Progress and Challenges in High-Resolution Refinement of Protein Structure Models

Kira M.S. Misura and David Baker*
Department of Biochemistry, University of Washington Health Sciences, Seattle, Washington

ABSTRACT **Achieving atomic level accuracy in *de novo* structure prediction presents a formidable challenge even in the context of protein models with correct topologies. High-resolution refinement is a fundamental test of force field accuracy and sampling methodology, and its limited success in both comparative modeling and *de novo* prediction contexts highlights the limitations of current approaches. We constructed four tests to identify bottlenecks in our current approach and to guide progress in this challenging area. The first three tests showed that idealized native structures are stable under our refinement simulation conditions and that the refinement protocol can significantly decrease the root mean square deviation (RMSD) of perturbed native structures. In the fourth test we applied the refinement protocol to *de novo* models and showed that accurate models could be identified based on their energies, and in several cases many of the buried side chains adopted native-like conformations. We also showed that the differences in backbone and side-chain conformations between the refined *de novo* models and the native structures are largely localized to loop regions and regions where the native structure has unusual features such as rare rotamers or atypical hydrogen bonding between β-strands. The refined *de novo* models typically have higher energies than refined idealized native structures, indicating that sampling of local backbone conformations and side-chain packing arrangements in a condensed state is a primary obstacle. Proteins 2005;59:15–29.**

Key words: protein structure prediction; model refinement; Rosetta; free energy function

INTRODUCTION

Substantial progress has been made in the area of *de novo* structure prediction; it is now possible to generate models with correct topologies for small proteins using several different methods, including the Rosetta *de novo* algorithm.[1–4] In many cases, features of native proteins such as turns, loops and relative orientations of secondary structure elements are captured in the *de novo* models. However, the overall accuracy of the models is not sufficient for applications requiring high-resolution detail. Even for small proteins of fewer than 100 amino acids, the root mean squared deviation (RMSD) over alpha carbon atoms of the native structure to the *de novo* models is typically greater than 3 Å. In addition, most energy functions cannot reliably distinguish models with the correct topology from those with non-native topologies. This is illustrated by the CASP4 and CASP5 experiments; while one of the five models generated by Rosetta and submitted as predictions often had the correct topology, it was frequently not the best-ranked model.

To increase the accuracy and reliability of protein structure models, it is necessary to develop methods that sample high-resolution details of native structures as well as potential energy functions that recognize the native state as the lowest energy conformation. One approach to this problem is to refine low-resolution models produced by *de novo* or template-based modeling methods. Successful refinement would improve *de novo* or template-based models by shifting their conformations closer to the native state; equally importantly, they would allow models with correct topologies and side-chain packing to be distinguished from non-native models based on their relative energies. Energy based discrimination of models greater with than 3 Å RMSD to the native structure is problematic as the native side-chain packing arrangement is unlikely to be captured.[5] High-resolution refinement would benefit *de novo* structure prediction as well as comparative modeling applications, where it is desirable to generate models that are more similar to the native structure than the starting template.

High-resolution refinement is a difficult task that requires an effective sampling strategy as well as an accurate energy function to guide the search through conformational space. Attempts to refine protein structure models into native-like conformations have been made previously. Lee et al. used molecular dynamics simulations with an explicit solvent model to refine Rosetta *de novo* models followed by scoring with the Poisson–Boltzman surface area solvation model.[6] Their results showed that native structures could be distinguished from low-resolution models and that the native state is stable. Lu et al. used a combination of local constraints, knowledge-based potentials and molecular dynamics approaches.[7] While these results were promising and showed improvements over

*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Box 357350, J-567 Health Sciences, Seattle, WA 98195-7350. Email: dabaker@u.washington.edu

Received 22 July 2004; Accepted 16 September 2004

Published online 2 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20376

previous studies using standard molecular dynamics methods,[6] the goal of producing models of atomic-level accuracy was not achieved. Fan and Mark also applied molecular dynamics methods to Rosetta *de novo* models.[8] These authors showed that molecular dynamics methods increased the RMSD value of the starting model during short simulations, but longer simulations appeared to generate tighter packing of helices and regularization of β-strands in some cases.

Previously, our group developed and used a refinement protocol to create a structurally diverse set of models with all heavy atoms represented explicitly, and we optimized the energy function to discriminate near-native from non-native models.[9] Building on this work, in this study we have further developed the energy function and refinement protocol. To facilitate refining protein models towards their native conformations, we designed four tests of increasing difficulty and evaluated the performance of the Rosetta refinement protocol in each of the tests. We then examined the lowest RMSD and a low energy refined *de novo* model for each of the test set proteins and characterized the structural and energetic differences between the native structure and the refined models.

MATERIALS AND METHODS

Test Set and Generation of *De Novo* Models

The test set consisted of 10 proteins ranging from 49 to 106 amino acid residues. We selected small proteins due to the large amount of computational time required for each refinement simulation and to ensure that a significant number of near-native (<3 Å RMSD) *de novo* models were generated. Initial models were generated using the Rosetta *de novo* protein structure prediction algorithm.[1,10] For each protein 10,000 *de novo* models were generated with side chains approximated as centroids.

Refinement Protocol

The full refinement protocol consisted of a low-resolution step followed by a high-resolution step. The low-resolution step was designed primarily to remove steric clashes between backbone atoms and to improve β-sheet hydrogen bonding in the *de novo* models, and the high-resolution step was designed to sample backbone conformations more finely and search low energy side-chain conformations compatible with each backbone conformation. In the low-resolution refinement step, side chains were represented by centroid interaction centers. The parameters defining the closest approach of centroids and backbone atoms were taken from the 25th closest approach distance observed in high-resolution crystal structures, and the penalty for violating these constraints was slowly increased during the simulation. The bond lengths and angles were fixed at ideal values, and the polypeptide chain was represented in internal coordinates as the backbone torsion angles ϕ, ψ and ω. To minimize the energy as a function of these degrees of freedom, we utilized a Monte Carlo minimization (MCM) strategy with the following steps:[11] (1) a random perturbation to the current values of the backbone torsion angles, (2) optimization of the torsion angles flanking the site(s) of the original perturbation using the Davidon-Fletcher-Powell (DFP) algorithm[12] and (3) acceptance or rejection of the new angles based on the energy difference between the final minimized conformation and the initial conformation prior to the random perturbation using the standard Metropolis criterion. The initial random perturbation consisted of either a series of small random changes in backbone torsion angles or a replacement of the torsion angles of one or three consecutive residues with those from a fragment from the Protein Data Bank (PDB) followed by variation of adjacent torsion angles to minimize the mean square deviation of atoms brought about by the insertion.[11]

In the high-resolution step, the representation of the polypeptide chain was the same as in the low-resolution step except that all side-chain atoms were represented explicitly and the side-chain torsion angles $\chi1$, $\chi2$, $\chi3$, and $\chi4$ were included. The energy was minimized as a function of these degrees of freedom using the same MCM strategy as in the low-resolution refinement protocol, with the addition of a side-chain rotamer optimization step after the initial random perturbation. Low energy side-chain conformations were obtained by cycling through the Dunbrack rotamer library[13] for each residue. The gradient computation for the DFP optimization was facilitated by an efficient recursive algorithm for computing the derivatives of the orientation dependent hydrogen bonding term with respect to the backbone and side-chain torsion angles.[14]

The high-resolution refinement step was divided into three sub-sections. The first aimed at removing atomic clashes while maintaining compactness of the structures. The weight on the repulsive component of the Lennard–Jones potential was gradually increased after each cycle of 10 attempted perturbations, starting at 1/50 of its final value. Conservative backbone perturbations ("small" and "shear" moves) were attempted, followed by optimization of the backbone and side-chain torsion angles. During the second part, sets of conservative and moderate angle perturbations were attempted ("small," "shear," "wobble" and "crank" moves), and complete combinatorial optimization of the side-chain conformations was performed after every 25 attempted moves using a simulated annealing method described previously.[15] We used an expanded rotamer set for all repacking trials compared to our previous study,[9] and buried residues were allowed additional rotamer choices relative to exposed residues. For residues with more than 10 side chains within 7–11 Å (dependent on amino acid type), all rotamers with frequencies of occurrence of greater than .01 in the Dunbrack backbone dependent rotamer library were included. These conformations were supplemented with additional rotamers generated by varying $\chi1$ and $\chi2$ by ±1 standard deviation for aliphatic residues and by $\pm1/3$ and $\pm2/3$ standard deviation for aromatic residues. The number of attempted MCM moves of each type was equal to the number of residues in the protein. The final part of the simulation is similar to the second part except that side-chain χ-angles were included with the backbone tor-

sion angles in the DFP optimization to allow off-rotamer conformations to be sampled.

Free Energy Function

The free energy function in the low-resolution refinement step was similar to that used in the *de novo* folding protocol[16] with the addition of orientation and secondary structure dependent backbone hydrogen bonding potentials.[17] The atomic radii and free energy function used in the high-resolution step is described in Kuhlman et al.,[18] with the following modifications: the weight given to the implicit solvation model energy was reduced from 1.0 to 0.5, the weight given to the Ramachandran energy was reduced from 0.10 to 0.05 and the weight given to the Dunbrack energy was reduced from 1.0 to 0.5. The repulsive region of the Lennard–Jones 12–6 potential was modified such that the energy increased linearly from 0.6× the sum of the Van der Waals radii to 0.0 Å. This functional form reduced the number of clashes in the refined models relative to those generated in Tsai et al.[9] (data not shown). After the first round of refinement, we found that a large percent of the hydrogen bond energy was derived from short-range hydrogen bonds where the donor and acceptor were separated by fewer than four residues, corresponding mainly to α-helices. Reducing the weight on backbone hydrogen bonds in helices by a factor of two in proteins predicted to be helical and by a factor of four in all other proteins improved discrimination of near-native from non-native models (data not shown). These changes were incorporated into the second round of refinement.

Idealized Native Refinement Tests (Tests 1 and 2)

The native bond lengths and angles of the native structure were replaced with ideal values, and the distance matrix error to the native structure was minimized for residue pairs within 10 Å of each other in the native structure by optimizing the backbone torsion angles using the DFP algorithm.[12] The resulting idealized native structures were then subjected to the refinement protocol. To include native side-chain information, the Dunbrack rotamer library was supplemented with the native rotamer conformations when the side chains were repacked. Native rotamers were replaced by substitution if a non-native rotamer was found to have a lower energy at that position. To generate refined idealized native structures without side-chain information, the side chains were removed from the idealized native structure prior to the refinement step described above.

Perturbed Native Refinement Test (Test 3)

To generate the starting structures, native bond lengths were replaced with idealized values, but the native ϕ and ψ torsion angles were retained. In contrast to the idealized native structures described above, no attempt was made to minimize structural changes. For each protein in the test set a single starting conformation was subjected to 50 independent simulations using the refinement protocol.

De Novo Model Refinement Test (Test 4)

For each of 10,000 *de novo* models generated for a target sequence, five independent simulations were performed using the refinement protocol. For the larger and more complex proteins in the test set, computational considerations limited our ability to refine all the initial models, and in these cases random subsets of the initial set were chosen for refinement. The resulting refined model populations contained 25,000 to 50,000 models for each protein.

RESULTS AND DISCUSSION

With the goal of refining protein structure models to more closely resemble their native conformations, we devised four tests of increasing difficulty to evaluate the effectiveness of the Rosetta search strategy and energy function. We applied each test to a set of 10 small proteins for which x-ray structures have been determined. The first and second tests subjected the native structures to bond length idealization and rotamer optimization, testing the stability of the idealized native state under the Rosetta energy function. They were designed to test the protocol in the energy landscape immediately surrounding the native state. Rosetta uses approximations such as idealized bond lengths and angles as well as a backbone dependent rotamer library in order to represent side-chain conformations, and these tests also allow us to assess the effects of these approximations. The third test involved perturbations to the native structures. The fourth and most challenging test applied the refinement protocol to *de novo* models generated using Rosetta.

Test 1: Idealization and Native Structure Refinement Including Native Side-Chain Information

This test was designed to explore the effects of bond length and angle idealization on the native structures and determine whether the idealized native state is stable under our energy function and simulation conditions. For the idealized native state to be considered stable, we required that the backbone move no farther than 0.5 Å RMSD from the native structure and that the majority of the side chains remain in their native conformations during the whole of the simulation. We generated native structures with idealized backbones and native side-chain torsion angles for each protein in the test set (described in Methods) and carried out 10 independent refinement runs for each idealized native structure. During the simulation, the side-chain χ-angles were replaced with rotamers from the library[13] if lower energy conformations were found.

The refined models were found to be structurally similar to the native conformation (Fig. 1, Δ RMSD $\approx$0.5 Å) and to each other [Fig. 1, Fig. 2(a,c), Δ RMSD $\approx$0.1 Å] and to have a narrow range of energies (Fig. 1, black points). These results show that using idealized bond lengths and angles is not a prohibitive constraint when attempting to accurately model the finely sampled ranges of torsion angles observed in experimentally determined protein structures. The side chains remain in their native conformations, as shown in Figure 2(a,c), and are not replaced with rotamers

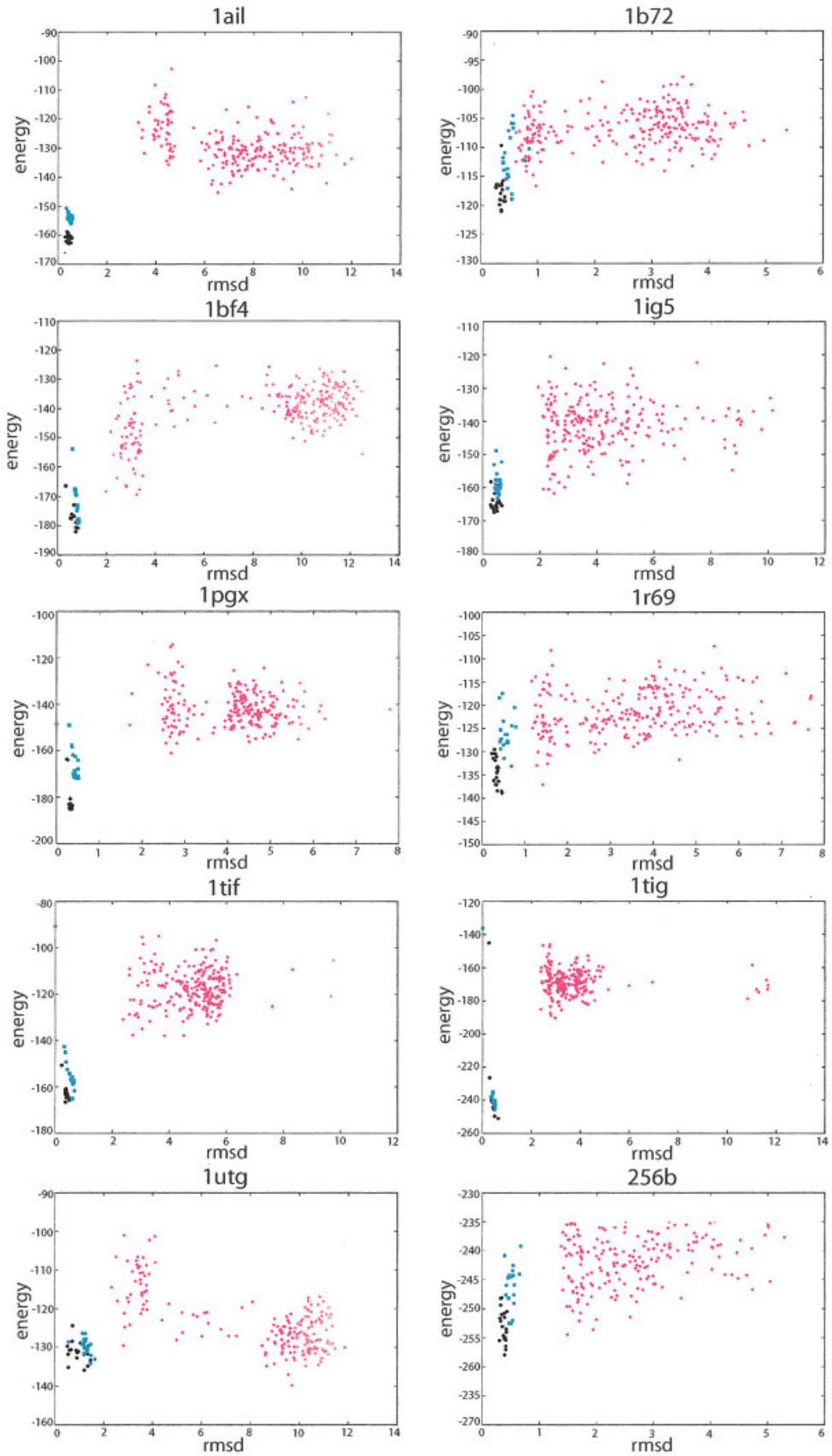
1ail
1b72
1bf4
1ig5
1pgx
1r69
1tif
1tig
1utg
256b
energy
rmsd

Figure 1.

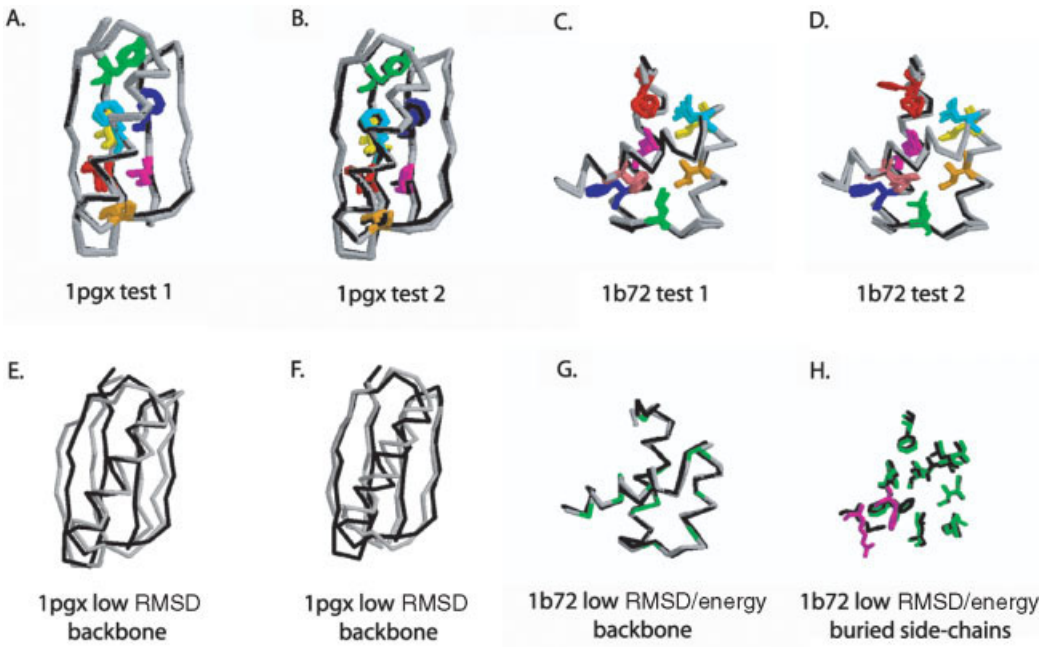

Fig. 2. Cartoon representations of representative 1pgx and 1b72 results from Tests 1, 2 and 4. (a–d) Native structure superimposed with the lowest RMSD* refined idealized native structures from Test 1 (a,c) and Test 2 (b,d), showing selected buried side chains. For clarity, only four models are shown superimposed with the native structure. The native structure is shown in black, and the models are colored. (e,f) Native 1pgx backbone superimposed with (e) the lowest RMSD refined *de novo* model backbone and (f) the lowest RMSD* refined *de novo* model backbone from Test 4. (g) Native 1b72 backbone superimposed with the lowest RMSD/lowest RMSD* refined *de novo* model backbone from Test 4. The lowest RMSD* and lowest RMSD models are identical in this example. (h) Native 1b72 side chains superimposed with the lowest RMSD/lowest RMSD* refined *de novo* model side chains from Test 4. Refined model side chains that adopt the native rotamer are green, and those which adopt the incorrect rotamer are magenta. Orientation is the same as in panel (g), and the colored side chains correspond to colored marked positions on the backbone in panel (g). For panels (e–h), the native structure is shown in black. This figure and Figure 4 were made using Rasmol 2.7 (http://RasMol.org).

from the rotamer library, which indicates that the native side-chain conformations are low in energy compared to alternate possibilities available in the library. Taken together, this indicates that the native state is stable under our energy function given the length and conditions of the refinement simulation.

In all 10 test cases, the energies of the native and native structures with idealized backbones were higher than the refined native structures with idealized backbones, which can be attributed to small numbers of atomic clashes and outlying ϕ/ψ-angles or unusual side-chain rotamers that are typically present in experimentally determined structures. However, the refined idealized native structures were structurally very similar to the experimentally determined structure. Therefore, the energy of the refined idealized native structure provides a useful standard of comparison for the energies of the *de novo* models discussed later in the article.

Fig. 1. Energy versus RMSD for refined models from Tests 1, 2, and 4. Energy is plotted on the *y*-axis, RMSD to native on the *x*-axis. Black points represent refined models from Test 1 for which native side-chain information was included in the starting coordinates, and cyan points represent refined models from Test 2 for which native side-chain information was excluded from the starting coordinates. Magenta points represent the 200 low-energy and the 50 low-RMSD refined *de novo* models from Test 4. This figure and Figures 4 and 5 were made using Gnuplot (http://www.gnuplot.org).

Test 2: Native Structure Refinement Excluding Native Side-Chain Information

In order to accurately refine protein structure models, it must be possible to search and identify the native side-chain conformations. This relatively simple test was designed to assess the effectiveness of the side-chain search procedure coupled with a fixed side-chain rotamer library in the context of an idealized native backbone. This test was identical to Test 1 except that native side-chain χ-angle information was discarded prior to refinement. We carried out 10 independent refinement simulations using the idealized native structures as starting models. The magnitudes of the backbone changes were in general slightly larger those observed in Test 1 ($\approx$0.5 Å RMSD). This is reflected in the energies of the refined idealized native structures; for all of the test proteins, the lowest energy model produced in this test was slightly higher in energy than the lowest energy model produced using native backbone and side-chain coordinates from Test 1 (Fig. 1, cyan points). In most cases, the conformations of the buried side chains were recovered [Fig. 2(b,d)]. The native conformations of other side chains were also recovered, but the displacements from the native positions were larger than seen in Test 1 (data not shown). These results show that side-chain conformations close to the native conformations can be located reasonably well

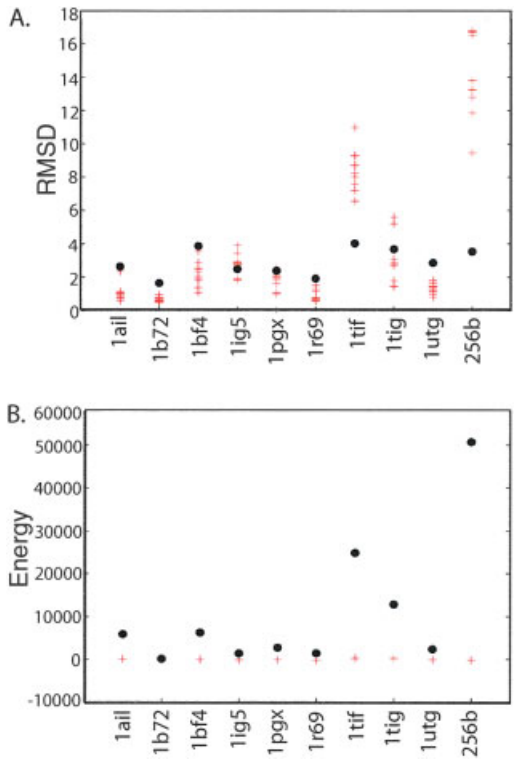

Fig. 3. Results from Test 3. (a) Comparison of RMSD values of starting perturbed native structures and refined perturbed native structures. RMSD to native is plotted on the *y*-axis with the PDB code of the test set proteins on the *x*-axis. Solid circles (●) represent the starting perturbed native structure, and crosses (+) represent the 10 lowest energy refined models from a total of 50 refined models. (b) Comparison of energies of starting perturbed native structures and refined perturbed native structures. Energy is plotted on the *y*-axis and the PDB code of the test set proteins on the *x*-axis. Solid circles represent the starting perturbed native structure and crosses represent the average energy of 50 refined models. (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)

given only an idealized native backbone and that the system is stable using the same criteria defined for Test 1 and given the conditions and length of the simulation. Approximations such as idealized backbone bond lengths and angles as well as the use of a rotamer library[13] to represent protein conformations do not appear to be prohibitive constraints in accurately modeling protein structures.

Test 3: Perturbed Native Structure Refinement

The third test was designed as an easy version of *de novo* model refinement (Test 4). The perturbed native structures have correct topologies and secondary structure assignments, but the relative orientations of the secondary structure elements are altered from the native orientations. To generate the perturbed native structures, the native bond lengths and angles were replaced with ideal values as in Tests 1 and 2, but the resulting coordinate displacement was not minimized through compensating changes in the backbone torsion angles. Native side-chain coordinates were initially included in the perturbed idealized model, as in Test 1. In most cases, the starting backbone conformations deviated significantly from the native conformations [Fig. 3(a)], resulting in significant atomic overlaps, as can be seen from their high energies [Fig. 3(b)]. The backbone deviations are larger than in Test 1 and Test 2, resulting in a larger conformational space that must be searched is in order to find the native state. We carried out 50 independent refinement simulations starting with each perturbed native structure. The RMSD distributions of the starting structure and the 10 lowest energy refined models are shown in Figure 3(a). For most proteins in the test set, the refinement protocol reduced the RMSD between the starting model and the native structure in many of the 50 independent simulations. These included 1ail and 1bf4, in which the RMSD of the perturbed native structures were 2.6 and 3.9 Å, while the RMSD of the lowest 10 energy refined models were 0.6 and 1.1 Å, respectively [Fig. 3(a)]. The perturbed native structures of 1tif and 256b had unusually high energies [Fig. 3(b)], and in these cases refinement was not successful. However, these results indicate that the refinement protocol can bring about significant decreases in RMSD.

Test 4: *De Novo* Model Refinement

The fourth and most challenging test was the refinement of *de novo* models generated by Rosetta. In the previous tests, the backbone torsion angles generally varied by only a few degrees from those of the native structure, but the differences were much larger for *de novo* models even when the topology was correct. Therefore the conformational space that must be searched to find the native state is larger for this test. This is especially noticeable in loops and turns, where the conformations in the models can be significantly different from the native conformation even when the flanking secondary structural elements have roughly correct relative orientations (see discussion below).

The scale of motion produced during the refinement simulation is small (Table I), such that we do not expect non-native starting *de novo* models to sample near-native conformations during the refinement simulation. In an effort to decouple *de novo* model generation from refinement, the proteins in our test set were selected such that we could generate significant numbers of near-native *de novo* models. We generated 10,000 *de novo* models for each protein with a range of RMSD distributions (Table II), and five independent refinement simulations were carried out for each starting *de novo* model. The 250 lowest energy and 50 lowest RMSD refined models were selected and subjected to a second round of refinement (see Methods, Free Energy Function). Using the new refined population, we were able to examine the properties of the near-native models, from which we can determine how successful the protocol is at achieving native-like side-chain packing and backbone conformations. Examination of the non-native models can allow us to identify shortcomings in the potential function.

Properties of Starting *De Novo* and Refined Model Populations

The refinement protocol typically changed the backbone coordinates by 1.5–4.0 Å RMSD from the starting *de novo*

TABLE I. Average Change in RMSD Between Starting *De Novo* and Refined Models

	RMSD of Refined Model								
PDB	<2 Å	2–3 Å	3–4 Å	4–5 Å	5–6 Å	6–7 Å	7–8 Å	8–9 Å	>9 Å
1ail	—	—	2.3	2.5	3.1	3.1	3.1	3.1	3.5
1b72	2.1	2.3	2.5	2.7	2.8	2.5	4.1	3.2	2.6
1bf4	—	2.4	3.0	2.8	3.0	3.4	3.1	2.9	3.0
1ig5	—	2.6	2.8	3.1	3.5	3.9	4.4	4.3	4.7
1pgx	—	1.8	2.2	2.2	2.2	2.4	2.7	2.3	2.4
1r69	2.6	2.5	2.6	3.0	3.3	3.7	3.9	4.0	5.7
1tif	—	1.5	2.2	2.1	2.1	2.3	2.3	2.3	2.6
1tig	—	2.2	2.2	2.6	3.0	3.3	3.4	3.7	3.2
1utg	—	2.3	5.1	4.0	3.8	4.2	4.1	4.2	4.0
256b	2.4	2.5	2.7	2.9	3.2	3.5	4.0	4.5	5.6

TABLE II. RMSD Distribution of Starting *De Novo* Models

	% Models in Test Set with RMSD in Ranges								
PDB	<2 Å	2–3 Å	3–4 Å	4–5 Å	5–6 Å	6–7 Å	7–8 Å	8–9 Å	>9 Å
1ail	0.00	0.00	0.04	0.39	3.74	12.20	20.61	21.10	41.91
1b72	6.82	34.30	39.91	12.92	3.57	1.45	0.33	0.24	0.45
1bf4	0.00	0.10	0.43	0.77	0.67	0.93	2.44	6.10	88.57
1ig5	0.00	2.37	14.89	21.95	17.45	12.93	8.78	8.04	13.60
1pgx	0.01	0.81	4.80	34.30	27.66	11.90	3.00	2.87	14.65
1r69	0.51	8.62	26.91	29.33	19.36	10.01	3.52	1.24	0.50
1tif	0.00	0.03	1.15	21.01	48.00	11.63	9.56	6.58	2.04
1tig	0.00	0.62	9.72	8.72	6.93	5.82	5.15	5.28	57.77
1utg	0.00	0.03	0.28	2.21	5.82	5.36	5.97	9.56	70.78
256b	0.43	5.45	12.15	18.37	19.46	13.74	9.37	6.50	14.51

model coordinates, depending on the protein sequence and the similarity of the models to the native structure (Table I). In general, the magnitude of the change was smaller when the models were <3 Å RMSD from the native structure; the 1r69 and 1ig5 models between 2 Å and 3 Å RMSD moved an average of 2.6 Å and 2.5 Å from their starting *de novo* coordinates, respectively, while models between 6 and 7 Å RMSD from native moved an average of 3.9 Å and 3.7 Å, respectively. When starting from a near-native conformation, the magnitude of the movement during the refinement simulation was of the order required to sample the native state.

Two simple ways to improve the sampling strategy are (1) increasing the number of starting *de novo* models or (2) increasing the number of independent refinement simulations carried out for each template. We generated 500,000–1,000,000 *de novo* models for five of the 10 test set proteins and did not observe improvement in the RMSD distribution of the population over that observed for the initial population of 10,000 *de novo* models used in this study (data not shown). Therefore it seemed unlikely that this approach would improve our sampling strategy. To evaluate the effectiveness of increasing the number of refinement simulations for each starting *de novo* model, we compared the values of the lowest RMSD models from the complete refined population (corresponding to five independent refinement simulations for each starting *de novo* model) to a subset of the refined population (corresponding to a single refinement simulation for each starting *de novo* model). For all proteins in the test set, a lower RMSD model was present in the complete refined population, indicating that increasing the number of refinement simulations carried out for each starting *de novo* model improves the search strategy (Table III).

TABLE III. Comparison of Low-RMSD Models Between Complete Refined Population and Refined Population Subsets

	Low-RMSD Model (Å)			Low-RMSD* Model (Å)		
PDB	Complete	Subset	Δ	Complete	Subset	Δ
1ail	3.3	4.5	−1.1	6.0	6.5	−0.5
1b72	0.6	0.7	−0.1	0.8	0.8	0.0
1bf4	1.9	2.7	−0.8	2.3	2.8	−0.5
1ig5	1.9	2.2	−0.3	2.1	2.2	+0.1
1pgx	1.7	2.7	−1.0	4.2	4.2	0.0
1r69	1.1	1.3	−0.2	2.2	1.4	+0.8
1tif	2.4	2.5	−0.1	4.2	4.2	0.0
1tig	2.4	2.5	−0.1	3.1	2.8	+0.3
1utg	2.3	3.0	−0.7	5.6	9.3	−3.7
256b	1.4	1.5	−0.1	1.5	1.6	−0.1

To evaluate the effectiveness of the refinement protocol at improving the structural quality of the *de novo* models, we compared the RMSD distributions of the starting *de novo* models to an equal sized subset of refined *de novo* models. When binned according to RMSD, for nine of the 10 test set proteins there was a larger percent of the refined population in the lowest RMSD bin relative to the starting *de novo* population (Table IV). To evaluate the effectiveness of the high-resolution full-atom potential

TABLE IV. Enrichments of Initial and Refined Models in Lowest RMSD Bin

PDB	RMSD Range of Lowest Bin[a] (Å)	% Enrichment[a,b] Initial Models	% Enrichment[a,b] Refined Models
1ail	3.0–4.0	3.03	5.28
1b72	0.0–1.0	0.10	8.21
1bf4	1.0–2.0	24.89	8.92
1ig5	2.0–3.0	9.15	12.86
1pgx	1.0–2.0	6.25	9.19
1r69	1.0–2.0	3.54	11.08
1tif	2.0–3.0	5.06	7.30
1tig	2.0–3.0	33.86	32.39
1utg	2.0–3.0	4.24	4.32
256b	1.0–2.0	12.92	36.93

[a]The lowest RMSD bin is the lowest RMSD range in increments of 1 Å for which some models were produced.
[b]The enrichment is defined as [(the number of models in the intersection of the low RMSD population with the low energy population/number of models in the low energy population)/(number of models in the low RMSD population/total number of models)]. The low RMSD population is defined as 5% of the total population with the lowest RMSD values, and low energy population is defined as 5% of the total population with the lowest energies. Enrichment values greater than 1 indicate an enrichment over a uniform distribution. Both the initial models and refined models have associated energies, and we calculated the percent enrichment of both for comparison.

energy function relative to the low-resolution centroid-based energy function, we compared the enrichments of low RMSD models in the subset of low energy models between the starting *de novo* model population and the refined model population. For eight of the 10 test set proteins, the enrichment was larger for the refined models than for the starting *de novo* models (Table IV). Taken together, these results show that generation of *de novo* models followed by the refinement protocol is more successful at producing and recognizing low RMSD models than *de novo* modeling alone.

We also compared the RMSD values after refinement of the lowest RMSD and lowest RMSD* refined models to those of their corresponding *de novo* models in order to identify structural improvement in these select models. The lowest RMSD* model is defined as the lowest RMSD model of the 10 lowest energy models in the complete refined population. We chose to examine the RMSD* model because we considered the energy function successful if one of the 10 lowest energy models were also one of the lowest RMSD models of the 25,000–50,000 refined models. In all cases, the refined lowest RMSD model was closer to the native structure than its starting *de novo* model (Table IV). In nine of the 10 cases, the RMSD improved for the lowest RMSD* refined model from the corresponding starting *de novo* model (Table V), indicating that the energy function is capable of recognizing improvement of the model for some cases. This was not true for 1utg; however, the low energy models are significantly different from the native conformation (Fig. 1) and we do not expect the refinement protocol to sample a near-native conformation and significantly improve the *de novo* model.

TABLE V. Change in RMSD between Starting *De Novo* Models and Refined Lowest RMSD and Lowest RMSD* Models

PDB	Low-RMSD Model (Å) Start	Refined	Δ	Low-RMSD* Model (Å) Start	Refined	Δ
1ail	4.2	3.3	−0.9	7.2	6.3	−0.9
1b72	1.4	0.6	−0.8	2.1	1.0	−1.1
1bf4	4.5	1.9	−2.6	4.5	1.9	−2.6
1ig5	2.8	1.9	−0.9	2.2	2.1	−0.1
1pgx	2.3	1.7	−0.6	3.4	2.9	−0.5
1r69	2.3	1.1	−1.2	3.0	2.0	−1.0
1tif	2.7	2.4	−0.3	4.1	3.9	−0.2
1tig	2.6	2.4	−0.2	2.6	2.4	−0.2
1utg	5.1	2.3	−2.8	6.2	7.4	+1.1
256b	2.5	1.4	−1.1	3.3	1.5	−1.8

Comparison of Limited and Expanded Rotamer Sets

Increasing the number of side-chain conformations made available during the refinement simulation increases the chances that the native conformation will be sampled. However, the computational requirements of using large rotamer sets can be prohibitive. Small rotamer sets are less computationally demanding but are unlikely to contain the diversity of native side-chain conformations observed in the PDB. To determine the effectiveness of different sized rotamer sets, we repacked the side chains on the refined idealized native structures and the combined population of the 50 lowest RMSD models and the 200 lowest energy models using reduced, standard and expanded rotamer sets. The reduced rotamer set was equivalent to that used in Tsai et al.,[9] which allowed three rotamers at buried positions and two at exposed positions. The standard rotamer set allowed up to 45 rotamers and included all rotamers with a frequency of occurrence greater than 0.01.[13] The expanded set was used to supplement the standard set with additional rotamers generated by increasing or decreasing $\chi 1$ and $\chi 2$ by one standard deviation of the variance observed in the PDB.[13]

We evaluated the three rotamer sets by computing the energy gap between the lowest energy repacked idealized native structure and the lowest energy repacked refined model and normalized the energy gap by the standard deviation of the energies of the repacked model population. The most effective rotamer set will maximize the normalized energy gap between the repacked idealized native structure and the models. The results are shown in Table VI; a negative value indicates that the refined idealized native structure is lower in energy than the refined models. The reduced rotamer set was not effective, producing a positive energy gap in all but one case. The standard rotamer set was more effective, producing a negative energy gap in five of the 10 cases, while the expanded rotamer set was successful in all 10 cases. For the cases in which the energy gaps were negative, their magnitudes were found to be larger using the expanded rotamer set. This shows that the additional conformational space

TABLE VI. Normalized Energy Gaps Between Lowest Energy Idealized Refined Model and Lowest Energy Refined Model Overall After Repacking with Different Rotamer Sets

PDB	Reduced	Standard	Expanded
1ail	−0.061	−0.077	−0.176
1b72	+0.432	+0.127	−0.028
1bf4	+0.132	−0.034	−0.031
1ig5	+0.152	+0.737	−0.097
1pgx	+0.662	−0.402	−1.110
1r69	+0.927	+0.109	−0.976
1tif	+0.845	−0.957	−1.666
1tig	+0.064	+0.193	−0.574
1utg	+0.529	−0.012	−0.024
256b	+0.759	+0.507	+0.014

searched using larger rotamer sets is important to successfully discriminate the native conformation from non-native models.

Energetic Properties of Low RMSD and Low RMSD* Refined Models

When the refined low RMSD and low RMSD* models from Test 4 were compared to the refined idealized native structures from Test 1, we observed an energy gap in nine of the 10 test set proteins, with the refined idealized native structures having the lower energies (Fig. 1). Since each refined model represents a local minimum in the energy landscape, the native minimum is lower in energy relative to the other minima identified in the conformational space search. Ideally, the energy of the refined models in the vicinity of the native state would be lower than the energy of those far from the native state, allowing discrimination of native-like models. The energy function is moderately successful at discriminating near-native refined models from non-native ones, as shown by the correlation between energy and RMSD for 1b72, 1bf4, 1ig5, 1r69, 1tig and 256b (Fig. 1). Little correlation was observed for 1tif (a 59 residue α/β-protein) and 1ail (a 70 residue elongated α-helical protein), which may be due to the relatively poor quality of the starting *de novo* model population. The correlation of energy with structural quality of the model as measured by RMSD is most noticeable when the backbone conformation is less than 3 Å RMSD from the native conformation (Fig. 1). This provides an estimate of the accuracy required for low-resolution *de novo* models in order for the high-resolution atomic potential to be effective.

Qualitatively, the refined models do not appear to have as tight side-chain packing as the hydrophobic cores of native proteins (data not shown). The attractive component of the Lennard–Jones potential may provide a measure of the quality of side-chain packing and is also likely to be the most important optimization target for compact near-native models. To test this, we examined the correlation of RMSD with the attractive portion of the Lennard–Jones energies for the refined idealized native structures, the 50 low RMSD refined models and the 200 low energy models for each protein. For all proteins in the test set, the refined idealized native structures had lower Lennard–Jones attractive energies than the refined models. The energy gap in most cases was more pronounced than seen for the complete energy function. In several cases, the refined model backbone conformations approached the accuracy of the refined idealized native structures (256b, 1r69, 1bf4 and 1b72), and for 1b72 the RMSD values of the models overlapped those of the refined idealized native structures. The energy gap was also present in these cases, however, indicating that although the model backbone was native-like the side-chain packing was not optimized. This suggests that it may be useful to increase the weight given to the Lennard–Jones attractive potential during refinement.

Structural and Energetic Differences Between Refined Models and Native Structures

We have shown that the refined idealized native structures are energetically distinct from the refined *de novo* models. For each protein in the test set, we compared the lowest RMSD refined model (the most accurate) and the lowest RMSD* refined model (corresponding to the lowest RMSD model out of the 10 lowest energy models) to the corresponding native structure in order to correlate structural differences with differences in energy. We examined backbone and side-chain conformations to identify structural features of native proteins that are not well captured in the models. For cases where the lowest RMSD* refined model was not among the most accurate models, we attempted to identify deficiencies in the energy function that prevented discrimination of near-native from non-native models.

One particularly successful case was 1b72, where the low RMSD* model was also the most accurate (0.6 Å RMSD), and most of the buried side chains adopted the native conformations [Fig. 2(f)]. While the model was structurally very similar to the native and the refined idealized native structures from Test 1, the energy of the model was higher. On closer examination of the buried side chains, we found that F44 adopts incorrect $\chi1$ and $\chi2$ torsion angles. This conformation is not compatible with the native conformation of the exposed R48, which packs against the protein surface and makes contacts with L21 in the native structure. The backbone conformations of the refined idealized native and the model are very similar; therefore incorrect side-chain conformations are likely the major source of the energy difference between the model and the refined idealized native structure.

For 1pgx, several near-native models had low energies; however the lowest RMSD refined models were not among the lowest energy models. The structural differences between the lowest RMSD model and the refined idealized native structure are largely due to incorrect side-chain conformations and small offsets in the backbone. The native rotamer for I7 is rare ($\chi1 = -47°$, $\chi2 = 95°$) and therefore infrequently incorporated into models. A more common rotamer was chosen at this position in the refined model [$\chi1 = -60°$, $\chi2 = 171°$, Fig. 4(a)]. I7 makes extensive interactions with Y3, L5 and F52 in the hydrophobic core,

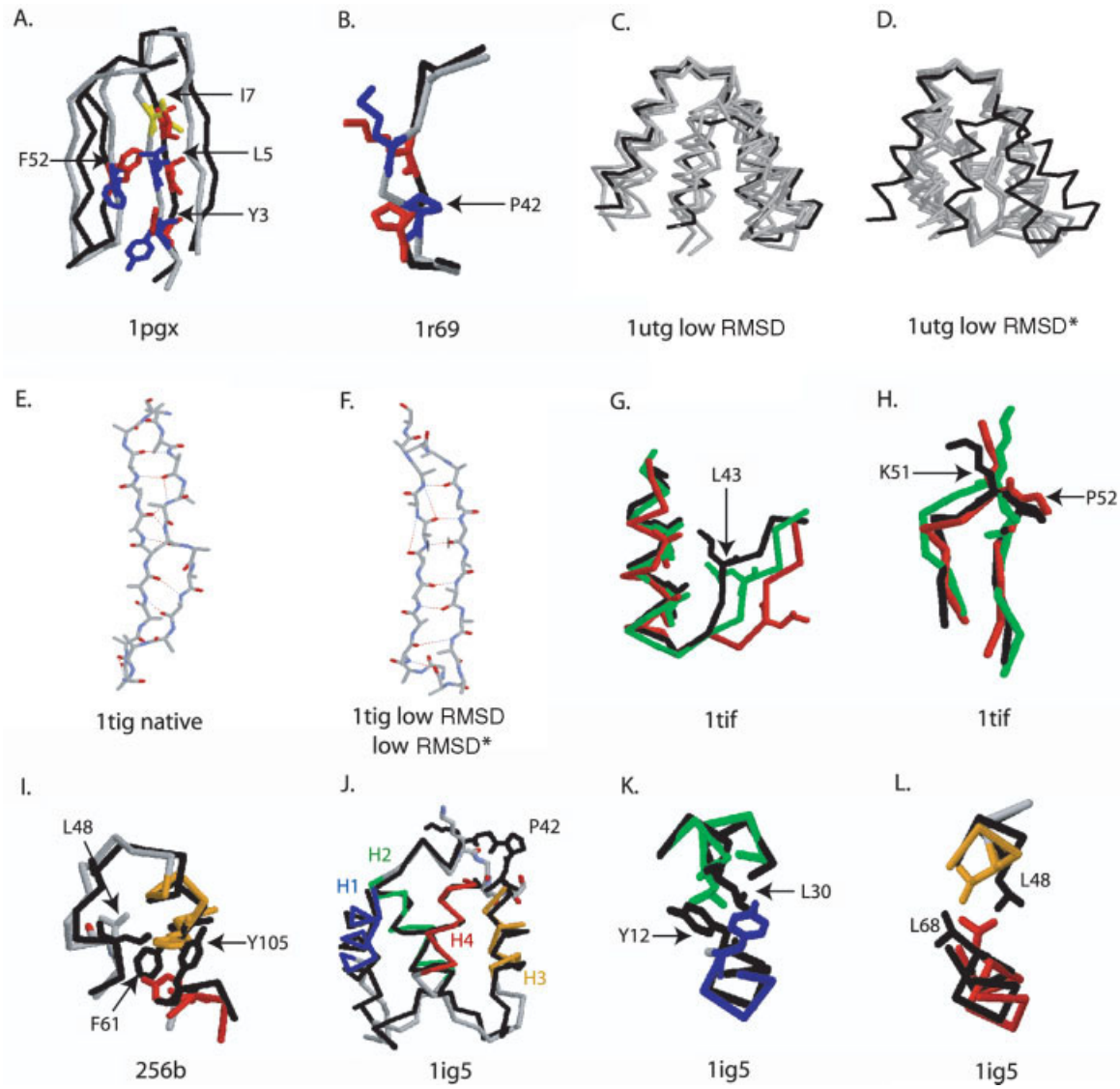

Fig. 4. Comparison of selected structural features from models produced in Test 4. In all panels, the native structure is shown in black. (a) 1pgx low RMSD refined model. Native side chains are shown in red, modeled side chains in blue and the rare-rotamer I7 side chain is shown in yellow. (b) 1r69 H3/H4 loop. Native side chains are shown in red, and selected incorrectly modeled side chains in blue. (c) A cluster of structurally similar refined low RMSD 1utg models, showing an expanded conformation. The backbones of the refined models are shown in gray for this panel and panel (d). (d) A cluster of refined low-energy 1utg models, showing a compact conformation. (e) Native 1tig backbone showing irregular strand pairing between strands 3 and 4. Hydrogen bonds in this panel and panel F are drawn in dashed lines. (f) Refined low RMSD* *de novo* model showing regularized strand pairing between strands 3 and 4. (g) 1tif helix/strand packing and conformation of the core residue L43. The low RMSD* refined model is shown in red, with the low RMSD refined model in green. (h) 1tif loop containing P52 and K51. Coloring is as in panel (e). (i) 256b, cluster of incorrectly modeled side chains (shown in red, orange and gray). (j) Low-RMSD refined model showing accurate arrangement of secondary structure elements. The helices H1, H2, H3 and H4 are shown in blue, green, orange and red, respectively. (k,l) Clusters of incorrectly predicted side-chain conformations in the low-RMSD 1ig5 model.

and the rotamers for these side chains are also incorrectly assigned in the model due to steric clashes with I7. The buried side chains F30, Y45 and W43 are assigned correct rotamers, but the backbone is shifted and slightly expanded in these regions in order to accommodate the incorrect Y3, L5 and F52 rotamers. The most notable difference between the low RMSD* refined model and the refined idealized native structure is an offset in the register in β-strands 1 and 2 that arises from a two residue displacement of the connecting β-turn. The strand register was correctly predicted in the low RMSD model.

In the case of 1r69, several accurate models were produced but did not receive low energies (Fig. 1). In the native structure, the loop connecting the third and fourth helices (H3/H4 loop) contains a proline residue at position 42 (P42), which adopts a β-conformation, but it adopts an α-helical conformation in the lowest RMSD model [Fig. 4(b)]. The incorrect proline conformation is not compatible with the native conformation of the H3/H4 loop, and furthermore it disrupts hydrophobic core packing in the low RMSD model. Fragment insertion moves were attempted in the refinement protocol and were found to

correct the proline error, but such large backbone adjustments were accepted infrequently. Analysis of the backbone ϕ and ψ angles at this position in the starting *de novo* model population compared to the refined model population confirmed this; only 0.05% of the proline residues at this position in the starting *de novo* models changed conformation during refinement (data not shown). The lowest RMSD* model (2.0 Å RMSD) contained backbone angles that deviated from the native conformation in the H3/H4 loop including P42 and additionally contained errors in the H1/H2 loop.

For 256b, the low RMSD and low score predictions were 1.3 Å and 1.4 Å RMSD from the native structure, respectively. The side-chain accuracy was high for both models, with only one cluster of incorrect side chains (Y105, F61 and L48) seen in both models that caused shifts in the backbone conformation from the native structure [Fig. 4(i)]. For 1bf4, the lowest RMSD* model was also the most accurate (2.0 Å RMSD). The overall backbone errors were small, but the W23 and R24 rotamers were incorrect in the model and resulted in a small displacement of the C-terminal helix.

For 1tig, the lowest RMSD* model was also the lowest RMSD model (2.4 Å RMSD). The structure consists of two parallel helices packed against a four-stranded sheet of topology 1243. The regions containing the largest coordinate error in the model are the outside two strands (S1 and S3), the loop connecting S3 and S4 and the long loop connecting S2 to H2. In the native structure, S4 contains a central proline residue (P72), which in addition to E69 form bulges in the hydrogen bonding network typically seen in regular β-sheet structures [Fig. 4(e)]. The potential energy function attempts to regularize β-sheets through enforcing hydrogen bonding networks, and as a result the outside strand in the refined low RMSD/low RMSD* model is more regular than the native strands (Fig. 4f).

For 1tif, the starting *de novo* model population was structurally more distant from the native conformation than for other proteins in the test set (Table II). The low RMSD and low RMSD* models are 2.4 and 3.9 Å RMSD from the native structure, respectively. The high energy of the low RMSD model is due primarily to the Lennard–Jones repulsive energy; L43 faces the interior of the protein as in the native structure but does not adopt the correct side-chain conformation and therefore clashes with other side chains in the hydrophobic core [Fig. 4(g)]. In the low RMSD* model this side chain is oriented towards the solvent and does not make close atomic contacts [Fig. 4(g)]. Atomic overlaps between K51 and P52 also contribute to the high Lennard–Jones repulsive energy. Although the overall RMSD is higher for the low RMSD* model, the loop containing K51 and P52 in the low RMSD* model is in a more native-like conformation [Fig. 4(h)]. The 1tif native structure contains a twisted and irregular four-stranded β-sheet, and, similar to the 1tig example, the low RMSD* model contains regularized strand pairings, particularly in strand 4.

The low RMSD and low RMSD* 1ig5 models are 1.9 Å and 2.1 Å from the native structure, respectively. The regular helices of the low RMSD model are closely superimposable with those of the native structure [Fig. 4(j)]; however the C-terminus of H1 is one turn shorter in the model relative to the native structure. The secondary structure prediction for the C-terminal region of H1 is not confident, resulting in a diverse population of fragments that were inserted at that position in the *de novo* models and creating a loop connecting H1 and H2 that is too long in the low RMSD model. Other deviations in the low RMSD model from the native structure include minor discrepancies in the backbone torsion angles in the loop connecting H2 and H3, which are localized to G41 and P42 [Fig. 4(j)]. The side-chain predictions are overall are accurate, but an incorrect rotamer for Y12 is accompanied by an incorrect rotamer for L30 [Fig. 4(k)], and an incorrect rotamer for L48 is accompanied by an incorrect rotamer at L68 [Fig. 4(l)].

Many of the 1ail *de novo* models contained secondary structure errors. Transitions between secondary structure assignments require large angle changes, which are rarely accepted during the refinement protocol. The secondary structure prediction methods which bias fragment selection in the Rosetta *de novo* simulations predict residues 44–54 to be a long loop, which is consequently seen in the majority of models produced for this protein. However, residues 44–49 are helical in the native structure. The native structure consists of three elongated helices with the C-terminal helix at a 90° angle to the N-terminal helices. The energy function favors compact globular structures and thus creates breaks in long helices. Partially as a result of these mistakes, the proportion of near-native starting *de novo* models in the population was low, and the energy function failed to discriminate near-native from non-native models.

For 1utg, the low RMSD* model is more compact and buries more hydrophobic residues than the low RMSD model or the native structure. Certain properties of the native structure are not currently modeled by the energy function; it forms a homo-dimer in solution, and many residues that are exposed in the monomer are buried in the dimer interface. The dimer is a globular protein with a hydrophobic core, but the monomer is elongated. Only approximately 10% of the side chains in the native monomer are buried, which is less than half the amount seen in typical globular proteins in our benchmark set. The complete refined population shows two distinct populations (data not shown), one low in energy but non-native and the other adopting the native topology but higher in energy [Fig 4(c,d)]. The energy function rewards models with typical protein-like features such as hydrophobic burial and compactness; therefore it is not surprising that the non-native models received lower energies than the near-native ones. Despite the less compact topology, the attractive component of the Lennard–Jones energy is more negative for the native and refined idealized native structures than

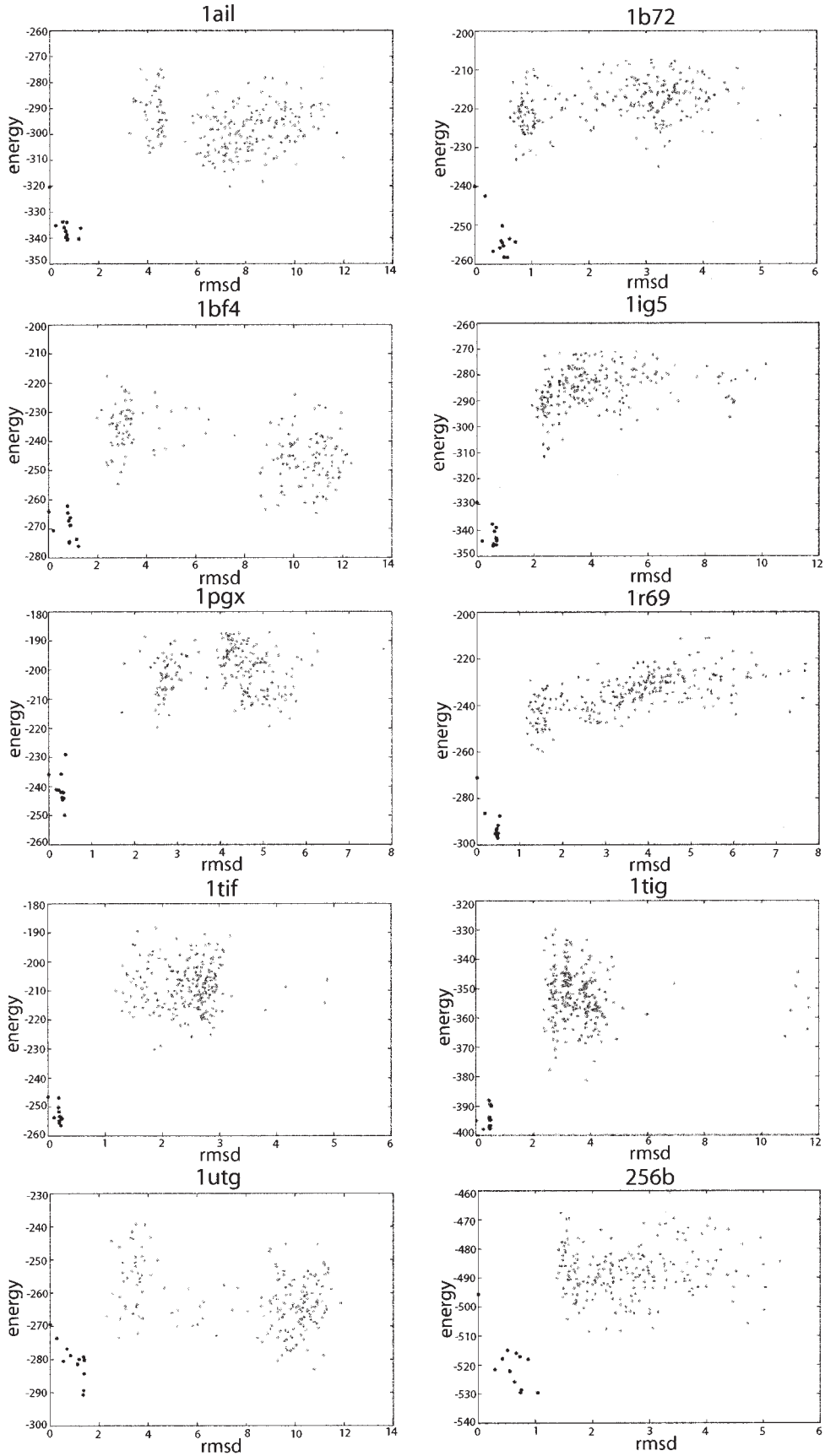

Fig. 5. Attractive component of the Lennard–Jones energy versus RMSD. Lennard–Jones energies of refined models from Test 1 (black points) and Test 4 (gray points) are plotted on the *y*-axis, while RMSD is plotted on the *x*-axis.

for the non-native low RMSD* models, similar to other proteins in the test set (Fig. 5).

Discussion

The four tests we have designed provide insight into both the energy function and sampling problems, and we discuss these two issues as well as the strengths and limitations of our method in the following sections.

Energy Function

The Rosetta energy function involves considerable approximations, such as modeling solvent implicitly rather than explicitly, and neglects long-range electrostatics. However, these approximations do not appear to be a primary limitation to successful model refinement. This is illustrated by comparing the energies of the models generated in Tests 1, 2 and 4, in which the refined idealized native models from Tests 1 and 2 are lower in energy and RMSD than the refined *de novo* models from Test 4 (Fig. 1). In addition, for the cases in which the refined *de novo* models from Test 4 are closer than 3 Å RMSD to the native structure, their energies are lower than those of the non-native refined *de novo* models. Test 3 shows that the refinement protocol can recover the native state for the perturbed native structures (Fig. 3), and the results from Test 4 show that the low RMSD and low RMSD* refined *de novo* models are generally shifted closer to the native conformation and energy from their starting unrefined models. Poor results might be expected for proteins that dimerize or bind ligands, but the results are surprisingly good for the calcium binding protein 1ig5, the DNA binding proteins 1bf4 and 1b72, the heme binding protein 256b and the dimeric protein 1utg, which was modeled as a monomer in our tests and has an extensive exposed hydrophobic surface in the monomeric state. In many cases the one of the 10 lowest RMSD* models is also one of the most accurate in the population. These results are an improvement over previous model discrimination tests carried out using Rosetta.[9,15]

The free energy function fails in some cases, assigning higher energies to the near-native models than to the non-native models. These cases primarily involve native structures that are not globular, such as 1utg, or that have differences in helix and sheet content between the native structure and the unrefined *de novo* models, as in 1ail. Frequently, low energy models with an excess of helical secondary structure relative to the corresponding native structures are favored by the free energy function. There may be several explanations for this phenomenon, one of which is that initial configurations of helices are well modeled by fragments inserted during the *de novo* protocol. Well-formed helices are compact, contributing more favorable Lennard–Jones attractive energies, and have hydrogen bonding networks, contributing favorably to the overall hydrogen bond energy. In order for β-strands to achieve the same energetically favorable structures as helices, they must pair with adjacent strands. This process involves searching a considerably larger conformational space than for helices; therefore there is a lower probability of forming energetically favorable structures containing β-strands. We have attempted to compensate for this by modifying the weights on the short-range hydrogen bonds relative to other hydrogen bonds, but capturing the subtle differences in backbone torsional free energies is clearly a challenging problem.

Sampling

While the energy function shows adequate performance in each of the four tests, the sampling protocol appears to be a considerable limitation. The rotamer search protocol and the backbone conformational search algorithm are reasonably successful in the relatively simple tests (Tests 1, 2 and 3); however, refinement of *de novo* models (Test 4) is less successful. To identify the major deficiencies in our search algorithm, we analyzed the non-native structural features in the refined low RMSD and low RMSD* models generated in Test 4. We also examined the energies of the models in order to correlate the non-native structural features with the energy gap between the models and the refined idealized structures. Consistent with the generally poorer packing in the refined model structures compared to the native structures, the refined *de novo* models have consistently worse Lennard–Jones attractive energies (Fig. 5) than the native and refined idealized native structures. This indicates that the refinement protocol does not effectively minimize the Lennard–Jones attractive component of the energy function, and this is a target for future research.

The structural differences between the refined models and the refined idealized native structures consist of errors in local backbone conformation, errors in side-chain conformation and errors in strand alignment. The first two problems are most prominent when the native structure contains an unusual feature, such as a strained side chain or local backbone conformation. The problems associated with irregular backbone features in the native structure are illustrated by the 1tif and 1tig examples. The structures contain twisted and irregular four-stranded β-sheets with unusual hydrogen bonding networks. The β-sheets formed in the low RMSD* and low RMSD models for these proteins resemble canonical β-strand pairings and contain more favorable backbone torsion angles. An example of sampling problems associated with rare rotamers in the native structure can be seen in the 1pgx low RMSD model from Test 4. A common rotamer was chosen at position I7 rather than the unusual rotamer seen in the native structure, thus attributing a high energy to the low RMSD model. In general, the most frequently sampled conformations in the simulations will reflect the most common features of native proteins, while the unique and unusual features, perhaps relating to functional constraints, will be sampled rarely.

However, even in the absence of unusual features, the size of the conformational space is problematic. The 1pgx example described above also illustrates the challenging combinatorial problem associated with rotamer selection. The incorrect rotamer choice at position I7

causes a cascade of incorrect rotamers to be chosen for nearby side-chains due to steric clashes with I7 [Fig. 4(a)]. As evident from the energy gap between the refined idealized native structures and the refined *de novo* models, it appears that for the energy to decrease substantially almost all core side-chains must adopt roughly the native conformations. Many of the refined *de novo* models generated in Test 4 have correct topologies but considerably higher free energies than the refined idealized native structures; this is perhaps not surprising given the importance of the precise jigsaw puzzle-like packing of side-chains in protein hydrophobic cores. Even subtle differences in the backbone conformation are sufficient to prevent native-like side-chain packing, which is illustrated by the 1b72 example. A subset of the 1b72 refined *de novo* models are in the same RMSD range as the refined idealized natives but have incorrect side-chain conformations and higher energies. It is possible that each correct backbone feature was sampled individually but unlikely that the correct conformations were sampled simultaneously.

Local backbone conformations containing glycine and proline present challenges to the refinement protocol. Both residues are frequently found in loops and turns, which adopt a wider range of conformations in the PDB than regular secondary structure elements. In at least two cases, 1r69 and 1ig5 [Fig. 4(b,i)], the most notable mistake in the low RMSD model is an incorrect loop conformation due to non-native proline torsion angles. Significant changes to the backbone torsion angles, such as those accompanying changes from alpha to beta conformations, are difficult to achieve for other residues as well, as illustrated by the 1ail helix example. Clashes between the side chain and backbone atoms limit the range of motion that can be achieved through small angle adjustments. Fragment insertion moves are included in the refinement protocol and can cause larger backbone changes by replacing the torsion angles of several residues simultaneously; however, these changes are not frequently accepted.[11] The examples discussed above highlight the difficulties of sampling a large volume of conformational space in the context of a compact polypeptide chain.

CONCLUSIONS

We have shown that the extent of sampling can be improved somewhat by increases in the number of independent refinement simulations starting from each unrefined *de novo* model. There are a number of avenues for improvement of the sampling protocol itself. First, the constant-temperature Monte Carlo minimization procedure may profit from a replica exchange approach in which different trajectories are carried out at different temperatures, and at intervals exchanges in energy are allowed. It should be noted, however, that Monte Carlo minimization is in a sense a zero-temperature approach, as minimization is carried out at each step, so the thermodynamic analogy to a multi-canonical ensemble no longer holds. Second, the side-chain packing algorithm uses the same potential function with a steep Lennard–Jones repulsive component as the overall refinement procedure; hence rotamers may be rejected that could with small adjustments lead to significant energy decreases. As the steep repulsive Lennard–Jones potential was one of the primary sources of improved discrimination of refined native structures from models compared to our previous results,[9] a softer potential in the overall protocol is not desirable. However, it may be useful to use a softer potential in the repacking step. Alternatively, it may be useful to increase the strength of the attractive Lennard–Jones potential during the repacking step to favor the formation of tightly packed cores, which are absent in most of the refined model structures generated in this study. In addition, preliminary results suggest that briefly minimizing the energy of each rotamer in the context of other fixed side-chain conformations during the packing procedure increases the probability of selecting the correct rotamer.[19] Finally, a direct enumeration approach in which the distribution of local structural features in the *de novo* model population are systematically sampled may prove useful in tackling the combinatorics of searching backbone and side-chain conformations.

ACKNOWLEDGMENTS

The authors thank Bill Schief and Phil Bradley for comments on the manuscript, Carol Rohl for managing the Rosetta code base, and members of the Baker laboratory for helpful advice and discussion. We also acknowledge Keith Laidig for expert management of computational resources. K.M.S.M. gratefully acknowledges funding from the Helen Hay Whitney Foundation. D.B. was supported by the Howard Hughes Medical Institute.

REFERENCES

1. Bradley P, Chivian D, Meiler J, Misura KMS, Rohl CA, Schief WR, Wedemeyer WJ, Scheuler-Furman O, Murphy P, Schonbrun J, Strauss CEM, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Prot: Struct Funct Genet 2003;53:457–468.
2. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D. TOUCHSTONE: a unified approach to protein structure prediction. Prot: Struct Funct Genet 2003;53(S6):469–479.
3. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. Prot: Struct Funct Genet 2003;53(S6):480–485.
4. Fang Q, Shortle D. Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments. Prot: Struct Funct Genet 2003;53(S6):486–490.
5. Lee MR, Tsai J, Baker D, Kollman PA. Molecular dynamics in the endgame of protein structure prediction. J Mol Biol 2001;313:417–430.
6. Lee MR, Baker D, Kollman PA. 2.1 and 1.8 Å average C_α RMSD structure predictions on two small proteins, HP-36 and S15. J Am Chem Soc 2001;123:1040–1046.
7. Lu H., Skolnick J. Application of statistical potentials to protein structure refinement from low resolution *ab initio* models. Biopolym 2003;70(4):575–584.
8. Fan H, Mark AE. Refinement of homology-based protein structures by molecular dynamics simulation techniques. Prot Sci 2004;13(1):211–220.
9. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl C, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Prot: Struct Funct Genet 2003;53(1):76–87.
10. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein

tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268(1):209–225.
11. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. Meth Enzym 2004;383:66–93.
12. Press WH, Flannery BP, Teukolsky SA, Vettering WT. Minimization or maximization of functions. In: Numerical recipes in FORTRAN: The art of scientific computing (2nd ed). Cambridge: Cambridge Univ Press; 1992. p 428-429.
13. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
14. Wedemeyer WJ, Baker D. Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. Prot: Struct Funct Genet 2003;53(2):262–272.
15. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97(19).
16. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Prot: Struct Funct Genet 1999;34:82–95.
17. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. J Mol Biol 2003;326:1239–1259.
18. Kuhlman B, Dantas D, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic level accuracy. Science 2003;302(5649):1364–1368.
19. Wang C. Unpublished data.