Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling

Jian Zhang,[1] Yu Liang,[2] and Yang Zhang[1,3,*]
[1]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Mathematics and Computer Science, Central State University, Wilberforce, OH 45384, USA
[3]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
*Correspondence: zhng@umich.edu
DOI 10.1016/j.str.2011.09.022

SUMMARY

One of critical difficulties of molecular dynamics (MD) simulations in protein structure refinement is that the physics-based energy landscape lacks a middle-range funnel to guide nonnative conformations toward near-native states. We propose to use the target model as a probe to identify fragmental analogs from PDB. The distance maps are then used to reshape the MD energy funnel. The protocol was tested on 181 benchmarking and 26 CASP targets. It was found that structure models of correct folds with TM-score >0.5 can be often pulled closer to native with higher GDT-HA score, but improvement for the models of incorrect folds (TM-score <0.5) are much less pronounced. These data indicate that template-based fragmental distance maps essentially reshaped the MD energy landscape from golf-course-like to funnel-like ones in the successfully refined targets with a radius of TM-score ∼0.5. These results demonstrate a new avenue to improve high-resolution structures by combining knowledge-based template information with physics-based MD simulations.
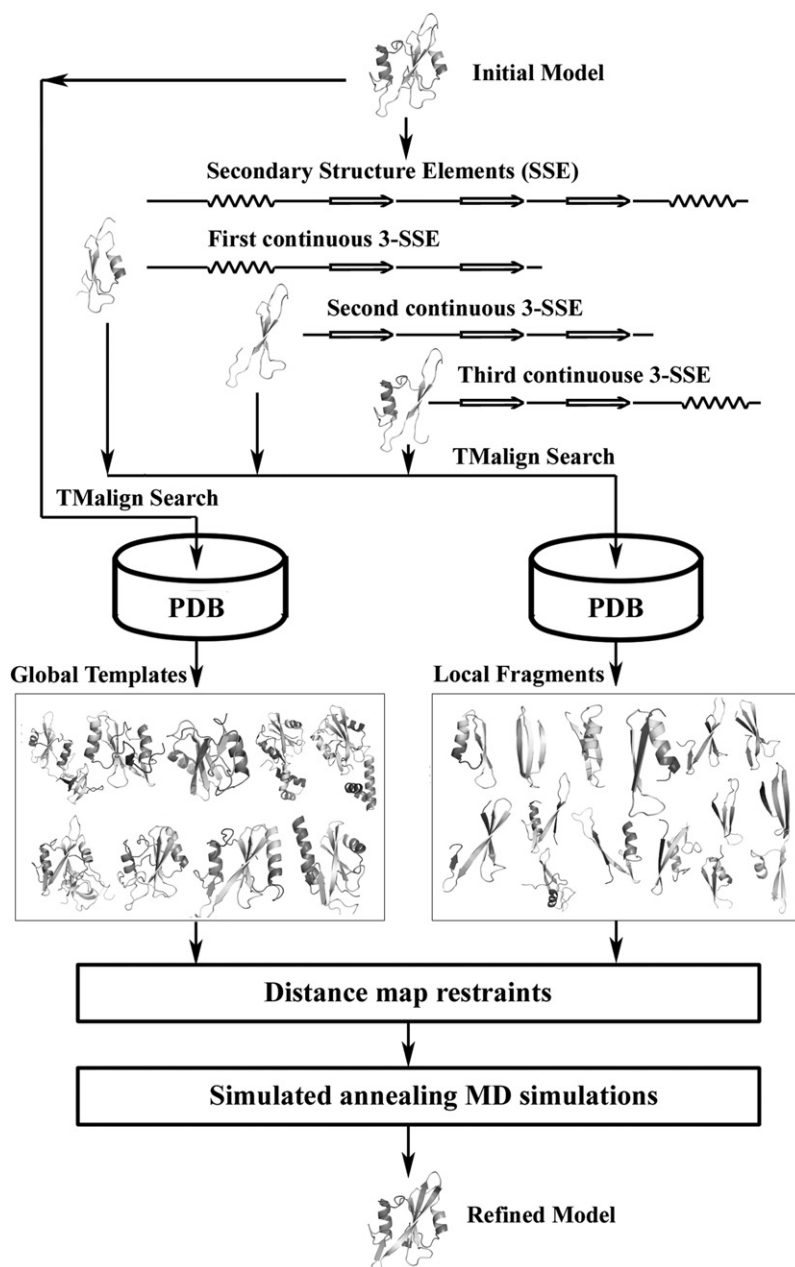
INTRODUCTION

Template based modeling (TBM) represents the most accurate method in protein structure prediction. In traditional TBM, the structural model is built by aligning the query sequence to a single protein template and then copying the structure information from the template in the aligned regions (Sali and Blundell, 1993). Thus, the final structural models are often closer to the template than to the native structure (Tramontano and Morea, 2003). In the contemporary TBM, multiple templates are often identified through metathreading techniques (Fischer, 2003; Ginalski et al., 2003; Wu and Zhang, 2007), and full-length models are built by combining the structural fragments/restraints from multiple templates (Cheng, 2008; Raman et al., 2009; Zhang et al., 2010; Zhang, 2009a). Although the multiple template approach can build models with topology closer to the native

structure than individual templates, the assembled structures often contain significant local distortions, including steric clashes, unphysical phi/psi angles, and irregular hydrogen-bonding (H-bonding) networks, which render the structure models less useful for high-resolution functional analysis (Arakaki et al., 2004; Keedy et al., 2009). How to refine the protein structure closer to the native while keeping the physically meaningful atomic details of local structure remains a significantly unsolved problem in the field of protein structure predictions (Zhang, 2009b).

Different from the template-based structure assembly simulations that are usually implemented in reduced-level modeling, molecular dynamics (MD) simulations try to relocate every protein atom following Newton's laws of motion (Alder and Wainwright, 1957). Because of the advantage of direct sampling of protein atoms as guided by physics-based force field, MD simulations have been widely used in the atomic-level protein structure refinements (Chen and Brooks, 2007; Fan and Mark, 2004; Floudas, 2007; Lee et al., 2001). Except for some isolated instances, however, no systematic structural improvement has been achieved (Lee et al., 2001). Although it is efficient for removing steric clashes, MD simulations without restraints often drive the structure away from the native state, especially when relaxing the overly compact structural models, such as the structures constructed by the recombination of multiple homologous templates (Zhang, 2007, 2009b).

Recently, Zhu et al. (2008) performed replica-exchange molecular dynamics (REMD) simulations to refine 21 protein structures built by comparative modeling. The authors found that the REMD simulations could produce structures with a lower RMSD than the initial models and the best in top five models has a RMSD improvement of 0.24 Å in the secondary structure region. Although encouraging, the experiment highlights a key issue of physics-based structure refinements—that is, no atomic potential could distinguish the near-native structures from nonnative structures. However, the energy of the native structures was often found to be lower than that of all structure decoys. These data indicate that the current energy landscape is similar to a golf court, with the native state as the deepest hole, but lacks a middle-range funnel that could guide the simulation to the native state (Zhang, 2009b).

Another noteworthy observation was recently made by Summa and Levitt (2007), who exploited atomic potentials from AMBER99, OPLS-AA, GROMOS96, and ENCAD, on the

particular the possibility of reshaping the middle-range funnel of physics-based energy landscapes. We developed a Fragment-Guided Molecular Dynamics (FG-MD) algorithm, which combines the physical-based force field AMBER99 (Wang et al., 2000) with knowledge-based H-bonding and repulsive potentials. Distance maps taken from high-resolution experimental fragments are used as restraints to guide the simulated annealing MD simulations. FG-MD was extensively tested on both benchmark and CASP refinement experiments and demonstrated significant potential in atomic-level protein structure refinements.

RESULTS AND DISCUSSION

Benchmark of FG-MD on Structural Refinement of I-TASSER Models

The flowchart of FG-MD is showed in Figure 1, which consists of three steps of fragment structure identification, simulated annealing MD refinement simulation, and final model selection (see Experimental Procedures for detailed descriptions).

To benchmark the performance of the FG-MD approach, we collected a test set of 181 nonredundant proteins from the PDB that are nonhomologous to the training proteins of our MD force field. It includes 82 α, 28 β, and 71 α/β single-domain proteins ranging in length from 64 to 222 residues. The initial structural models were generated by the I-TASSER pipeline, which represents a typical multiple template reassembly approach (Roy et al., 2010; Zhang, 2007). Although the I-TASSER assembly simulations have drawn the threading templates significantly closer to the native (with the average RMSD reduced by 1.15 Å and average TM-score increased by 12% compared to the best threading template), the structure models were found to be overly compact and have considerable unphysical distortions in the local structures. On average, there are 50 steric clashes between the heavy atoms of the models, and only 48.07% of native H-bonds are retrieved on the I-TASSER models (Table 1).

To examine in detail the effect of various energy terms and spatial restraints, we split FG-MD into five different runs:

refinement of 75 proteins by *in vacuo* energy minimization. The authors found that a knowledge-based atomic contact potential outperformed all the traditional physics-based potentials by moving most protein models closer to the native state, while the physical potentials, except for AMBER99, essentially drove the decoys away from the native. The vacuum environment without solvent may be part of the reason for the failure of the molecular mechanics minimization. However, this observation demonstrated the potential of combining knowledge-based potentials with physics-based force field to improve the funnel shape of the energy landscape of protein folding force field (Wroblewska et al., 2008).

In this work, we will systematically examine the ability of MD simulations to refine protein structural models and check in

1) MD: Simulated annealing MD simulations using AMBER99 force field and a knowledge-based Cα repulsive potential (Equations 3 and 4);

Table 1. Summary of FG-MD Refinement on a Benchmark Set of 181 Proteins

Method	GDT[a]	TM[b]	Rmsd[c]	HB[d]	No. of Clash[e]
I-TASSER	0.4901	0.7163	4.971	0.4807	50
MD	0.3746	0.6379	5.721	0.3668	2
MD+HB	0.4035	0.6588	5.552	0.5312	2
MD+HB+DRM	0.4908	0.7164	4.980	0.5194	0
MD+HB+DRM+DRT	0.4916	0.7167	4.977	0.5240	2
MD+HB+DRM+DRFG	0.4946	0.7173	4.992	0.5250	3
MD+HB+DRM+DRT+DRFG	0.4971	0.7187	4.940	0.5340	2

[a] Average GDT-HA score.
[b] Average TM-score.
[c] Average C_α RMSD.
[d] Average hydrogen-bonding score.
[e] Average number of steric clashes between heavy atoms.

2) MD+HB: MD simulations with H-bonding optimization guided by a knowledge-based HB potential (Equation 2);

3) MD+HB+DRM: MD+HB simulation guided by the distance restraints taken from the initial I-TASSER model (First term of Equation 1);

4) MD+HB+DRM+DRT: MD+HB+DRM simulation with additional restraints taken from global template searched by TM-align from the PDB (second term of Equation 1);

5) MD+HB+DRM+DRFG: MD+HB+DRM simulation with additional restraints taken from structural fragments searched by TM-align from the PDB (third term of Equation 1); and

6) MD+HB+DRM+DRT+DRFG: MD+HB+DRM simulation with additional distance restraints taken from both global templates and local structural fragments searched from the PDB.

Table 1 summarizes the average results on the 181 target proteins. First, the MD simulations show an apparent power for removing the steric clashes between all heavy atoms (with the average number of clashes per protein reducing from 50 to 2), because of the strong repulsive term of Lennard-Jones potential in AMBER99 force field and the knowledge-based C_α repulsive term. However, the quality of the global topology as measured by GDT-HA (Zemla, 2003) and TM-score (Zhang and Skolnick, 2004) was considerably degraded (i.e., GDT-HA reduces by 30.8% and TM-score by 12.3%; the HB-score is also reduced by 31%). The RMSD of the initial models was increased by 0.75 Å. We note that running MD simulations without the knowledge-based C_α repulsive potential results in models with similar RMSD, GDT-HA, TM-, and HB-scores. But the additional repulsive potential significantly speeds up the convergence of the MD relaxing simulations.

The MD+HB approach was formulated by using backbone H-bonding optimization with a newly developed knowledge-based HB potential (Figure 8). Because part of the native hydrogen-bonding network was recovered, the corresponding HB-score increases by 45% (from 0.3668 to 0.5312). The GDT-HA and TM-scores are also increased by 7.7% and 3.2%, respectively, and RMSD decreases by 0.17 Å, which demonstrate a correlation between the hydrogen-bonding network and the global topology score, especially for beta-proteins, because of the long-range H-bonding between beta strands.

Although the models were improved by MD+HB, the global topology was still further away from the native than the initial models, as judged by the GDT-HA and TM-score and RMSD. We therefore applied the distance maps collected from the starting models to guide the MD simulations with the purpose of constraining the simulation near the initial structures. Indeed, the restraints improved GDT-HA, TM-score, and RMSD to a similar level of the starting models (GDT-HA, 0.4901 vs. 0.4908; TM-score, 0.7163 vs. 0.7164, and 4.971 vs. 4.980 Å). However, the HB-score was slightly lower than that by MD+HB because of the bending to the starting models, which have a distorted H-bonding network.

In the MD+HB+DRM+DRT simulation, we added the distance restraints taken from the high-resolution PDB structures that are closest to the initial model, as searched by the structure alignment program TM-align (Zhang and Skolnick, 2005). Because the experimental structures have ideal local structures with standard H-bonding networks, we anticipated that the inclusion of the template-based restraints could improve the quality of the local structures; this was indeed verified by the increased HB-score and a slight improvement in the TM- and GDT-HA scores and RMSD.

In the MD+HB+DRM+DRFG simulation, we split the sequences into smaller fragments and used the substructures of the initial model as probes to search for high-resolution template structures from the PDB. The distance map restraints taken from the fragmental templates were then incorporated into the MD simulations. Overall, the restraints from local fragments outperformed those from global templates, and better TM-score, GDT-HA, RMSD, and HB-scores were achieved in the final model structures.

Finally, in the MD+HB+DRM+DRT+DRFG, we combined the distance restraints taken from both global templates and local structural fragments. The models generated by this approach have achieved the highest TM-score, GDT-HA and HB-scores, and lowest RMSD among all the simulations (Table 1).

In Figures 2A–2C, we present scatter plot distributions of GDT-HA, TM-score, and HB-score improvements versus TM-score of the starting models. Although the points were scattered both below and above the dashed reference line, there were obviously more proteins with improved scores than that with deteriorated ones. The average improvement was showed
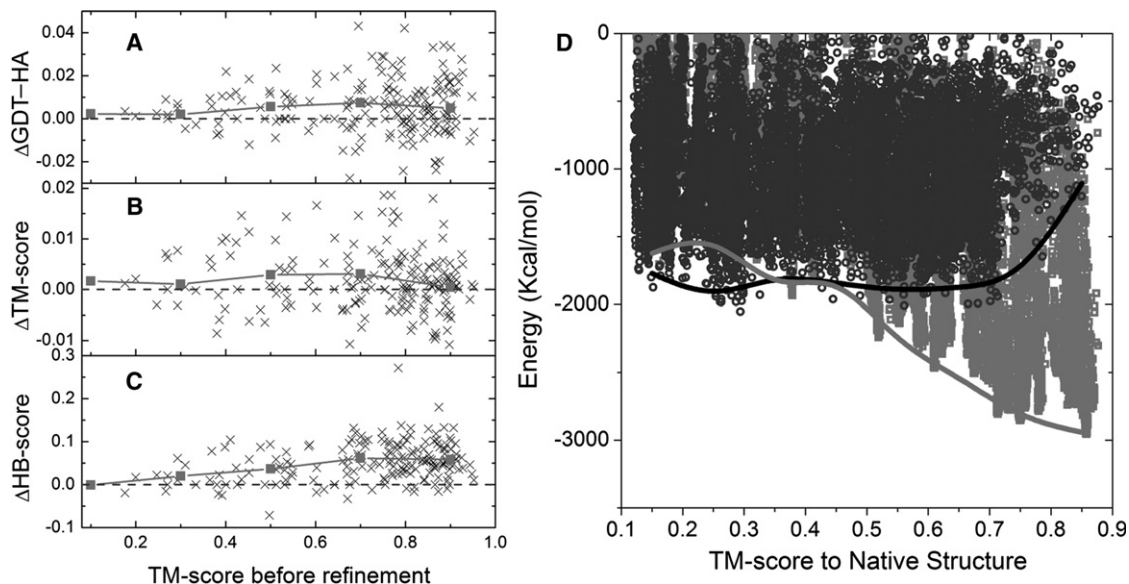
Figure 2. Scatter Plot of the FG-MD Improvements versus TM-Score of Initial Models

(A–C) GDT-HA (A), TM-score (B), and HB-score (C) are shown.

(D) AMBER99 (black circles) and FG-MD (gray squares) energy versus TM-score for 77 × 200 refined models by MD and FM-MD simulations. The fitting curves are connection of the medians of the 10 lowest-energy models in each of the TM-score bins (0.1-0.2, 0.2-0.3, …). See also Figures S1 and S2.

in the solid line in each plot. For 181 benchmarking targets, 71% (128/181) of the targets have a GDT-HA improvement, 65% (117/181) of the targets have a TM-score improvement, and 92% (167/181) of the targets have a HB-score improvement. A detailed histogram of the score difference is showed in Figure S1 (available online), which is available with this article online.

Although the quality improvement of local structures demonstrated by the HB-score seems to occur on all ranges of protein models, the most significant improvement of GDT-HA and TM-score was observed for the models of good starting structures (e.g., TM-score >0.5). For the models of incorrect topology (e.g., TM-score <0.5), the GDT-HA and TM-score improvements by FG-MD are less pronounced. This may indicate that the FG-MD force field landscape has a sensitive funnel shape in the region of TM-score >0.5; within the region of TM-score <0.5, however, the correlations of energy and TM-score are much weaker, and the influence of local geometry repacking on the global topology refinement therefore tends to be randomized.

As an illustration, in Figure 2D we calculated the AMBER99 and FG-MD energy force fields according to the models from a successfully refinement example of 2bwfA, which had the TM-score improved from 0.798 to 0.804. For this target, 77 decoy models with TM-score ranging from 0.17 to 0.84 were taken from the I-TASSER simulation trajectory; 200 refinement models were then generated by MD and FG-MD simulations separately for each of the 77 initial models. The corresponding AMBER99 (black circles) and FG-MD (red squares) energy potentials for the 77 × 200 refined models are plotted versus the TM-score. As shown by the fitting curves (the median of the 10 lowest energy models per TM-score bin), there was almost no correlations between AMBER99 and the global topology. After the introduction of the better fragment-based distance restraints (with a DMRMSD 0.036Å lower than that of

the initial model), there appears an apparent funnel-like shape starting from TM-score ~0.5 in the FG-MD energy landscape, which is critical to the success of structure refinement of this example.

Although encouraging, the funnel-like landscape was not always achieved in the FG-MD force field even when TM-score of the initial model is >0.5. The reshaping of energy landscape was found to strongly rely on the quality of fragment templates, although other energy terms in FG-MD contribute as well. Among the successfully improved targets, the majority of the targets were found to have a funnel-like landscape similar to Figure 2D (with lower energy in the near-native regions). For the deteriorated targets, most of targets have still an energy landscape that lacks energy-TM-score correlations. In Figure S2, we present such an example from 1z3e, where the DMRMSD of the best fragment template is 0.1 Å worse than the initial model and the FG-MD energy landscape shifts the lowest energy away from the native state, compared with the AMBER99 potential. Consequently, the TM-score and GDT-HA of the initial model were deteriorated by 0.0074 and 0.0089, respectively, in this example.

Why Can FG-MD Refine the Global Topology of the Protein Models?

Except for the adjustment of local structures and H-bonding networks, the major driving force for the global structural refinement of FG-MD is the distance map restraints taken from the initial model, the global and fragmental templates as searched by TM-align from the high-resolution PDB structures. Generally, the distance maps from the initial models do not directly contribute to improving the topology of protein models because these restraints do not contain new information to the initial structures. In Figure 3A, we plot the histogram of ΔTM-scores that is defined as the TM-score of the TM-align templates subtracting
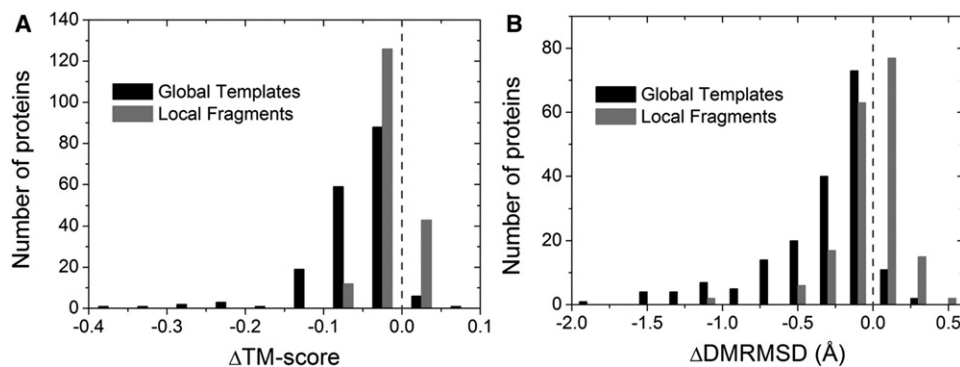
Figure 3. Topology and Restraint Accuracy from Global and Fragment Templates
(A) Histogram of ΔTM-score of global templates (black) and fragmental templates (gray).
(B) Histogram of ΔDMRMSD of distance map restraints taken from the global templates (black) and fragmental templates (gray). See also Figure S3.

the TM-score of the initial models, where a positive ΔTM-score value indicates a better quality of template structures than the initial models. Because the initial models were used as probes in the template search, most of the templates are closer to the initial models than to the native (i.e., there are more templates with a TM-score lower than that of the initial models). These data are consistent with the observation we obtained in the previous study (see Figure 4A in Zhang and Skolnick, 2005). However, because there are much more appropriate fragmental templates than the global ones for the given initial models, we obtained a better quality of fragment templates (see the red bars in Figure 3A). Overall, 24% of fragment templates have a higher TM-score than the corresponding initial models, and only 4% of global templates do so.

In Figure 3B, we present the distribution of ΔDMRMSD, which is defined as the distance map root-mean-squared-deviation (DMRMSD, Equation 6) from initial models subtracting the DMRMSD from the templates, where a positive ΔDMRMSD indicates a more accurate distance map restraint from templates than that from initial models. Again, most of the restraints (90%) from the global template have a lower accuracy than those from the initial templates. Interestingly, there are slightly more distance restraints from the fragment templates (52%) that have a better accuracy than that from the initial models. Because the average TM-score of the fragment templates to the native is still lower than that of the initial models, this excess in distance maps is probably attributed to the more proteinlike and regular secondary structures in the experimental structures, which constitutes the major driven force for the FG-MD to refine the high-resolution protein structures.

To examine the detailed data in different categories, we split the targets into "successful" (TM-score increased after refinement) and "failed" (TM-score decreased after refinement simulation) cases, and compared the restraints from fragment and global templates separately, as shown in Figure S3. For the successful cases, 56% (65/117) of the targets have better fragments (ΔDMRMSD > 0) and 10% (12/117) of the targets have better global templates (ΔDMRMSD > 0). For failed cases, only 44% (28/64) of the targets have better fragment (ΔDMRMSD > 0) and 2% (1/64) of the targets have better global templates (ΔDMRMSD > 0). On average, the fragment templates

are always better than global templates for both successful and failed cases. However, only in the successful cases, the quality of restraints outperformed that from initial models.

In Figure 4, we look into the details of a typical refinement example from the *Escherichia coli* 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (PDB ID: 1eqmA). Figures 4A and 4B are the long-range distance maps ($|i-j| > 10$) from the initial model and that from fragmental templates, respectively, versus the distance map of the native structure, where only the residue pairs with a distance < 8Å are presented. Although all 199 restraints from the initial model appeared in that from the fragment structures, there were 57 new distance restraints predicted only by fragments (circles). If we define a correct distance restraint as that with $|D_{ij} - D_{ij}^N| < 0.5$ Å, where D_{ij} and D_{ij}^N are distances from model and native, there are 91 and 188 correct restraints from initial model and fragmental templates, respectively. Thus, the fragment-based restraints outperform the initial models in both accuracy (188/256 vs. 91/199) and coverage. As a result, the GDT-HA, TM-, and HB-scores of the initial models were increased from 0.573, 0.791, and 0.391, respectively, to 0.6, 0.796, and 0.529, respectively (Figure 4C).

Figure 5 shows three additional, typical examples with major improvements from loop, helix, and strand regions, respectively. For 1elw, a high-quality fragment at the loop region (19I-66L) with a DMRMSD of 0.228 Å drew the RMSD/DMRMSD of the corresponding region in the initial model from 0.439/0.306 Å to 0.365/0.211 Å, which results in the GDT-HA and TM-scores of the global model increasing from 0.686 and 0.885 to 0.720 and 0.898, respectively. Similar improvement was observed for the other two proteins with PDB ID 1hnl and 1djr, one with improvements occurring in the helical region and another in the stranded region, both attributing to the better fragmental templates identified in the corresponding regions (rows 2 and 3 in Figure 5).

Performance of FG-MD on CASP8 Refinement Targets
The structure refinement category was a new addition to CASP since CASP8 (MacCallum et al., 2009). In this category, predictors were given starting models that had been generated by the CASP structural prediction servers and judged by organizers to be among the best for each targets, and requested to refine
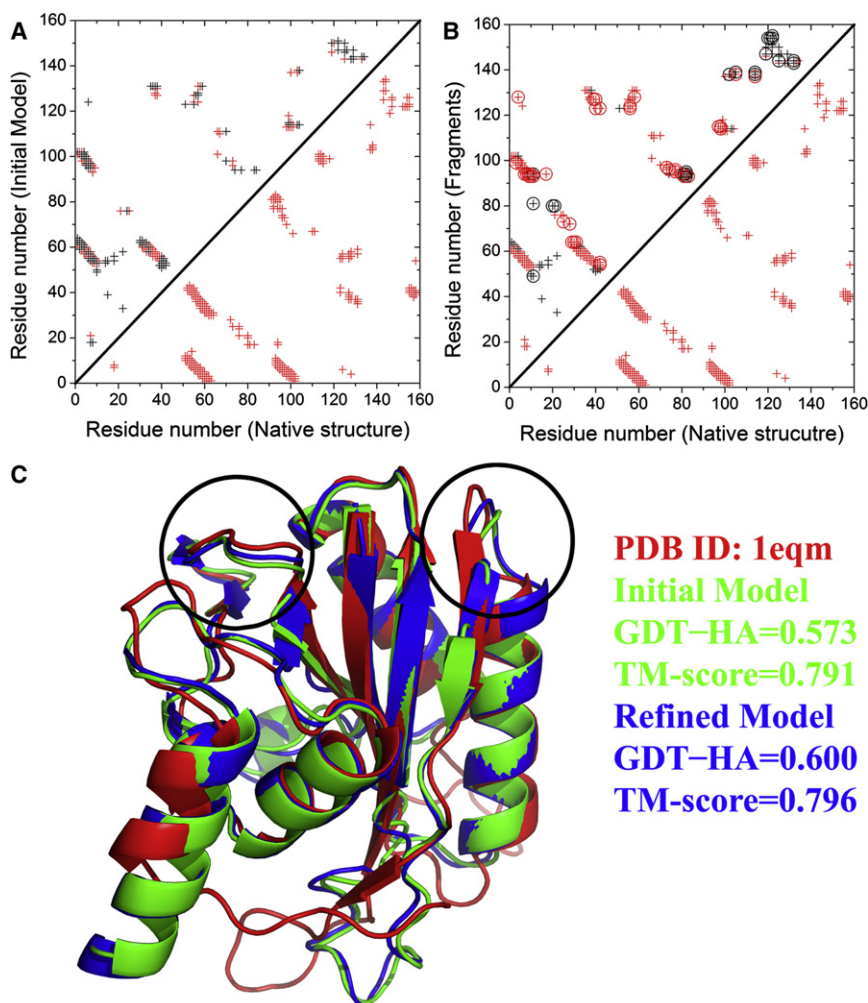
Figure 4. An Example of FG-MD Refinement on 1eqm

(A) Long-range distance map from initial model versus that from native ($|i$-$j| > 10$).

(B) Long-range distance map from fragments versus that from native. Only the distances below 8 Å are shown. Distance restraints with error below 0.5 Å are shown in red and others in black spots. The new restraints from fragments, which are not in the initial model, are highlighted by circles.

(C) Superposition of the refined model (blue) and the initial model (green) on the native structure (red). The black circles highlight the regions of more pronounced structural improvements.

PDB ID: 1eqm
Initial Model
GDT−HA=0.573
TM-score=0.791
Refined Model
GDT−HA=0.600
TM-score=0.796

the models in a blind mode. We tested FG-MD on refining all the 12 target proteins from the CASP8 refinement experiment. Although the FG-MD run was performed after the CASP8, we note that all templates proteins deposited in PDB after CASP8 were excluded to ensure that we are in strictly the same modeling condition as the CASP8 blind predictors.

The FG-MD refinement result is summarized in the upper part of Table 2 together with the best five groups as ranked by the cumulative GDT-HA score of refined models for all the 12 models. A full list of 25 groups is given in Table S1. As illustrated in the overall results, FG-MD is the only method that could drive the initial protein models closer to the experiment structure according to both the cumulative GDT-HA and TM-scores and the average RMSD. Because some groups submitted fewer proteins, we also calculated the average TM-score and GDT-HA on the submitted proteins and found that none of the CASP8 groups have a TM-score, GDT-HA, or RMSD better than the initial models, which indeed highlights the difficulty in refining global topology of protein models.

Overall, the GDT-HA and TM-score are 1.2% and 1.6% higher and RMSD is 0.164 Å lower than that by the second best LEE group. The HB-score of the FG-MD models was ranked the

third position following SAM-T08-human and LevittGroup. The average number of steric clash is 0.1, lower than the models from all other top five groups. A comparison of the FG-MD models with the initial models for all 12 CASP8 protein targets are listed in Figure 6A, where FG-MD improved the GDT-HA and TM-scores in nine of 12 cases and the HB-score in 10 of 12 cases. Again, the most significant improvements were observed in the region of TM-score larger than 0.5.

In the CASP8 experiment (MacCallum et al., 2009), the assessor also used the MolProbity score (Chen et al., 2010) to count for the physically unfavorable steric overlaps, rotamer, and Ramachandran outliers. We listed the MolProbity score in the last column of Table 2 and Figure 6B. FG-MD improves the MolProbity score of the initial models for all but two targets with an average reduction by 5%.

In Figures 6C and 6D, we present two representative examples of the CASP8 refinements. For Target TR432, the improvements were scattered among the entire sequence, including loop and regular secondary structures, which resulted in a 3.2% increase in GDT-HA, a 0.9% increase in TM-score, an 11.1% increase in HB-score, and an 8.8% increase in MolProbity score (Figure 6C). For TR461, however, the improvement was mainly in the beta strand region, which resulted in an increase of GDT-HA, TM-score, and HB-score by 2.0%, 0.2%, and 8.7%, respectively (Figure 6D).

Blind Test of FG-MD in CASP9 Refinement Experiment

We participated (as "ZHANG") in the structure refinement section in the CASP9 experiment, which contained 14 protein targets with length from 69 to 159 residues. Two, five, and seven targets belong to α, β, and αβ proteins, respectively. A summary of the refinement results for the top five groups is listed in the lower part of Table 2 according to the cumulative GDT-HA score. A complete list of the CASP9 groups is listed in Table S2. We note that the algorithm generating the ZHANG models in CASP9 was not identical to the FG-MD reported in this work because the models submitted in CASP9 were further refined
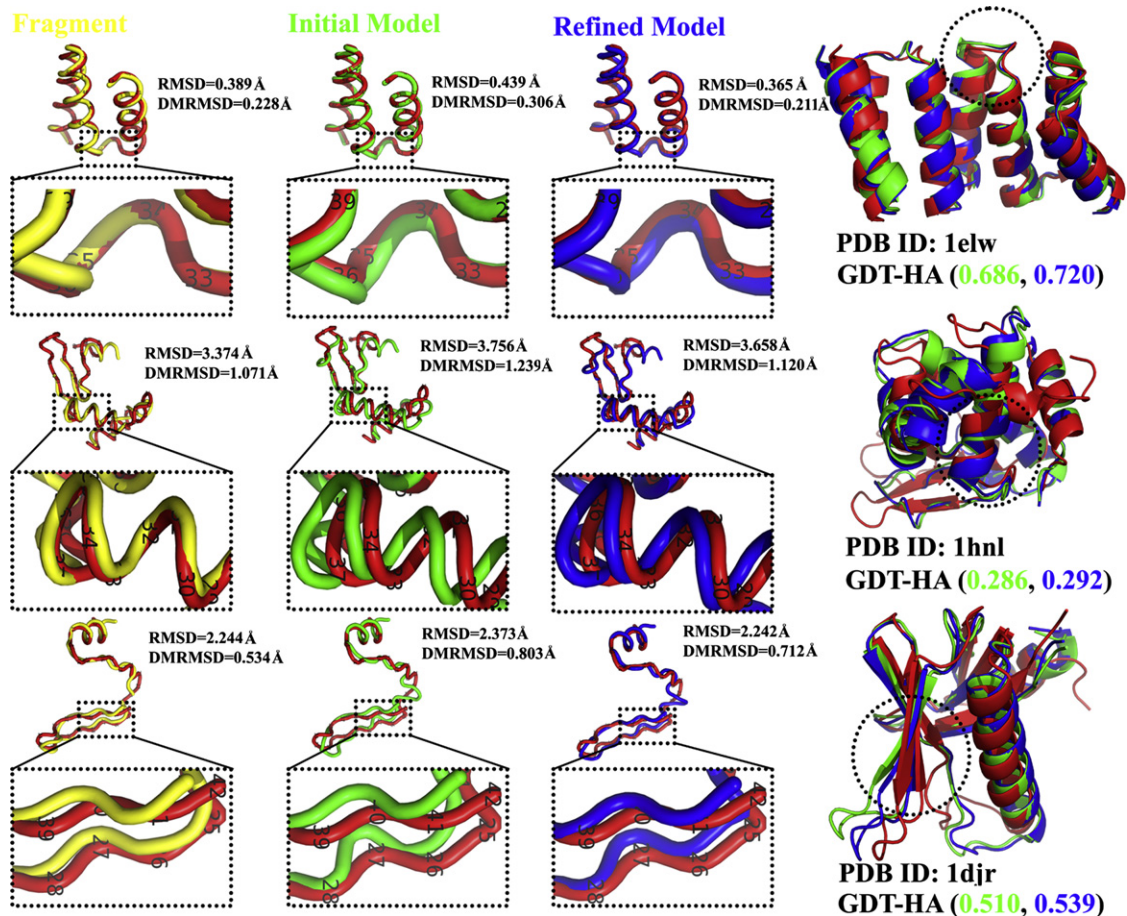
Figure 5. Three Examples of Refinement by FG-MD Simulations in the Benchmark Test Set
The superposition of fragmental template (Column 1), initial model (Column 2), and refined model (Column 3) on the experimental structure in the fragmental region are shown. Column 4 is the superposition of the full-length refined model and initial model on the native. The examples are from 1elw (Row 1), 1hnl (Row 2), and 1djr (Row 3). The black dotted circles highlight the regions of more pronounced structure improvements.

by MODELER (Sali and Blundell, 1993), where the FG-MD refined model was used as the single template and then a simple AUTOMODEL step was performed in MODELER. The reason for us to use MODELER was that it could slightly improve the rotamer torsion angles according to our benchmark test. However, we found that MODELER refinements could result in a significant increase of H-atom clashes and therefore deteriorated the MolProbity score. Therefore, we modified the FG-MD algorithm without using MODELER in this work. In Table 2, we also listed the models by the current version of the FG-MD program. All the FG-MD refined models for the CASP9 targets, together with that for CASP8 targets and the 181 I-TASSER models, were uploaded to http://zhanglab.ccmb.med.umich.edu/FG-MD/.

The groups in the lower part of Table 2 were ranked according to the cumulative GDT-HA score of the first model. Only two groups, ZHANG and SEOK in CASP9, could refine the initial structures on the basis of GDT-HA and TM-scores. However, both ZHANG and SEOK models have the MolProbity score higher than those of the initial models, indicating a deterioration of the local structure qualities. In particular, SEOK models showed a significant number of heavy atomic overlaps (Column

7). After removing the MODELER refinement step, FG-MD improves the MolProbity score by 39% compared to ZHANG, which is 15% lower than that of the initial models. Overall, the FG-MD models demonstrate improvement in all aspects of GDT-HA, TM-score, RMSD, HB-score, heavy atom clash, and MolProbity score.

In Figures 7A–7D, we present the improvement of GDT-HA, TM-score, HB-score, and MolProbity score over the initial models for all individual targets by ZHANG and FG-MD. Again, the results for ZHANG and FG-MD are similar for GDT-HA, TM, and HB-scores, but FG-MD without the MODELER step reduces significantly the MolProbity score over the ZHANG models. There are 11, 10, 10, and five of 14 cases, respectively, in which the ZHANG models show improvement over the initial models according to GDT, TM, HB, and MolProbity scores, and the FG-MD models do so in 12, 10, five, and 13 cases, respectively.

Figures 7E and 7F show two typical examples of ZHANG models in CASP9. For TR614, the major improvement is in region 82N-91E, although minor improvement were scattered along the whole chain (Figure 7E), which resulted a final model

Table 2. Results of FG-MD and the Top Five Groups in CASP8 and CASP9 Experiments

	Group	No. Targ[a]	GDT[b]	TM[c]	Rmsd[d]	HB[e]	No. of Clash[f]	Molprob[g]
CASP8	FG-MD	12	6.979	9.362	2.953	6.758	0.1	2.575
	NULL[h]	12	6.898	9.316	3.004	6.290	27.9	2.706
	LEE	12	6.860	9.195	3.117	6.624	18.4	2.613
	LevittGroup	12	6.701	9.160	3.047	6.730	5.3	2.875
	SAM-T08-human	12	6.523	9.084	3.056	6.773	0.7	2.762
	YASARARefine	12	6.407	9.155	3.105	6.853	0.3	1.071
	Bates_BMM	12	6.167	8.734	3.359	6.060	2.9	2.738
CASP9	FG-MD	14	7.387	10.386	4.331	6.928	0.0	2.183
	ZHANG	14	7.365	10.396	4.338	7.084	0.0	3.042
	SEOK	14	7.359	10.399	4.259	6.819	15.8	3.436
	NULL[h]	14	7.319	10.368	4.344	6.867	3.7	2.521
	FAMSD	14	7.284	10.348	4.309	6.790	1.1	2.550
	KNOWMIN	14	7.194	10.182	4.440	7.590	1.2	2.179
	TASSER	14	7.164	10.259	4.740	6.814	47.4	3.160

See also Tables S1 and S2.
[a] Number of targets.
[b] Cumulative GDT-HA score of the first models.
[c] Cumulative TM-score of the first models.
[d] Average RMSD of the first models to the native structure (Å).
[e] Cumulative hydrogen-bonding score of the first models.
[f] Average number of heavy atom steric clashes in the first model.
[g] Average MolProbity score of the first model.
[h] The initial models for the CASP refinement experiment.

of GDT-HA score 0.535, which was 1.7% higher than in the initial model. For the second target TR530, the major refinement occurs at 49R-57E and 108D-115K, which results in a final model of GDT-HA score 0.709, which was 2.6% higher than in the initial model (Figure 7F).

Conclusions

Atomic-level protein structural refinement represents a significantly unsolved problem in protein structure prediction. Most of the current methods based on MD simulations drive the structural models away from the native state, mainly because of the lack of long-range correlation between the topology and the energy of the physics-based atomic potentials (Summa and Levitt, 2007; Zhang, 2009b; Zhu et al., 2008). The multiple-template-based homology modeling has the potential to improve the structure of the threading templates, and a fundamental issue is that the local structures of the resultant model can be seriously distorted (Fischer, 2003; Zhang, 2007). In this work, we developed a protocol of FG-MD for high-resolution and atomic-level protein structure refinements, where spatial restraints from fragmental templates were exploited to reshape the energy funnel of the simulated annealing MD simulations. The knowledge-based H-bonding potential is incorporated for improving the local structure refinements.

FG-MD was tested in a large-scale set of 181 benchmark proteins with initial models generated by the I-TASSER template structural reassembly approach. It was found that the distance map restraints extracted from fragmental templates had a significant higher accuracy than those obtained from the global templates; the accuracy was also higher on average than that from the initial models. As a result, progressive improvements were observed when H-bonding energy term and the distance map restraints were separately introduced to guide the refinement simulations. The average GDT-HA score of the FG-MD models was 0.7 units higher than that of the initial models, and RMSD was reduced by 0.031 Å. The majority of the improvements happened for the initial models of correct topology (i.e., TM-score > 0.5). This corresponds to a funnel-shaped improvement of the physics-based energy landscape from a golf-course-like shape to a funnel-like shape with an approximate radius of TM-score of 0.5, which had been seen in most of the successfully refined targets. As part of local structure measurements, the H-bonding score of the FG-MD models increased by 12%, and the number of steric clashes was reduced from 50 to 2, mainly because of the introduction of the H-bonding and repulsive energy terms in the MD simulations.

The FG-MD method was also tested in the blind test of CASP refinement experiments, where starting models were generated by a variety of structure assembly methods. FG-MD was among the very few methods that could consistently bring the initial models closer to the native structure as assessed by the improved GDT-HA and TM-scores. The local structural quality of the H-bonding score, the number of steric clashes, and the MolProbity score were also significantly improved, compared to the initial models. These data demonstrate a promising approach of atomic protein structure refinements by using analogical fragment templates from other experimentally solved protein structures.
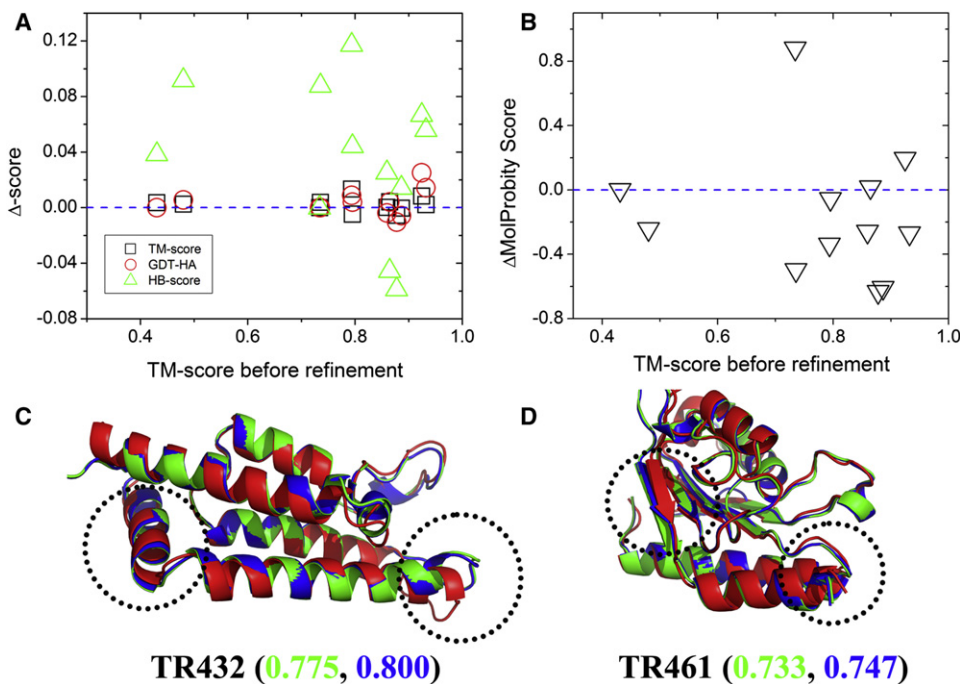
Figure 6. FG-MD Structural Refinement Results on CASP8 Targets

(A) Scatter plot of improvements on TM-score, GDT-HA, and HB-score.

(B) Scatter plot of MolProbity improvements.

(C) Structural superposition of the initial model (green) and refined model (blue) on the native structure (red) for Target TR432.

(D) Structural superposition of the initial model (green) and refined model (blue) on the native structure (red) for Target TR461. The dotted circles highlight the regions of more pronounced structural improvements.

EXPERIMENTAL PROCEDURES

FG-MD Refinement Protocol

The FG-MD protocol is showed in Figure 1. Starting from a target protein structure, the sequence was split into separate secondary structure elements (SSEs). The substructures of every three consecutive SSEs, together with the full-length structure, were used as probes to search through a nonredundant PDB library by TM-align (Zhang and Skolnick, 2005) for structure fragments closest to the target. The top 20 template structures with highest TM-scores (Zhang and Skolnick, 2004) were used to collect spatial restraints. Simulated annealing molecular dynamics simulations were implemented using modified LAMMPS (Plimpton, 1995), as guided by the distance map restraints and a knowledge-based hydrogen-bonding potential. The final refined models were selected on the basis of the sum of Z-score of hydrogen bonds, Z-score of the number of steric clashes, and Z-score of FG-MD energy. The procedure is fully automated, and the running time for each refinement target is less than 2 hr at a 2.4 GHz CPU. The FG-MD server is freely available at http://zhanglab.ccmb.med.umich.edu/FG-MD.

FG-MD Force Field

The FG-MD potential contains four energy terms.

Distance Map Restraints

The C_α distance maps were collected from three sources of initial models, global structure templates, and fragmental structure templates. The C_α distance restraint potential is written as:

$$E(r_{ij}) = \begin{cases} k_1\left(r_{ij} - r_{ij}^1\right)^2 + k_2\left(r_{ij} - r_{ij}^2\right)^2 + k_3\left(r_{ij} - r_{ij}^3\right)^2 & r_{ij} \leq 15 \\ 0 & r_{ij} > 15 \end{cases}, \quad (1)$$

where r_{ij} is the distance between ith and jth C_α atoms. r_{ij}^1, r_{ij}^2, and r_{ij}^3 are the distance maps from the initial model, global structure template, and fragmental template, respectively. k_1, k_2, and k_3 are the corresponding force constants with the value equal to 0.5, 0.5, and 2.0 Kcal/mole, respectively.

TM-align was used to build the structural alignment between the template structures and the initial model. We found that restraints from residues in close distance are most efficient and thus we only considered the distance map with r_{ij} below 15 Å from the top 20 templates. The force parameters (and others shown below) were decided on the performance of FG-MD on an independent training set of proteins.

Explicit Hydrogen Binding

The definition of backbone hydrogen bond is illustrated in Figure 8A. A knowledge-based, explicit H-bonding potential is constructed as:

$$E_{HB}(d_{ij}, \alpha, \beta) = \begin{cases} k_4(d_{ij} - d_0)^2 + k_5(\alpha - \alpha_0)^2 + k_6(\beta - \beta_0)^2 & d_{ij} \leq 3.0 \\ 0 & d_{ij} > 3.0 \end{cases}, \quad (2)$$

where d_{ij} is the distance between hydrogen of the donor and oxygen of the acceptor. α is the angle of N-H-O, and β is the angle of C-O-H (Figure 8A). The standard values of the H-O distance, N-H-O, and C-O-H angles are derived from the statistics average of 1,383 nonredundant, high-resolution experimental structures (see Figures 8B–8D). This protein set was constructed with the PISCES server (Wang and Dunbrack, 2003), with a percentage identity cutoff 20%, a resolution cutoff 1.6 Å, and an R-factor cutoff 0.25 Å. The average result is $d_0 = 1.95 \pm 0.17$ Å, $\alpha_0 = 160.0 \pm 12.2°$, and $\beta_0 = 150.0 \pm 17.5°$. Because the fluctuations of the experimental values are relatively small, we took the average value with a harmonic restraint for the H-bonding in Equation 2. K_4, k_5, and k_6 are the force constant with values equal to 2.0, 0.5, and 0.5, respectively. Only short-range H-binding potential is considered, with a cutoff of $d_{ij} \leq 3Å$.

Repulsive Potential

The C_α repulsive potential was designed to rapidly relax compact structural models with severe C_α clashes, as follows:

$$E(r_{ij}) = \begin{cases} k(3.6 - r_{ij}) & r_{ij} \leq 3.6 \\ 0 & r_{ij} > 3.6 \end{cases}, \quad (3)$$

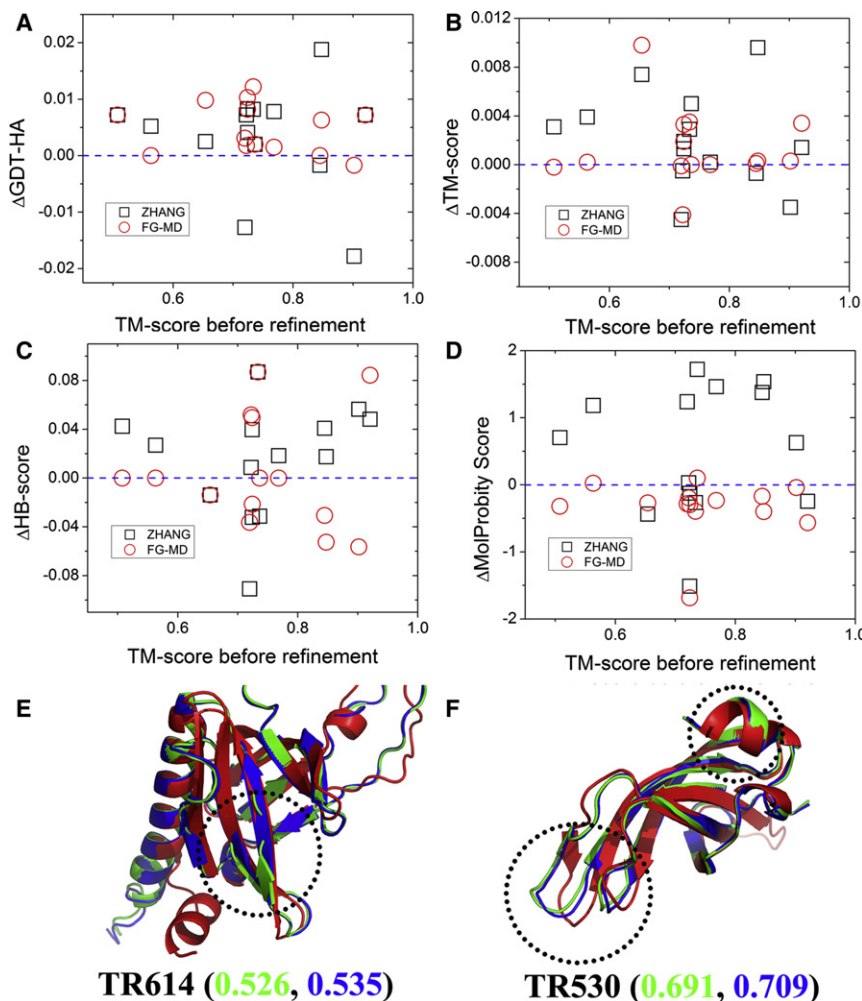where the force constant $k = 200$ Kcal/mole.

Figure 7. Structural Refinement Results on the CASP9 Targets

(A) GDT-HA score improvements.
(B) TM-score improvements.
(C) HB-score improvements.
(D) MolProbity score improvements.
(E) Structural superposition of the initial model and final model on the native structure for Target TR614.
(F) Structural superposition of the initial model and final model on the native structure for Target TR530. The black dotted circles highlight the regions of more pronounced structure improvements.

$$TM - score = Max \left[\frac{1}{L_N} \sum_{i=1}^{L_T} \frac{1}{1 + \left(d_i / d_0 \right)^2} \right], \quad (5)$$

where L_N is the length of the native structure, L_T is the number of common residues appearing in both compared structures, d_i is the distance between the ith pair of residues, and d_0 is a scale to normalize the match difference. Both GDT-HA and TM-score lie in [0, 1] with higher values indicating better accuracy. However, the GDT-HA score focuses more on the accuracy of local structure ($d < 4$ Å), and the TM-score counts the residues of all structures. Another difference is that TM-score is length independent, with TM-score > 0.5 indicate proteins of the same fold (Xu and Zhang, 2010). Historically, the GDT-HA score was used as the standard topology measurement for high-resolution protein structure prediction and refinements in CASP experiments (Cozzetto et al., 2009; Keedy et al., 2009; Kopp et al., 2007; MacCallum et al., 2009).

Accuracy of Distance Map
The RMSD of C_α distance map, called DMRMSD, is introduced to evaluate the accuracy of distance restraints:

$$DMRMSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left[\left(D_i - D_i^N \right)^2 \right]}, \quad (6)$$

where D_i is the ith distance restraints collected from structural models and D_i^N is the corresponding distance on the native structure. n is the total number of collected distance restraints.

Hydrogen-Bonding Score
The accuracy of the hydrogen-bonding network (or secondary structures) of the protein models is evaluated by the HB-score:

$$HB - score = \frac{Number\ of\ consensus\ hydrogen\ bonds\ in\ model\ and\ native}{Number\ of\ hydrogen\ bonds\ in\ the\ native\ structure}. \quad (7)$$

The hydrogen bonds in full atomic structures are defined by HBPLUS (McDonald and Thornton, 1994).

Steric Clashes
To count the unphysical overlaps of the model structures, we define a steric clash between two heavy atoms if the distance is less than 80% of the sum of Van der Waals radius, which is taken from Amber99 force field (Cornell et al., 1995).

AMBER99 Force Field
The standard AMBER99 force field (Wang et al., 2000) was used for local conformation stiffness:

$$E_{Amber} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$$

$$+ \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\varepsilon R_{ij}} \right], \quad (4)$$

where r, θ, and φ are bond length, bond angle, and torsion angle, respectively; and r_{eq}, θ_{eq}, and γ are the corresponding equilibrium values. K_r, K_θ, and V_n are the force constants for bond length, bond angle, and torsion angle, respectively. A_{ij} and B_{ij} are the Lennard-Jones parameters; q_i and q_j are the partial charge of atom i and j. R_{ij} is the distance between atom pair i and j.

Metrics for Structure Assessments
Accuracy of Global Topology
Although the root-mean-squared deviation (RMSD) between two structures was often used to measure the modeling accuracy, the value can be very sensitive to the local structural variations. In this work, we mainly used GDT-HA score (Zemla, 2003) and TM-score (Zhang and Skolnick, 2004) to evaluate the global topology refinement of the structural models. The GDT-HA score counts the average percentage of residues with C_α distance from the native below 0.5, 1, 2, and 4 Å, respectively, after the optimal structure superposition. The TM-score is defined as follows:
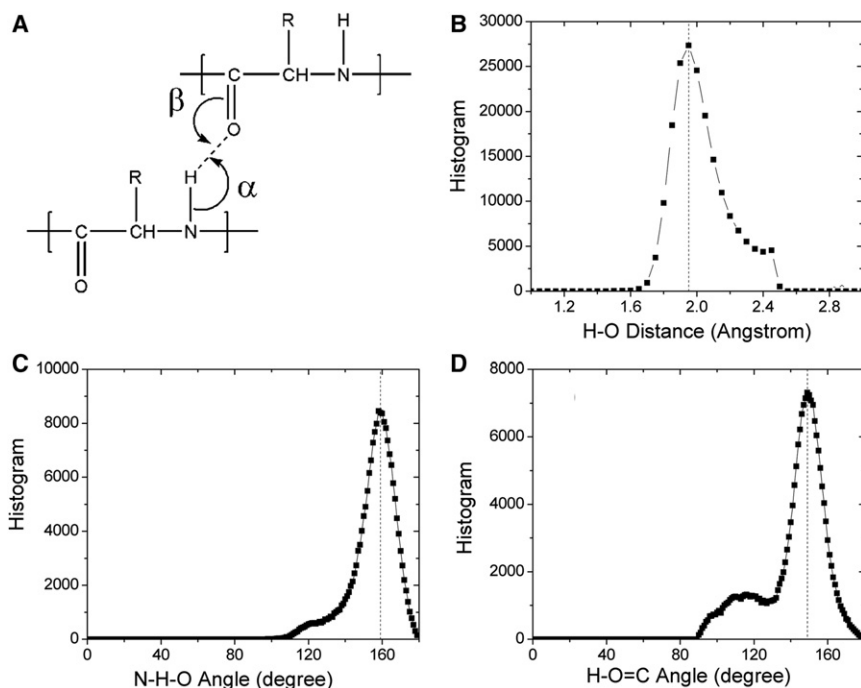
Figure 8. The Illustration and Statistics of the H-Bonding Potential Used in FG-MD
(A) Definition of H-O distance and inner angles for hydrogen-bond.
(B) Histogram of H-O distance.
(C) Histogram of the N-H-O angle.
(D) Histogram of the H-O-C angle.
All histograms were obtained from a set of 1,383 non-redundant, high-resolution experimental structures.

MolProbity Score

The MolProbity score is calculated as (Chen et al., 2010):

$$MolProbity - score = 0.426 \times \ln(1 + clashscore)$$
$$+ 0.33 \times \ln(1 + \max(0, rota\_out - 1)) \qquad (8)$$
$$+ 0.25 \times \ln(1 + \max(0, rama\_iffy - 2)) + 0.5,$$

where clashscore counts the number of unfavorable steric overlaps ≥ 0.4 Å, including H-atoms, and rota_out and rama_iffy are the percentages of the outliers of the side-chain rotamers and the backbone torsion angles, respectively. Lower MolProbity scores indicate more physically realistic models.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two tables and three figures and can be found with this article online at doi:10.1016/j.str.2011.09.022.

REFERENCES

Alder, B.J., and Wainwright, T.E. (1957). Phase transition for a hard sphere system. J. Chem. Phys. *27*, 1208–1209.

Arakaki, A.K., Zhang, Y., and Skolnick, J. (2004). Large-scale assessment of the utility of low-resolution protein structures for biochemical function assignment. Bioinformatics *20*, 1087–1096.

Chen, J., and Brooks, C.L., 3rd. (2007). Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins *67*, 922–930.

Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21.

Cheng, J. (2008). A multi-template combination algorithm for protein comparative modeling. BMC Struct. Biol. *8*, 18.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. *117*, 5179–5197.

Cozzetto, D., Kryshtafovych, A., Fidelis, K., Moult, J., Rost, B., and Tramontano, A. (2009). Evaluation of template-based models in CASP8 with standard measures. Proteins *77 (Suppl 9)*, 18–28.

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

Fischer, D. (2003). 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins *51*, 434–441.

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

Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics *19*, 1015–1018.

Keedy, D.A., Williams, C.J., Headd, J.J., Arendall, W.B., 3rd, Chen, V.B., Kapral, G.J., Gillespie, R.A., Block, J.N., Zemla, A., Richardson, D.C., and Richardson, J.S. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins *77 (Suppl 9)*, 29–49.

Kopp, J., Bordoli, L., Battey, J.N., Kiefer, F., and Schwede, T. (2007). Assessment of CASP7 predictions for template-based modeling targets. Proteins *69 (Suppl 8)*, 38–56.

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

MacCallum, J.L., Hua, L., Schnieders, M.J., Pande, V.S., Jacobson, M.P., and Dill, K.A. (2009). Assessment of the protein-structure refinement category in CASP8. Proteins *77 (Suppl 9)*, 66–80.

McDonald, I.K., and Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. *238*, 777–793.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular-dynamics. J. Comp. Physiol. *117*, 1–19.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins *77* (*Suppl 9*), 89–99.

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

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Summa, C.M., and Levitt, M. (2007). Near-native structure refinement using in vacuo energy minimization. Proc. Natl. Acad. Sci. USA *104*, 3177–3182.

Tramontano, A., and Morea, V. (2003). Assessment of homology-based predictions in CASP5. Proteins *53* (*Suppl 6*), 352–368.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

Wang, J.M., Cieplak, P., and Kollman, P.A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J. Comput. Chem. *21*, 1049–1074.

Wroblewska, L., Jagielska, A., and Skolnick, J. (2008). Development of a physics-based force field for the scoring and refinement of protein models. Biophys. J. *94*, 3227–3240.

Wu, S.T., and Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. *35*, 3375–3382.

Xu, J., and Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics *26*, 889–895.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. *31*, 3370–3374.

Zhang, J., Wang, Q., Barz, B., He, Z., Kosztin, I., Shang, Y., and Xu, D. (2010). MUFOLD: A new solution for protein 3D structure prediction. Proteins *78*, 1137–1152.

Zhang, Y. (2007). Template-based modeling and free modeling by I-TASSER in CASP7. Proteins *69* (*Suppl 8*), 108–117.

Zhang, Y. (2009a). I-TASSER: fully automated protein structure prediction in CASP8. Proteins *77* (*Suppl 9*), 100–113.

Zhang, Y. (2009b). Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. *19*, 145–155.

Zhang, Y., and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins *57*, 702–710.

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309.

Zhu, J., Fan, H., Periole, X., Honig, B., and Mark, A.E. (2008). Refining homology models by combining replica-exchange molecular dynamics and statistical potentials. Proteins *72*, 1171–1188.