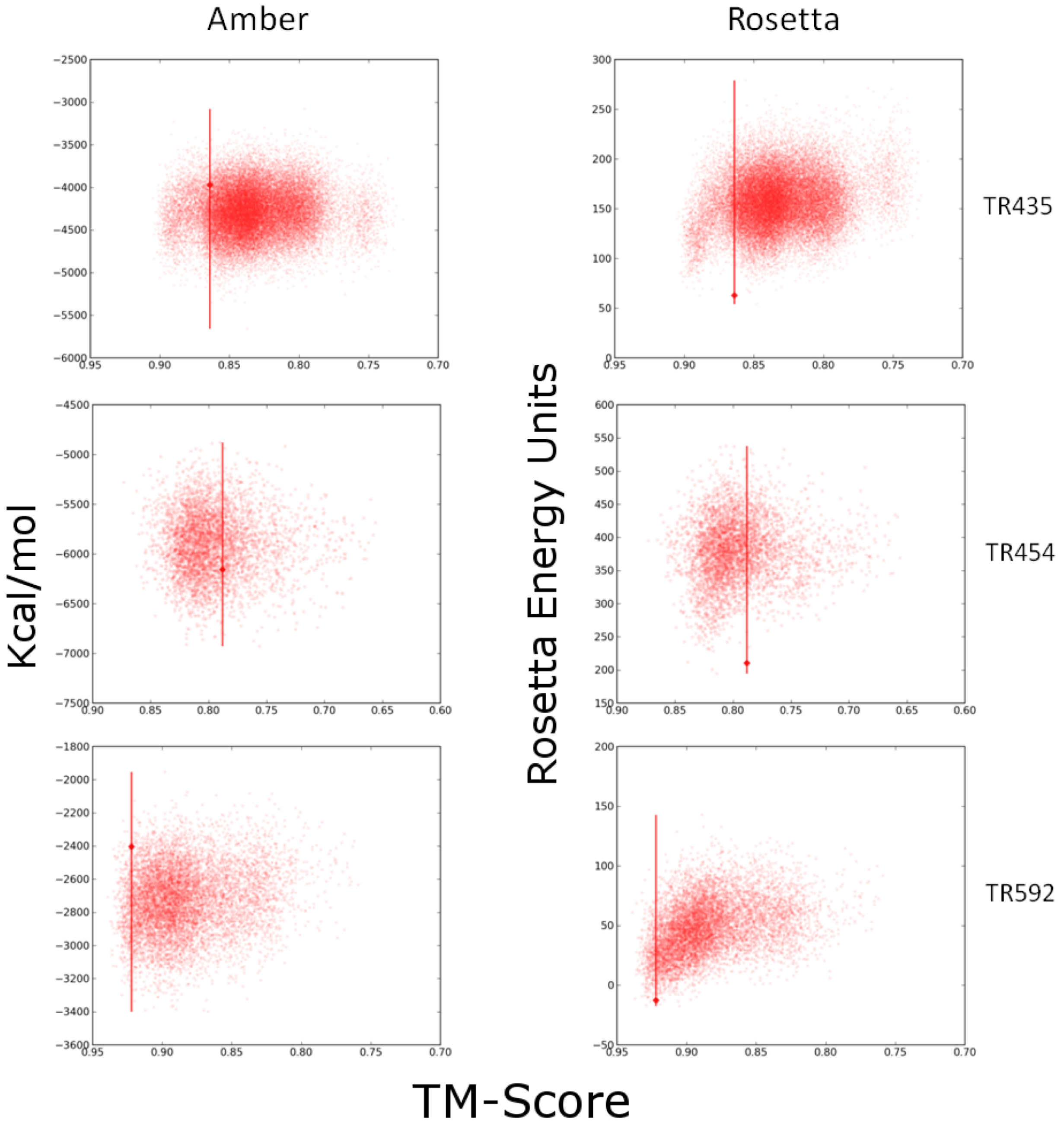


Figure 3. Potential energies of the conformations generated in the GNEIMO-REXMD simulation trajectories calculated using (a) AMBER99SB all atom forcefield and (b) Rosetta energy function.

targets, and Figure S2 of the Supporting Information shows the same for all the targets.

It is seen that, for TR429 and TR454, a significant (over 50%) population shift occurs toward the native structure compared to the starting decoy (denoted by the dotted line in Figure 2) in both GDT and TM scores. Structurally, both of these proteins are notable in that they have over 50% secondary structure content. Both TR568 and TR624 show a small percentage (about 10−20%) of the population getting refined. The targets TR576, TR606, TR614, and TR622 show little to no population shift toward the native structure. These structures have less than 40% secondary structure content and have large loops, thus leading to poor refinement. Thus, GNEIMO shows effective refinement of secondary structure regions and their packing with the failures occurring in the refinement of the loop regions. We are exploring the use of GNEIMO simulations with side chain replacement methods to improve the refinement of loop regions.

Comparison of GNEIMO performance to Cartesian MD Method. In this section we compare the effectiveness of the conformational sampling of GNEIMO method to Cartesian MD method. However, we are not examining the effectiveness of energy functions that are used for selecting the best structure for cases where the experimental structure is unknown. Our future goal is to combine GNEIMO with other refinement methods, namely those methods such as torsional Monte Carlo methods, that complement the torsional dynamics based conformational search of GNEIMO.

Figure 2 shows the comparison of the population of structures in the GNEIMO ensemble that get refined with respect to the starting decoy structures as measured by the GDT scores, compared to the corresponding populations in the Cartesian MD. The MD simulations for both GNEIMO and Cartesian MD were done with the identical forcefield and GBSA solvation. The dotted line in the figure shows the position of the starting decoy. Figure 2 shows that the relative proportion of GNEIMO ensemble that gets refined is more than the population of refined structures from Cartesian simulations. This shows that the conformational sampling afforded by sampling torsional angles using the same forcefield is more effective in getting closer to the native structures than the Cartesian all-atom simulations. Shaw and co-workers[13] have tested the ability of several microseconds of Cartesian MD simulations to refine homology models for 25 CASP refinement targets, 21 of which are common to our study. They observed that the long Cartesian MD simulations lead to unraveling of the homology model away from the crystal or NMR structures. Using the same forcefield we have shown in this paper that GNEIMO torsional dynamics method leads to more refinement than the Cartesian MD simulations. Using an energy function that would preserve and funnel toward the native structure is a critical component for structure refinement. In the next section, we compare the all-atom AMBER energy function that we have used with GNEIMO MD in this study to the knowledge based energy function in Rosetta.[46]

AMBER versus Rosetta Energy Function. Scoring functions are important in selecting the most refined structure from the ensemble of conformations generated during the GNEIMO-REXMD simulations. In this paper, we have not addressed this issue. Briefly, we have calculated the all-atom AMBER energies[37] and energies from the Rosetta energy function[11] for all the conformations generated in the GNEIMO-REXMD trajectories of three target proteins. The Rosetta energy function is based largely on the CHARMM energy function with additional knowledge based hydrogen bond terms.[47] As shown in Figure 3, the Rosetta energy function showed a more funnel like character for some of the targets, i.e., the near-native structures showed the lowest energy. Thus, use of the Rosetta energy function could improve the selection of the best refined structure. However, for many other targets, both the AMBER and Rosetta energy functions did not show a funnel like behavior. In the next stage of our study, we will examine many other knowledge based energy functions[18,48] for rescoring the conformations generated by GNEIMO-REXMD simulations. Also, we will explore the use of a force field derived from the Rosetta potential energy function for driving the GNEIMO dynamics.

Assessment of GNEIMO as a Refinement Tool. An advantage of using the GNEIMO method is the time required to perform each sampling. By taking stable time steps of 10 fs, GNEIMO combined with REXMD is able to explore more regions of conformational space in the same number of processor cycles compared to Cartesian simulations. While the targets in this study were run for up to 100 ns (for all replicas combined), existing CASP teams which replace Cartesian MD with GNEIMO could simulate about 1 order of magnitude longer in the same clock time. Unlike torsional Monte Carlo method where the moves are random and scored by energy, the forces govern the moves in GNEIMO torsional MD. Performing conformational search in the torsional degrees of freedom appears to focus the search in the low frequency degrees of freedom. Coupling GNEIMO torsional MD with REXMD provides enough thermal energy to overcome barriers that can arise from the stiffness in the dynamic model from freezing high frequency degrees of freedom. Giving such high thermal energy to all atoms in the Cartesian all-atom dynamics can however result in an unraveling of structured regions in the starting decoy.

The GNEIMO approach is also highly extensible. Some Monte Carlo based methods that were used in CASP restrict sampling of certain regions or attempt to rigidly dock individual domains of the same protein. The ability to perform dynamics of coarsened bodies is inherent within GNEIMO, and the generalized coordinate system can naturally incorporate constraints into the equations of motion to rigidify or free any desired degree of freedom. Further, GNEIMO is not restricted to any specific force field but rather has a modular design with an extensible interface class for any force field that can be wrapped to fit a template. Numerical integration methods for time propagation are also optional modules in GNEIMO, and methods, which rely on torsional Monte Carlo sampling (like Rosetta), can use GNEIMO to directly sample coordinates in phase space based on any definable coordinate system.

In this paper, we report some of the promising developments on the application of torsional molecular dynamics method (GNEIMO) to structure refinement of CASP target proteins. We have applied brute force torsional MD without any restraints on any part of the structure from known structural information to refining homology models. The torsional MD refinement yields results that are better compared to the all-atom MD simulations performed under the same conditions. There is still much progress and issues that need to be addressed to improve GNEIMO as a refinement tool. As observed in Table I, the extent of refinement by GNEIMO is the same whether the starting decoy is of low resolution (greater than 5 Å) or of high resolution (less than 3 Å). One possibility is to test the extent of refinement using different clustering schemes in GNEIMO in combination with side chain refinement methods. The use of a force field tailored for protein structure prediction can also improve refinement with GNEIMO. If distance restraints are available from known experimental data, refinement can be improved at low resolution.[12,13] Presently, we are working on all these aspects to make GNEIMO a robust and generic refinement tool.

CONCLUSIONS

We have shown that the GNEIMO-REXMD simulation technique leads to refinement of up to 1.3 Å for 30 CASP target proteins starting from their homology models of variable resolution. GNEIMO method leads to refinement for 21 out of 23 refinement targets although the average refinement for 23 targets is 4.0 in GDT score, 0.04 in TM score, and 0.5 Å in RMSD in coordinates. These torsional MD simulations were done without using any experimental data. Significantly, we did observe that GNEIMO with REXMD simulations enable focused conformational sampling in the low frequency torsional space that is essential for structure refinement. However, the overall extent of refinement was modest and needs further improvement. Further testing of this method, applying residue-based distance restraints obtained from experiments and testing a suitable energy function that provides identification of the native structure, is ongoing. Additionally Feig and co-workers have demonstrated that an ensemble average shows better refinement than rather than a single structure from MD simulations.[20] Further testing of GNEIMO ensembles using an ensemble average could provide better refinement than seen in this study. Enhancement in side chain sampling can be obtained by combining the GNEIMO simulations with side chain reassignment from a rotamer library. Our ultimate goal is to combine GNEIMO torsional dynamics with a torsional Monte Carlo method and test the method in a future CASP in the refinement category.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Results demonstrating the extent of refinement in the secondary structure regions of the targets, as well as the population of refined structures in the GNEIMO MD trajectories. This information is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: (626)-301-8408. Fax: (626)-301-8186. E-mail: nvaidehi@coh.org.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding for this work was provided by NIGMS R01-GM082896. We thank Simbios, the NIH Center for Biomedical Computation at Stanford University, for providing us with the GB/SA solvation module. Part of the research described in this paper was performed at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under contract with the National Aeronautics and Space Administration (NASA). A.B.L. thanks the California Institute of Regenerative Medicine for fellowship.

■ REFERENCES

(1) Webb, B.; Sali, A. *Protein structure modeling with MODELLER*; Methods in Molecular Biology, Humana Press, 2013; in press.

(2) Dunbrack, R. L. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.* **2006**, *16*, 374−384.

(3) Wu, S.; Zhang, Y. Protein Structure Prediction. *Bioinformatics: Tools Appl.* **2009**, *11*, 225−242.

(4) Sippl, M. J.; Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins* **1992**, *13*, 258−271.

(5) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. ICM: a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(6) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **2003**, *31*, 3381−3385.

(7) Misura, K. M.; Chivian, D.; Rohl, C. A.; Kim, D. E.; Baker, D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5361−5366.

(8) Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protocols* **2010**, *5*, 725−738.

(9) MacCallum, J. L.; Pérez, A.; Schnieders, M. J.; Hua, L.; Jacobson, M. P.; Dill, K. A. Assessment of protein structure refinement in CASP9. *Proteins−Struct. Funct. Bioinf.* **2011**, *79*, 74−90.

(10) Raman, S.; Vernon, R.; Thompson, J.; Tyka, M.; Sadreyev, R.; Pei, J.; Kim, D.; Kellogg, E.; DiMaio, F.; Lange, O.; Kinch, L.; Sheffler, W.; Kim, B. H.; Das, R.; Grishin, N. V.; Baker, D. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins−Struct. Funct. Bioinf.* **2009**, *77*, 89−99.

(11) Das, R.; Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363−382.

(12) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* **2010**, *18*, 923−933.

(13) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins−Struct. Funct. Bioinf.* **2012**, *80*, 2071−2079.

(14) Chen, J.; Im, W. Brooks III. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comput. Chem.* **2005**, *26*, 1565−1578.

(15) Fan, H.; Mark, A. E. Refinement of homology based protein structures by molecular dynamics simulation techniques. *Protein Sci.* **2004**, *13*, 211−220.

(16) Floudas, C. A. Computational methods in protein structure prediction. *Biotechnol. Bioeng.* **2007**, *97*, 207−213.

(17) Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. Molecular dynamics in the endgame of protein structure prediction. *J. Mol. Biol.* **2001**, *313*, 417−430.

(18) Summa, C. M.; Levitt, M. Near Native Structure Refinement Using in vacuo Energy Minimization. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3177−3182.

(19) Zhang, J.; Liang, Y.; Zhang, Y. Atomic level protein structure refinement using fragment guided molecular dynamics conformation sampling. *Structure* **2011**, *19*, 1784−1795.

(20) Mirjalili, V.; Feig, M. Protein structure refinement through Structure selection and averaging from molecular dynamics ensembles. *J. Chem. Theory Comput.* **2013**, *9*, 1294−1303.

(21) Jain, A.; Vaidehi, N.; Rodriguez, G. A fast recursive algorithm for molecular dynamics simulation. *J. Comput. Phys.* **1993**, *106*, 258−268.

(22) Vaidehi, N.; Jain, A.; Goddard, W. A. Constant Temperature Constrained Molecular Dynamics: The Newton−Euler Inverse Mass Operator Method. *J. Phys. Chem.* **1996**, *100*, 10508−10517.

(23) Wagner, J. R.; Balaraman, G. S.; Niesen, M. J.; Larsen, A. B.; Jain, A.; Vaidehi, N. Advanced techniques for constrained internal coordinate molecular dynamics. *J. Comput. Chem.* **2013**, *34*, 904−914.

(24) Balaraman, G. S.; Park, I. H.; Jain, A.; Vaidehi, N. Folding of small proteins using constrained molecular dynamics. *J. Phys. Chem. B* **2011**, *115*, 7588−7596.

(25) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.

(26) Park, I. H.; Gangupomu, V.; Wagner, J.; Jain, A.; Vaidehi, N. Structure refinement of protein low resolution models using the GNEIMO constrained dynamics method. *J. Phys. Chem. B* **2012**, *116*, 2365−2375.

(27) Tress, M. L.; Ezkurdia, I.; Richardson, J. S. Target domain definition and classification in CASP8. *Proteins−Struct. Funct. Bioinf.* **2009**, *77*, 10−17.

(28) Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 target classification. *Proteins−Struct. Funct. Bioinf.* **2011**, *79*, 21−36.

(29) Nugent, T.; Cozzetto, D.; Jones, D. T. Evaluation of predictions in the CASP10 model refinement category. *Proteins−Struct. Funct. Bioinf.* **2013**, DOI: 10.1002/prot.24377.

(30) Jain, A. *Robot and Multibody Dynamics: Analysis and Algorithms*; Springer, 2010.

(31) Tobias, D. J.; Brooks, C. L., III Molecular dynamics with internal coordinate constraints. *J. Chem. Phys.* **1988**, *89*, 5115−5127.

(32) Mazur, A. K.; Abagyan, R. New methodology for computer-aided modeling of biomolecular structure and dynamics. 1. Non-cyclic structures. *J. Biomol. Struct. Dyn.* **1989**, *6*, 815−832.

(33) Chun, H. M.; Padilla, C. E.; Chin, D. N.; Watanabe, M.; Karlov, V. I.; Alper, H. E.; Soosaar, K.; Blair, K. B.; Becker, O. M.; Caves, L. S. D. MBOND: A multibody method for long time molecular dynamics simulations. *J. Comput. Chem.* **2000**, *21*, 159−184.

(34) Schwieters, C. D.; Clore, G. M. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.* **2001**, *152*, 288−302.

(35) Flores, S. C.; Sherman, M. A.; Bruns, C. M.; Eastman, P.; Altman, R. B. Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1247−1257.

(36) Gangupomu, V. K.; Wagner, J. R.; Park, I.; Jain, A.; Vaidehi, N. Mapping conformational dynamics of proteins using torsional dynamics simulations. *Biophys. J.* **2013**, *104*, 1999−2008.

(37) Case, D. A.; Cheatham, T. E.; Darden, T., III.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A., Jr.; Simmerling, C.; Wang, B.; Woods, R. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(38) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins−Struct. Funct. Bioinf.* **2004**, *55*, 383−394.

(39) Eswar, N.; Eramian, D.; Webb, B.; Shen, M. Y.; Sali, A. Protein structure modeling with MODELLER. *Methods Mol. Biol.* **2008**, *426*, 145−159.

(40) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acid. Res.* **2000**, *28*, 235−242.

(41) Engh, R. A; Huber, R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.* **1991**, *47*, 392−400.

(42) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.

(43) Gromiha, M. Inter-residue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.

(44) Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins−Struct. Funct. Bioinf.* **2004**, *57*, 702−710.

(45) Bordner, A. J. Forcefields for homology modeling. *Methods Mol. Biol.* **2012**, *857*, 83−106.

(46) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucl. Acid. Res.* **2004**, *32*, 526−531.

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

(48) Yang, Y.; Zhou, Y. Specific interactions for *ab* initio folding of protein terminal regions with secondary structures. *Proteins−Struct. Funct. Bioinf.* **2008**, *72*, 793−803.