Can Molecular Dynamics Simulations Provide High-Resolution Refinement of Protein Structure?

Jianhan Chen and Charles L. Brooks III*

Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

ABSTRACT **Recent advances in efficient and accurate treatment of solvent with the generalized Born approximation (GB) have made it possible to substantially refine the protein structures generated by various prediction tools through detailed molecular dynamics simulations. As demonstrated in a recent CASPR experiment, improvement can be quite reliably achieved when the initial models are sufficiently close to the native basin (e.g., 3–4 Å C_α RMSD). A key element to effective refinement is to incorporate reliable structural information into the simulation protocol. Without intimate knowledge of the target and prediction protocol used to generate the initial structural models, it can be assumed that the regular secondary structure elements (helices and strands) and overall fold topology are largely correct to start with, such that the protocol limits itself to the scope of refinement and focuses the sampling in vicinity of the initial structure. The secondary structures can be enforced by dihedral restraints and the topology through structural contacts, implemented as either multiple pair-wise C_α distance restraints or a single sidechain distance matrix restraint. The restraints are weakly imposed with flat-bottom potentials to allow sufficient flexibility for structural rearrangement. Refinement is further facilitated by enhanced sampling of advanced techniques such as the replica exchange method (REX). In general, for single domain proteins of small to medium sizes, 3–5 nanoseconds of REX/GB refinement simulations appear to be sufficient for reasonable convergence. Clustering of the resulting structural ensembles can yield refined models over 1.0 Å closer to the native structure in C_α RMSD. Substantial improvement of sidechain contacts and rotamer states can also be achieved in most cases. Additional improvement is possible with longer sampling and knowledge of the robust structural features in the initial models for a given prediction protocol. Nevertheless, limitations still exist in sampling as well as force field accuracy, manifested as difficulty in refinement of long and flexible loops. Proteins 2007;67:922–930. © 2007 Wiley-Liss, Inc.**

Key words: continuum electrostatics; generalized Born; replica exchange; structure prediction

INTRODUCTION

The last two decades has witnessed steady progress in both the prediction of protein structure from amino acid sequence and general understanding of the protein folding process.[1,2] In particular, conceptual aspects of protein folding are now considered understood, such as using the energy landscape paradigm.[3] Nevertheless, the prediction of protein structure remains one of the most important unsolved problems in computational biology and chemistry in this postgenomic era. At present, template-based modeling approaches, through either comparative modeling or threading/fold-recognition, continue to be the most reliable method for generating accurate structural models, as demonstrated in the past Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiments.[4,5] With growing efforts in structural genomics projects, more new folds are being experimentally discovered and template-based modeling is expected to become even more important compared to ab initio methods in practice.[6,7] In general, accuracy of models built by comparative modeling/threading is mainly determined by the availability of good structural templates and the ability to identify near-perfect alignments. It remains a challenging task to generate models that are closer to the native structure than the best template.[4,8] In addition, accurate prediction of the unaligned regions is typically limited to short segments.[8] Addressing these problems might require all-atom refinement with detailed energy functions. This has been recognized as one of the main areas for continual improvement of protein structure prediction.[1,8,9]

High-resolution refinement of protein structure requires both large-scale conformational sampling and accurate atomic energy functions. Much progress has been made using various statistics-based potentials that are augmented by energy terms motivated by considera-

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: National Institute of Health; Grant number: RR12255.

*Correspondence to: Charles L. Brooks III, Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037. E-mail: brooks@scripps.edu

Received 18 July 2006; Revised 14 November 2006; Accepted 16 November 2006

Published online 20 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21345

tion of important physical interactions, and sampling is often facilitated by low resolution initial conformation search and rigid body rearrangement of structural segments.[7,10] Despite this, general purpose molecular mechanical force fields[11] arguably should provide the ultimate potential functions for protein structure modeling. Particular difficulty in accurate modeling of the functional sites by templated-based approaches also argue for physics-based potentials.[8] However, their early application to structure prediction and refinement had not been successful, mainly due to difficulty in treatment of solvation effects, expensive computational cost, and prohibitive sampling requirement on highly rugged energy surface.[7,9,12] With steady increase of computational power, it has been recently shown that molecular dynamics (MD) simulations of tens to hundreds of nanoseconds (ns) in explicit water might provide meaningful refinement for proteins of small to medium sizes.[13] However, such a brute-force approach is expensive and suboptimal in terms of sampling. Instead, efficient implicit treatment of solvent[14,15] has been rapidly improved over the last few years, and, in combination with advance sampling techniques such as the Replica Exchange (REX) method,[16] might offer a reliable and effective way of high-resolution structural refinement in practice. In particular, the generalized Born (GB) approximation provides an accurate treatment of electrostatic solvation with only modest increase in the computational cost.[17,18] When parameterized carefully, GB can capture the delicate balance between solvation and intramolecular interactions and reproduce the experimental conformational equilibria of a range of peptides and small proteins.[19] The REX/GB refinement approach was recently shown to substantially improve the quality of NMR structures in presence of limited experimental data.[20,21] It is reasonable to expect a similar refinement protocol might be suitable for high-resolution refinement of predicted structures, especially when the initial models are native-like.

At current stage, there still exist severe limitations in the force field accuracy as well as sampling capability, and ab initio folding with physical energy functions has only been demonstrated with limited success for a few small and fast folding proteins.[22] Therefore, free simulation of proteins using these force fields is not expected to be an effective refinement tool in general. Instead, it is important to incorporate reliable structural information of the initial models into the refinement simulation. Without intimate knowledge of the underlying prediction protocols, one can assume that the overall topology and regular secondary structure elements are largely correct in the initial models. As such, the protocol is strictly confined to the scope of refinement and the sampling can be focused on the vicinity of the initial models. The secondary structures can be enforced through dihedral restraints and the topology through structural contacts. Such restraints also help to prevent the structures from melting at high temperature during the REX-MD simulation, which is also important for efficiency in practice. To allow sufficient flexibility for moderate structure rearrangement, all restraints should only be weakly imposed, such as with flat-bottom potentials. In the following, we will first describe the REX/GB protocol and investigate its application for high-resolution refinement of protein structures. The efficacy is illustrated using the five monomer targets in the recent CASPR model refinement experiment. Even though this is a relatively small test set, important lessons can be learned with regard to force field, sampling and other issues in refining protein structures.

METHODS AND MATERIALS

Consistent Implicit Solvent Force Field

A consistent GBSW implicit solvent force field[18,19] based on the CHARMM22/CMAP all-atom force field[23–25] was used in this work. GBSW is one of the latest GB models, which employs a van der Waals (vdW) based surface with a smooth dielectric boundary to allow stable force calculation. Born radii are calculated by a rapid volume integration scheme that includes a higher-order correction term to the Coulomb field approximation.[26] Default GBSW parameters were used with a 0.6 Å smoothing length (i.e., $w = 0.3$ Å) along with 50 Lebedev angular integration points and 24 radial integration points up to 20 Å for each atom.[18] The nonpolar solvation energy was estimated from the solvent-exposed surface area (SA) using a phenomenological surface tension coefficient of 0.005 kcal/mol/Å^2. The atomic input radii, the key physical parameters for defining the solvent boundary, have been carefully optimized together with peptide backbone torsion potentials to capture the delicate balance between solvation and intramolecular interactions.[19] Note that while parameterization of classical force fields traditionally relied primarily on experimental data and high-level quantum mechanics calculations of small molecules,[11] it recently began to benefit more directly from experimental structural and thermodynamics properties of peptides and proteins.[19,27] The consistent implicit solvent force field quantitatively reproduces the conformational equilibria of a helical peptide and four sequentially related β-hairpins, which range from largely unfolded to mostly folded at 300 K,[19] indicating proper balance of competing interactions.

Structural Restraints

For refinement, it is reasonable to assume that the initial structure is native-like and important to effectively utilize such information. Important structure features include secondary structural elements and their tertiary organization. A key consideration in implementing the structural restraints is balance between stability and flexibility: the structure should be reasonably stable even at the highest temperature simulated while allowed to substantially rearrange and move closer to the native basin. In practice, the secondary structural state of each residue can be first identified using the DSSP program[28] and then converted into weak flat-bottom harmonic restraints on the backbone ϕ and ψ torsion angles for residues within helices and β-strands,

$$E(\theta) = k_\theta \max(0, |\theta - \theta_0| - \Delta_\theta)^2, \tag{1}$$

where θ_0 is the minimum angle value and Δ is the half-width of the flat bottom potential. In this work, θ_0 is set to the value in the initial structure, $k_\theta = 50$ kcal/mol/radian^2 and Δ is set to 30 and 60° for helical and β residues respectively. These parameters are empirically chosen and they are similar to those used in NMR structure refinement.[20]

The protein fold topology can be described by the residue contact map.[29] There are multiple ways of realizing the topological restraint in term of pair-wise residue contacts. In the first implementation, a contact is formed when the minimum inter-residue distance of all pairs of heavy atoms is less than 4.2 Å in the initial model. Then, each contact is converted into a distance restraint between the C_α atoms, implemented as flat-bottom harmonic potentials similar to Eq. (1). The force constant is set to 5.0 kcal/mol/Å^2, the half-width of the flat bottom is 2.5 Å, and the minimum distance is set to the value of the initial model. In the second implementation, contacts are identified when the distance between sidechain geometric centers is within a cutoff of 6.5 Å. A single-side harmonic distance matrix restraint potential is then imposed,

$$E(\rho) = k_\rho \max(0, \rho - \rho_0)^2 \tag{2}$$

where the reaction coordinate is the fraction of native contacts broken,[30]

$$\rho = \frac{1}{N_c}\sum_i 1 - \frac{1}{1 + \exp[\gamma(r_i - r_{\text{cutoff}} - 5/\gamma)]}, \tag{3}$$

N_c is the total number of residue contacts and γ is a softness parameter that controls how rapidly the state of a contact is switched from being formed to being broken. In this work, $\gamma = 20$ is used, which allows sharp switching of contact states over about 1.0 Å range, $k_\rho = 5\ N_c$ kcal/mol/Å^2 and $\rho_0 = 0.3$ are used for all targets. Note that both realizations of topological restraint allow some balance between stability and flexibility. The sidechain distance matrix restraint in principle has an advantage of allowing some contacts to be completely broken and can accommodate large-scale local structural rearrangements. Such kind of local rearrangement might be necessary when the initial model contains large errors in local packing. On the other hand, the C_α distance restraint implementation enforces all contacts and better preserves the original topology, especially at high temperature.

REX/GB Refinement Protocol

We use the REX-MD method to achieve enhanced conformational sampling, which is enabled by the Multiscale Modeling Tools in Structural Biology (MMTSB) Tool Set[31,32] (available from http://mmtsb.scripps.edu) together with the CHARMM program.[33] Briefly, multiple copies (replicas) of the system are simulated at different temperatures independently and simultaneously. Exchanges of simulation temperatures are periodically attempted according to a Metropolis type algorithm. In the course of the REX simulation, replicas can travel up and down the temperature space in a self-regularized fashion, which, in turn, induces a nontrivial random walk in temperature space and greatly reduces the probability of being trapped in states of local energy minima. In addition, low energy conformations have a higher probability of occupying the low temperature windows and this provides a convenient way of selecting the best structures sampled during the course of the REX simulation.

In this work, we used 24 replicas covering a temperature range of 270–600 K for all targets. The temperatures were distributed exponentially within the specified ranges. SHAKE was applied to fix the lengths of all bonds with hydrogen atoms and a time-step of 2 fs was used. Exchanges of simulation temperatures were attempted every 2.0 ps of restrained MD. The total simulation length is around 3 ns unless otherwise specified. The overall exchange ratios of these simulations were about 30%. At the end, conformations sampled at the lowest temperature (270 K) during the last 1 ns (500 structures) were clustered to provide the refined models. A hierarchical clustering algorithm based on C_α mutual RMSD is used and the number of clusters is determined automatically up to a maximum number of four (MMTSB/cluster.pl). Ranking of the refined models is solely determined by cluster size, which is in principle governed by the free energy in REX simulations.[16]

Structure Preparation and Analysis

All five monomer targets of the CASPR model refinement experiment were refined using the REX/GB protocol. They are summarized in Table I. Note that two other targets (TMR02 and TMR03) with dimeric biological units are not included here as the structures seem to intimately depend on the dimerization. Residues in the target sequences that do not have correspondence in the PDB structures[34] were not included in this study, except for TMR06 where missing coordinates of residues 50–52 of PDB 1xg8 were rebuilt using CHARMM to provide a continuous peptide chain. The subsets of residues included in the refinement are given in Table I. Both termini were left uncapped for all targets. Initial models and PDB structures were relaxed in the GBSW implicit solvent by energy minimization with harmonic restraints on the backbone heavy atoms. The minimized models are within 0.2 Å backbone RMSD from the originals and all the structural properties shown in Table I were computed using the minimized structures. Additionally for targets TMR01 and TMR04, core domains are defined by excluding the terminal coils, as it is not clear how repre-

TABLE I. Summary of Backbone Structural Properties

				C_α RMSD (Å)			GDT_TS		
Target	N_{seq}	Subset	PDB	Init	Ref-1	Ref-2	Init	Ref-1	Ref-2
TMR01	116	26–116	1xe1	6.12 (2.24)	3.94 (1.53)	5.00 (1.54)	79.4	83.5	82.7
					5.70 (1.94)	4.29 (1.79)		78.6	83.0
						4.49 (2.57)			73.6
TMR04	70	1–70	1whz	2.18 (2.04)	1.88 (1.26)	1.57 (0.98)	76.1	85.7	90.0
					1.78 (1.55)	1.81 (1.52)		82.3	83.9
					2.52 (2.27)	4.01 (3.35)		72.5	82.1
TMR05	144	4–140	1tvg	3.92	3.31	3.30	74.1	70.4	69.7
					3.02	5.30		75.7	59.3
					2.81	4.04		72.3	64.1
					3.68	5.01		65.7	59.7
TMR06	102	2–102	1xg8	3.40	4.71	4.95	61.6	55.5	51.2
					4.56	5.17		58.9	48.8
					4.07	8.66		55.2	36.6
					3.85			57.9	
TMR07	124	1–107	1o13	2.22	2.72	3.29	78.0	81.1	68.1
					2.21	3.69		77.1	68.5
					2.82			68.7	

For each target, N_{seq} is the number of residues in the target sequence. All the structural refinement and evaluation only include the subset of residues with experimental data, given using the numbering convention of the initial model. The numbers in the parenthesis of C_α RMSD column are the values for the core domains: residues 34–116 for TMR01 and residues 6–70 for TMR04. Init columns show the properties of minimized initial models. Columns Ref-1 and Ref-2 are for the REX/GB refined models obtained using the two implementations of topological restraints: Ref-1 as C_α distance restraints and Ref-2 as sidechain distance matrix restraints. For each target, multiple rows correspond to the properties of the centroids of different clusters, given in the order of decreasing cluster size from top down (i.e., model 1 up to model 4).

sentative the X-ray structures are for the solution conformations of these residues. This issue will be discussed further in the next section. GDT_TS score,[35] a more reliable measurement of backbone structural similarity, was computed using the MMTSB/gdt tool. All the figures were created using VMD.[36]

RESULTS AND DISCUSSION

Backbone Refinement

Table I summarizes the results of REX/GB refinement using the C_α distance restraint (Ref-1) and sidechain distance matrix restraint (Ref-2) with a "standard" sampling time of 3 ns. The energy at the lowest temperature typically reaches a plateau value within 2 ns. However, convergence of conformational sampling is slower and less obvious. Issues regarding convergence and impact of sampling time will be discussed further in the next section. The backbone quality of the models is accessed in terms of C_α RMSD value and GDT_TS similarity score. Clearly, there is substantial improvement for most targets reflected in both structural measurements. With C_α distance restraints, lower RMSD values were obtained in the refined model 1 for three targets, TMR01, TMR04, and TMR05, and higher GDT_TS scores are observed in all targets except TMR06. In most cases, the first model, derived from the largest cluster of the production ensemble, is also the most native-like conformation, reflecting the ability of the GB implicit solvent force field in identifying the native fold. For target TMR05, the most native-like model is model 2 instead of model 1. This is related to limited convergence of sampling and will be further addressed in the next section. It is particularly encouraging that significant improvement is possible even for β-structures such as TMR01 and TMR05, which has been traditionally difficult for physical force fields. The amount of improvement depends on both the protein topology and nature of the initial models. Closer examination of the two most difficult targets (TMR06 and TMR07) shows that both targets contain several long and flexible loops and require substantially greater sampling. In addition, there exists ambiguity regarding whether the X-ray models are indeed representative of the solution conformations. These issues will be also discussed further in the following sections. Both implementations of the topological restraints seem to be effective in allowing the protein to move closer to the experimental native basin. For the "refinable" targets (such as TMR01 and TMR04), different implementations and parameters of restraint potential appear to have minimal impact. On the other hand, for the more difficult targets, there is significant dependence on the details of restraint potential. With sidechain distance matrix restraints, alternative basins of the energy landscape are more likely to be sampled. Such a in principle desirable feature might become disadvantageous in the refinement practice (such as for TMR05 and TMR07), mainly due to force field limitation. This will be further discussed in the following sections.

Figures 1 and 2 illustrate the backbone structural improvement of the refined models (model 1) of TMR01 and TMR04. TMR01 is a β-barrel protein. The initial model is actually more accurate than what an overall C_α

RMSD value of 6.12 Å indicates. The core domain, defined by excluding the coiled N-terminal tail (resides 26–33), has a C_α RMSD value of only 2.24 Å. After refinement, the core domain C_α RMSD is reduced to 1.53 Å, due to better packing of the β-strands as well as improved loop conformations [Fig. 1(a)]. While the N-terminal tail also swung toward the X-ray model and thus reduced the overall RMSD to below 3.94 Å, it does not necessarily represent a meaningful improvement. There is no evidence that the N-terminal tail adopts the same well defined packing against the core domain in solution. It is more likely that the terminal tails as well as long loops are mobile and such intrinsic flexibility might be essential for protein function, interaction and recognition.[37] TMR04 is a small α/β protein, for which a model of 1.6 Å C_α RMSD was built and celebrated as the most accurate blind de novo structure prediction so far.[7] The REX/GB refinement reduced the overall C_α RMSD from 2.18 Å to 1.88 and 1.57 Å, respectively, with two different implementations of topological restraints. By excluding the first five residues in the N-terminal coil, the core C_α RMSD is reduced from 2.04 Å to a respectable 1.25 Å in Ref-1 and 0.98 Å in Ref-2. Improved packing of the C-terminal helix as well as the β-strands is evident, as depicted in Figure 2. Consistent with these observations, the GDT_TS score is improved from 76.1 to 85.7 and 90.0, respectively.

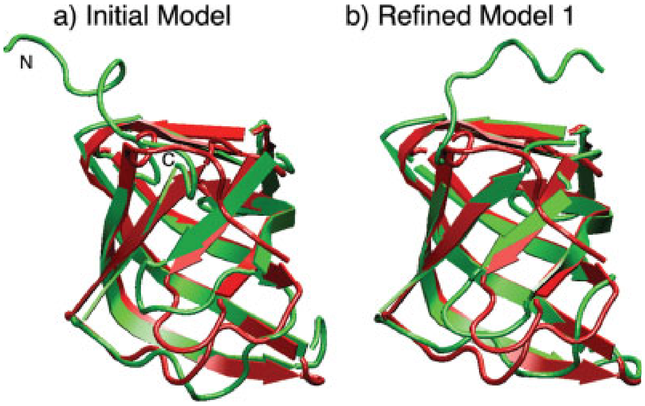

Fig. 1. Cartoon representation of the initial and refined models of target TMR01, show in green. The X-ray structure is shown in red for reference. The N- and C-terminals are marked with letters “N” and “C” respectively.

Convergence of Conformational Sampling

Control REX simulations were carried out for all targets to examine the convergence and impact of extended sampling on refinement. The control simulations were initiated from the PDB structures but with the same structural restraints derived from the initial CASPR models. The sampling time was 2 ns for each target and the last 300 conformations sampled (0.6 ns) at the lowest temperature were clustered. Then, the centroid of the largest cluster (i.e., control model 1) was compared with the PDB structure and REX/GB refined model. With sufficient sampling, the control simulation should converge to the same structural basin as the refinement simulation, such that the control and refined models are close and of similar distance from the PDB structure. Summarized in Table II, the results indicate that good convergence has only been achieved for targets TMR01 and TMR04. For the rest, models from the control simulations remain much closer to the PDB structures than to the refined models, indicating limited convergence. Interestingly, the control, refined, and PDB models of TMR06 have the largest mutual distances, reflecting particular difficulty of modeling this protein.

To investigate the impact of longer sampling on refinement, the REX/GB simulation was extended to 5 ns for TMR05. Clustering the last 500 structures sampling at the lowest temperature yielded two models, with C_α RMSD of 2.78 and 2.92 Å, respectively. The corresponding GDT_TS scores are 76.1 and 71.9. Therefore, for TMR05, extended sampling does not only move the structure closer to the native basin but also select the more native-like centroid as the first model. The structures of the initial and refined models are shown in Figure 3. Gradual improvement with longer sampling time is evident in loop conformation and packing of the structured core.

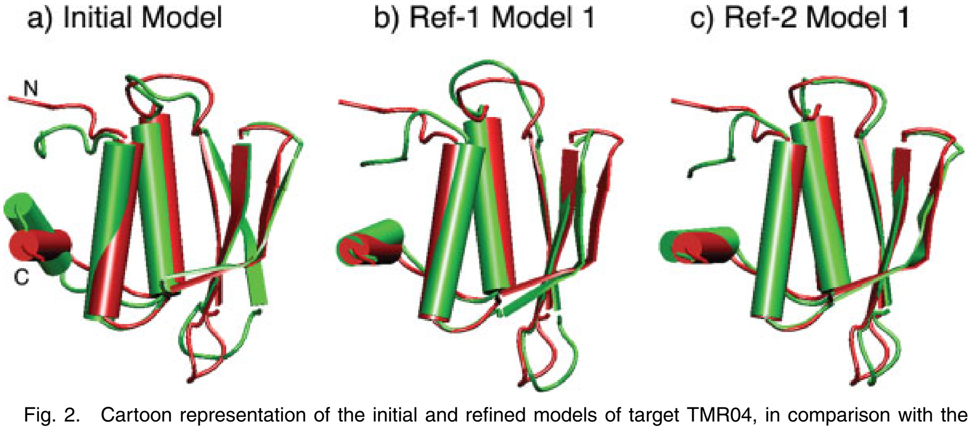

Fig. 2. Cartoon representation of the initial and refined models of target TMR04, in comparison with the X-ray structure shown in red.

TABLE II. Summary of Control Simulations

	C_α RMSD (Å)			GDT_TS		
Target	cntrl-Exp	ref-exp	cntrl-ref	cntrl-exp	ref-exp	cntrl-ref
TMR01	1.67	1.53	1.43	86.5	83.5	84.1
TMR04	1.04	1.26	1.16	91.8	86.8	89.3
TMR05	0.77	3.31	3.13	96.2	70.4	72.3
TMR06	2.58	4.71	3.73	73.3	55.5	63.6
TMR07	1.85	2.72	2.48	90.2	81.1	82.2

Notations: cntrl: model 1 from control simulations; ref: REX/GB refined model 1 with 3 ns sampling time; exp: experimental structure. RMSD values were computed from the core domains for TMR01 and TMR04 (as defined in Table 1) and all residues for TMR05, TMR06 and TMR07.

Accuracy of Sidechain Contacts and Rotamer States

The CASP experiments have shown that sidechain prediction strongly depends on the backbone quality and is typically less accurate.[8] Here, refinement of sidechains has been evaluated by examining sidechain contacts and χ_1 rotamer states, summarized in Table III. Indeed, the initial models generally contain low fractions of native sidechain contacts and native χ_1 rotamer states. Both properties were significantly improved with the REX/GB refinement in most cases. In particular, even though the REX/GB protocol failed to refine the backbone of target TMR06, it did improve the accuracy of χ_1 rotamer states to a level comparable to that of other targets. Improvement of sidechains appears to be smaller with the distance matrix realization of topological restraint for the difficult targets such as TMR06 and TMR07. This is probably due to fact that the distance matrix restraint is directly imposed on sidechains, which tends to restrict their movement.

Remaining Problems

Significant difficulty still exists for modeling coiled terminal residues and flexible long loops. Both targets TMR06 and TMR07 contain significant portion of residues in long loops, as depicted in Figure 4. The REX/GB refinement barely improved TMR07 and moved TMR06 further away from the experimental structure. The difficulty does not only arise from greater sampling requirement, but also reflects certain limitations of the underlying GB force field. These loops and tails are typically flexible and the conformational equilibrium is governed by delicate balance of solvation, intramolecular interactions and entropy. Small errors in the force field can easily translate into large shift of conformational equilibrium. For example, TMR07 contains several long loops connecting a structured core that consists of five strands and three helices. The structured core is only 0.94 Å away from the X-ray structure and the loops are the key regions for refinement. The REX/GB refinement simulation (Ref-1 of Table I) well maintained the structure of the core (within 1.5 Å C_α RMSD) and sampled a broad loop conformational space, which was clustered into three basins. The dominant cluster (64% populated) is centered around a basin that is further away from the experimental structure than the initial model. Even though the core C_α RMSD is reduced to 0.91 Å and the GDT_TS score is improved to 81.1, the loop C_α RMSD increases from 2.84 Å to 3.33 Å. The center of the second largest cluster (20% populated) is slightly closer to the native basin, with a reduced loop C_α RMSD of 2.71 Å but increased core C_α RMSD of 1.19 Å. It is not clear whether the force field accuracy or sampling is the main limiting factor in this case. Nevertheless, it should also be pointed out that there is ambiguity in the X-ray structure's ability of representing the solution conformation for flexible segments, which further complicates the analysis. Because of unknown reasons, TMR06 is unusually difficult to model using the GBSW implicit solvent field and the X-ray structure of TMR06 itself is not stable. Therefore, we will not attempt to further rationale the poor refinement performance other than simply noting the loose packing of the PDB structure.

Additional problem also remains in identifying the optimal way of imposing the topological restraints. Even with only five targets, there are already clear indications that different targets and initial structures might require different restraint potentials. Some targets such as TMR04 prefers the distance matrix restraint which provides more freedom for local structural rearrangement, while other targets such as TMR05 are more effectively refined with contact derived C_α distance restraints. There are further complications with choosing the exact parameters for restraint potentials, particularly in the distance matrix implementation. Two key parameters, maximum allowed fraction of native contacts broken, ρ_0, and, contact switching parameter, γ, control how much of and how strongly the initial contact pattern should be maintained. Even though we only present the results with a reasonable setting here ($\rho_0 = 0.3$, $\gamma = 20$), numerical experiments indicate some dependence of refinement performance on these parameters for more difficult targets such as TMR05. In practice, where the true native structure is not known, it will not be trivial to determine the optimal setting for achieving maximal improvement. Nonetheless, the contact derived distance restraints appear to be a conservative choice where substantial refinement can be quite reliably achieved in most cases.

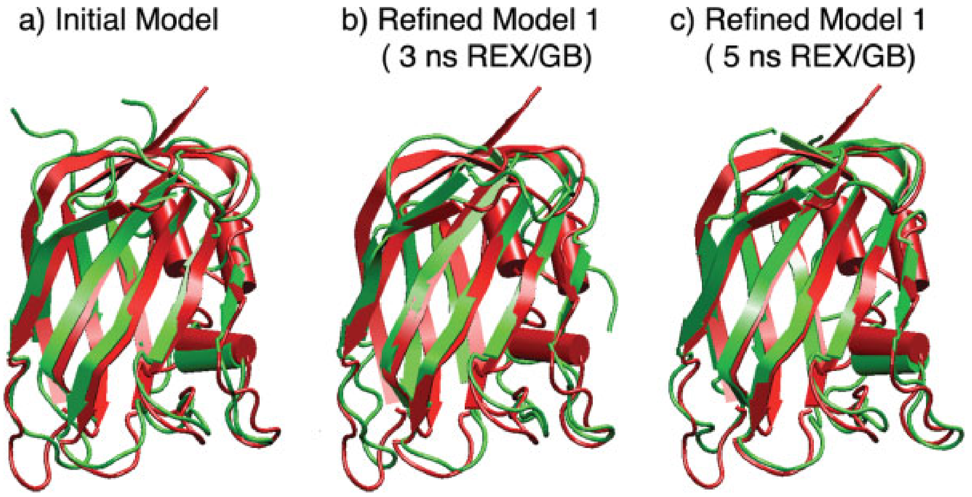
a) Initial Model
b) Refined Model 1 (3 ns REX/GB)
c) Refined Model 1 (5 ns REX/GB)

Fig. 3. Cartoon representation of the initial and refined models of target TMR05, in comparison with the X-ray structure shown in red.

TABLE III. Fraction of Sidechain Native Contacts and χ_1 Rotamer States

	Contacts			χ_1 Rotamer states[a]		
Target	Init	Ref-1	Ref-2	Init	Ref-1	Ref-2
TMR01	0.71	0.73	0.75	0.44	0.55	0.53
TMR04	0.52	0.73	0.81	0.48	0.62	0.72
TMR05	0.63	0.66	0.61	0.53	0.59	0.54
TMR06	0.52	0.50	0.40	0.27	0.52	0.42
TMR07	0.58	0.75	0.59	0.45	0.51	0.50

[a]Fraction of c angles in the models found within 40 degrees of the corresponding angles in the PDB structure.

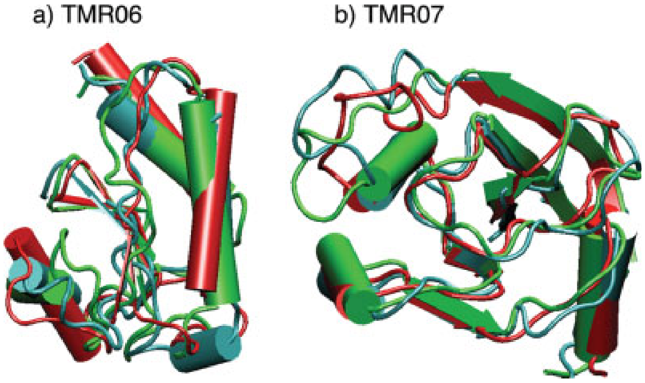

Fig. 4. Cartoon representation of the initial (blue) and refined models (green) of target TMR06 and TMR07, in comparison with the X-ray structures (red).

The limitations discussed earlier partially originate from imperfections in the underlying force field. If the force field were nearly perfect, sensitivity to the choice of restraint potential would diminish. In particular, as demonstrated previously in refinement of NMR structures,[20,21] the solvation energy term is maily responsible for driving the improvement in the models. Currently, there remain limitations in the implicit treatment of the nonpolar solvation in the current GB force field, where the nonpolar solvation free energy is estimated from the solvent exposed surface area using a single phenomenological surface tension coefficient. Such a simple nonpolar solvation model is not sufficient to capture the detailed balance between intramolecular dispersion interactions and solvent vdW screening.[38] Examination of pairwise and three-body interaction of nonpolar sidechains reveals a systematic over stabilization of pairwise interaction the GB/SA models and also points to difficulty in modeling cooperativity of hydrophobic association (Chen and Brooks, unpublished results). These issues are important in distinguishing the true native fold from protein-like compact misfolds and in high-resolution refinement of tertiary packing.

CONCLUSIONS

We have demonstrated that meaningful high-resolution structure refinement of protein structure can be achieved with careful MD simulations in modern molecular mechanical force fields. Improvement of backbone was achieved in four out of the five CASPR monomer targets and sidechain contacts and rotamer states were improved in all targets. In the best case, we successfully refined the core domain of a small protein (TMR04) to sub-Angstrom C_α RMSD. This has become possible due to recent advances in efficient characterization of solvent effects through the GB/SA implicit solvent and development of advanced sampling techniques such as the REX method. Another key to effective refinement is to incorporate reliable structural information during the simulation and thus focus the conformational sampling in the vicinity of the initial models. Such structural information may include secondary structure elements and the tertiary fold. The secondary structure state can be imposed as backbone dihedral restraints and the tertiary fold as residue contact restraints, implemented either as multiple C_α distance restraints or as a single sidechain distance matrix restraint. All restraints are weakly imposed with flat-bottom harmonic potentials to allow sufficient flexibility for structural movements. For small to medium-sized proteins, it appears that 3–5 ns of REX/GB refinement is sufficient to achieve substantial improvement when the initial model is sufficiently native-like (e.g., within 4 Å C_α RMSD). The consistent GBSW force field[19] employed in this study appears to be capable of identifying the native basin and the dominant cluster of the production structural ensemble is typically the most native-like.

Nevertheless, significant difficulty still exists in both sampling and force field accuracy. This is mainly reflected in the poor refinement performance of long loops and the sensitivity of performance to details of the topological restraints for difficult targets. One of the specific limitations of the GB/SA implicit solvent model is the nonpolar solvation term, where the balance between intramolecular dispersion interactions and solvent vdW screening is not well captured. Such force field limitations are the main reason for the dependence of refinement performance on restrain potentials. Based on the limited experience presented here, it appears that contact-derived C_α distance restraint is a conservative realization of the topological restraint that provides the most robust performance. In practice, one might expect further improvement with the knowledge of which structural features are the most accurate from the prediction protocols. It is clear that the REX/GB refinement protocol is still computationally expensive and can introduce meaningful improvement only when the starting model is reasonably close to the native basin. How close is enough is a difficult question and the answer depends on the specific protein and the models. Without the knowledge of the experimental structure, it will be difficult to determine whether the initial models are "refinable" or not. Furthermore, it is typically nontrivial to derive confidence criteria that indicates the refinement performance, while the limitations discussed earlier might be used to identify problematic cases. We are currently expanding the number of test targets by participating in the CASP7 bind prediction experiment as a refinement team. Lessons from these experiments will expand our current understanding of the high-resolution structural refinement problem and further improve the REX/GB refinement protocol. In conclusion, we expect physical force field based refinement protocols to play a more significant role in the protein structure prediction practice.

REFERENCES

1. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310:638–642.
2. Shakhnovich E. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. Chem Rev 2006; 106:1559–1588.
3. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science 1995;267:1619–1620.
4. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins 2003;53:585–595.
5. Kryshtafovych A, Venclovas C, Fidelis K, Moult J. Progress over the first decade of CASP experiments. Proteins 2005;61(S7):225–236.
6. Zhang Y, Skolnick J. The protein structure prediction problem could be solved by using the current PDB library. Proc Natl Acad Sci USA 2005;102:1029–1034.
7. Bradley P, Misura K, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 2005;309:1868–1871.
8. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A. Assessment of predictions submitted for the CASP6 comparative modeling category. Proteins 2005;61(S7):27–45.
9. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
10. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.
11. MacKerell AD, Jr. Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 2004;25:1584–1604.
12. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10:139–145.
13. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Sci 2004;13:211–220.
14. Roux B, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
15. Feig M, Brooks CL, III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol 2004;14:217–224.
16. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 1999;314:141–151.
17. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
18. Im W, Lee MS, Brooks CL, III. Generalized Born model with a simple smoothing function. J Comput Chem 2003;24:1691–1702.
19. Chen J, Im W, Brooks CL, III. Balancing solvation and intramolecular interactions: towards a consistent generalized Born force field. J Am Chem Soc 2006;128:3728–3736.
20. Chen J, Im W, Brooks CL, III. Refinement of NMR structures using implicit solvent and advanced sampling techniques. J Am Chem Soc 2004;126:16038–16047.
21. Chen J, Won H-S, Im W, Dyson HJ, Brooks CL. Generation of native-like models from limited NMR data, modern force fields and advanced conformational sampling. J Biomol NMR 2005;31: 59–64.
22. Kubelka J, Hofrichter J, Eaton WA. The protein folding 'speed limit'. Curr Opin Struct Biol 2004;14:76–88.
23. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
24. Feig M, MacKerell AD, Jr, Brooks CL, III. Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations. J Phys Chem 2003;107:2831–2836.
25. MacKerell AD, Jr, Feig M, Brooks CL, III. Improved treatment of the protein backbone in empirical force fields. J Am Chem Soc 2004;126:698–699.
26. Lee MS, Salsbury FR, Jr, Brooks CL, III. Novel generalized Born methods. J Chem Phys 2002;116:10606–10614.
27. MacKerell AD, Jr, Feig M, Brooks CL, III. Extending the treatment of backbone energetics in protein force fields: limitation s

of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400–1415.

28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
29. Mirny L, Domany E. Protein fold recognition and dynamics in the space of contact maps. Proteins 1996;26:391–410.
30. Sheinerman FB, Brooks CL, III. Calculations on folding of segment B1 of streptococcal protein G. J Mol Biol 1998;278:439–456.
31. Feig M, Karanicolas J, Brooks CL, III. 2001. MMTSB Tool Set, MMTSB NIH Research Resource, The Scripps Research Institute.
32. Feig M, Karanicolas J, Brooks CL, III. MMTSB tool set: Enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graph Model 2004;22:377–395.
33. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
35. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
36. William Humphrey, Andrew Dalke, Klaus Schulten. VMD—Visual Molecular Dynamics. J of Mol Graph 1996;14:33–38.
37. Uversky VN, Oldfield1y CJ, Dunker AK. Showing your ID: intrinsic disorder as an id for recognition, regulation and cell signaling. J Mol Recognit 2005;18:343–384.
38. Levy RM, Zhang LY, Gallicchio E, Felts AK. On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute-solvent interaction energy. J Am Chem Soc 2003;125:9523–9530.