Hui Lu*
Jeffrey Skolnick†
Laboratory of Computational Genomics, Donald Danforth Plant Science Center, 975 N Warson St., St. Louis, MO 63132

Received 5 June 2002; accepted 12 August 2003

Application of Statistical Potentials to Protein Structure Refinement from Low Resolution *Ab Initio* Models

Abstract: *Recently* ab initio *protein structure prediction methods have advanced sufficiently so that they often assemble the correct low resolution structure of the protein. To enhance the speed of conformational search, many* ab initio *prediction programs adopt a reduced protein representation. However, for drug design purposes, better quality structures are probably needed. To achieve this refinement, it is natural to use a more detailed heavy atom representation. Here, as opposed to costly implicit or explicit solvent molecular dynamics simulations, knowledge-based heavy atom pair potentials were employed. By way of illustration, we tried to improve the quality of the predicted structures obtained from the* ab initio *prediction program TOUCHSTONE by three methods: local constraint refinement, reduced predicted tertiary contact refinement, and statistical pair potential guided molecular dynamics. Sixty-seven predicted structures from 30 small proteins (less than 150 residues in length) representing different structural classes (α, β, α/β) were examined. In 33 cases, the root mean square deviation (RMSD) from native structures improved by more than 0.3 Å; in 19 cases, the improvement was more than 0.5 Å, and sometimes as large as 1 Å. In only seven (four) cases did the refinement procedure increase the RMSD by more than 0.3 (0.5) Å. For the remaining structures, the refinement procedures changed the structures by less than 0.3 Å. While modest, the performance of the current refinement methods is better than the published refinement results obtained using standard molecular dynamics.*
 Biopolymers 70: 575–584, 2003

Keywords: *structure refinement; statistical potential; structure prediction; molecular dynamics; Monte Carlo*

INTRODUCTION

Recent progress in *ab initio* structure prediction suggests that *ab initio* protein structure prediction, at least for small proteins, is becoming practical[1–5]. By way of illustration, the TOUCHSTONE program can predict a correct fold for 75% of 65 small test proteins.[5] The program has been further used in genome scale structure prediction of *mycoplasma genitalium*.[6] Some of the predicted structures have already been

Correspondence to: Jeffrey Skolnick; email: skolnick@buffalo.edu
**Current address*: Department of Bioengineering, University of Illinois at Chicago, Chicago, IL
†*Current address*: Center for Excellence in Bioinformatics, University of Buffalo, Buffalo, NY
Contract grant sponsor: National Institutes of Health; Contract grant number: RR12255
Biopolymers, Vol. 70, 575–584 (2003)

used in protein function prediction to identify biochemical functions and for ligand docking to find the active site of a known ligand.[7,8] But for drug ligand screening, these low-resolution predicted structures are not good enough if contemporary technology is used. Thus, structure refinement from low to high resolution is of great importance.

To make the conformational search practical in *ab initio* structure prediction, a reduced representation is often used. In TOUCHSTONE, for example, a side chain center of mass model is adopted.[9] Using this simplified representation in many cases gives us the correct topology, that is, the root mean square deviation (RMSD) of the predicted structure is less than 6.5 Å from native structure. This simplification of the model nevertheless limits the accuracy of the prediction to about 3 Å. Furthermore, the residue-based potentials used in reduced models are not sensitive to small structural changes. A more detailed representation of a protein is one natural way of attempting structure refinement. In this spirit, a full heavy atom representation of the protein is used in this work.

Structure refinement is a nontrivial problem. Despite great effort, so far no consistent method to improve protein structure has been reported. For example, in homology modeling, the best prediction resulted from using the template.[1] Attempts to further refine the structure usually make the predicted structure worse, that is, the structure will have a higher RMSD. In *ab initio* protein prediction, currently the best strategy is to use a cluster based method to group the generated conformations, then use the centroids of these clusters as the structure prediction.[10,11] A lot of effort has been expended in the use of molecular dynamics (MD) with either explicit or implicit solvent for the structure refinement, once a low-resolution structure with correct topology is identified. So far, improvements were seen for only a few limited cases.[12–14] Other refinement attempts have failed.[15] Furthermore, these MD simulation methods require intensive computations and their application to a genome would be impractical. Also, there is no guarantee that lengthier MD simulations would result in better structure. For example, a microsecond MD folding simulation of a 36-residue protein didn't produce the native structure.[16]

In the current work, we shall attempt to refine a set of predicted structures generated from the TOUCHSTONE *ab initio* prediction program. The TOUCHSTONE program can be summarized as follows. It has three components: a potential, a search engine, and a structure selection protocol. The potential used in TOUCHSTONE is based on a side chain center of mass model. Besides the generic (i.e., proteinlike) solvent accessible term and pairwise terms, the potential includes a novel term based on threading-based contact predictions. A contact between two residues is predicted when it appears in more than 25% of the top 20 scored template structures when the threading program PROSPECTOR[17] is applied. These contact predictions, although not always correct, greatly enhance the probability of finding a correct topology for the target protein. The search engine in TOUCHSTONE is based on a replica exchange Monte Carlo sampling method.[5,18] We have used two methods for structure selection. One is based on a clustering method[11]; the other introduces a combination of an atomic pairwise statistical potential with clustering[5] that yields slightly better results.

The atomic potential used in TOUCHSTONE is a distance dependent, heavy atom statistical pair potential.[19] This quasi-chemical atomic potential was developed using the similar strategy as those residue-based statistical potentials.[20,21] The performance of this atomic potential in the selection phase is better than when a residue-based potential is used. In a few cases, the atomic potential can select near-native structures even when the clustering method fails. Furthermore, we noticed that this potential is not very sensitive to minute structural details. Thus, intensive minimization of the rebuilt atomic structure is unnecessary. In some decoy test sets, we have shown that the current statistical potential performs similar to and sometimes a little bit better[19] than a molecular mechanics potential.[22–24] In this work, we take the next logical step and develop structure refinement procedures that are guided by these pairwise atomic statistical potentials.

This article is organized as follows: first, we will present the three methods we used for atomic statistical potential based structure refinement. Then, in the Results section, the performance of those methods is evaluated. In some cases, we directly compare our refinement results with MD simulations done by us and by other groups.[15] Finally, in the Discussion section, we will analyze possible reasons for the moderate success of current methods and suggest further improvements based on the use of statistical potentials.

METHODS

Test Cases

Thirty proteins less than 150 residues in length are selected as test cases; these are listed in Tables I, II, III. All structure categories of protein are included: all-α, all-β, and mixed α/β proteins. As a starting point, various qualities of struc-

Table I Comparison of Refinement Results from Molecular Dynamics Simulation and from Local Constraint Refinement Procedure

PDB Code	Class	Initial RMSD (Å)[a]	MD/GB Refinement (Å)[a,b]	Statistical Potential Refinement (Å)[c]
1cis	α	4.30	*4.75*	**3.75**
		5.07	*5.71*	**4.56**
		5.39	5.31	5.54
		6.32	6.49	**5.89**
		7.97	7.89	*8.35*
1tit	β	2.49	*3.57*	**2.23**
		2.98	3.24	**2.61**
		3.61	*4.42*	3.87
		4.12	*6.39*	**3.57**
		9.61	9.92	**9.28**
Average		5.19	5.77	4.97

[a] Starting structures are obtained with TOUCHSTONE program.

[b] MD/GB is molecular dynamics simulation with Generalized Born implicit solvent model. From each starting structure one 100 ps MD simulation is performed and the RMSDs of the final structures are reported.

[c] Statistical potential refinement runs are performed using LCR protocol. Monte Carlo simulations are performed, and the RMSD of the lowest atomic potential structure is reported. The cases where the RMSD improves by more than 0.3 Å are shown in bold; those where the RMSD deteriorates by more than 0.3 Å are shown in italic.

tures, mostly with an RMSD to native structures ranging from 2.5 to 6.5 Å, are selected. These structures are the top selections from the atomic potential based selection procedure described in TOUCHSTONE.[5] In total, 67 initial structures are used. Because our goal is to test the performance of the refinement procedure on near-native structures, we will mostly try our program on those cases for which the TOUCHSTONE-predicted structures have the correct native topology. These initial structures are posted on http://proteomics.bioengr.uic.edu/data/refine.data.tar.

Local Constraint Refinement (LCR)

The first method is based on the observation that, for the initial structures that have the correct topology (<6.5 Å RMSD), locally the structures are quite close to native. Here, local structure means the structure of a continuous piece of protein sequence that is contiguous and several residues in length. For refinement, we want to restrain the local pieces of the protein from straying too far from the initial conformations because these initial local conformations are already quite close to native. Yet, we want to give the structure enough freedom to locally relax. Thus, local structures are not fixed but biased toward the initial conformations. The procedure works as follows: an additional potential term is added to the Monte Carlo sampling program to constrain the local pieces of the initial structure and the side-chain-only (SICHO) reduced model[6] is used. In practice, a penalty in the potential is applied when a local structure is 2 Å from the initial structure. Constraints on three, five, seven, and nine residues have been tested, and we found that local pieces seven residues in length gave optimal results. This penalty energy term can be written as $\Sigma_{K=1}^{N-m}(\mathrm{RMSD}_K-2)$, where N is the length of the protein, m is the number of residues constrained, RMSD_K is the RMSD between initial predicted structure with current structure for residues between K and $K+m$. Here, we have used parallel hyperbolic Monte Carlo sampling for structure generation.[18] After the Monte Carlo, all structures are rebuilt with atomic detail and evaluated with the atomic pairwise statistical potentials. The final refinement results are the structures with the lowest atomic potential.

Reduced Contact Refinement (RCR)

The second method uses a reduced number of constraints derived from our threading-based contact prediction and used to assemble the global topology. Among those predicted contacts described in a previous article,[6] some are incorrect. In the refinement process, we first examine the predicted contacts in the TOUCHSTONE predicted structures. Distances of each pair of predicted contacting residues are calculated. Those predicted contacts that are obviously inconsistent with the initial predicted structure, that is, the distance between Cα atoms is larger than 12 Å, are removed. The refinement procedure consists of rerunning the Monte Carlo sampling program with the modified potential using the SICHO reduced model. Structures generated in the program are evaluated with atomic pairwise statistical potential. The lowest energy structure is reported as the refinement result.

Statistical Potential Guided MD

Here, we use our atomic knowledge-based pair potential as a guide to a MD simulation with an implicit solvent model. The procedure works as follows: from the initial predicted structure, 20 short molecular dynamics simulations at 300 K are started with different initial conditions (randomly assigned velocities), each 10 picoseconds long. The MD simulations are performed using the CHARMM program[25] with the CHARMM27 potential plus Generalized Born (GB) parameters.[26] During this process, the predicted secondary structure is restrained. At the beginning of each of the 20 simulations, one piece of secondary structure is randomly chosen and moved by a short distance or rotated by a few degrees. At the end of these simulations, the 20 output structures are evaluated with the atomic statistical pair potential, and the structure with the lowest energy is then used as the starting structure for next round of simulations. Usually after about 10 rounds, the value of atomic potential on the final selected structure doesn't decrease in subsequent

Table II Results from Local Constraint Refinement

Name	Length	Class	Initial RMSD (Å)[a]	Best RMSD (Å)[b]	Final RMSD (Å)[c]
1ixa	39	Small	4.5	3.5	**4.1**
1rpo	61	Small	3.8	2.7	**3.2**
1c5a	66	α	5.7	4.6	5.5
1pou	71	α	3.2	2.8	3.0
1pou	71	α	4.0	3.2	**3.5**
1kjs	74	α	4.8	4.6	5.2
1aoy	78	α	4.5	3.4	**3.8**
2af8	86	α	4.4	3.8	4.5
2lfb	100	α	5.8	5.4	6.0
256bA	106	α	3.4	3.1	3.3
1hlb	138	α	3.2	2.6	**2.7**
1pgx	56	α/β	2.4	2.2	2.5
2ptl	60	α/β	2.6	2.2	**2.2**
2fmr	65	α/β	4.0	3.4	3.8
1ife	91	α/β	10.3	9.6	10.3
1shaA	103	α/β	4.1	3.4	**3.7**
1cewl	108	α/β	5.7	5.5	5.9
1gpt	47	β	3.5	2.7	*4.1*
1shg	57	β	4.9	4.3	**4.6**
1vif	60	β	3.7	2.8	**3.3**
1csp	64	β	5.0	3.4	**4.4**
1sro	66	β	4.6	4.0	**4.3**
1sro	66	β	6.6	4.4	**4.8**
1ah9	71	β	5.4	4.8	**4.9**
1tit	89	β	2.6	1.9	2.4
1ksr	100	β	5.6	4.4	**5.1**
ave			4.6	3.8	4.3

[a] Initial RMSD is the RMSD of the starting structure that was generated with TOUCHSTONE program.

[b] Best RMSD is the best structure generated during the sampling in the refinement. Of course, without the knowledge of native structures, we cannot always pick out the best structures.

[c] Final RMSD is the RMSD of the output of the refinement procedure. These are the lowest energy structures selected by the pairwise atomic statistical potential. The cases where the RMSD improves by more than 0.3 Å are shown in bold; those where the RMSD deteriorates by more than 0.3 Å are shown in italic.

simulations. When this convergence occurs, the simulations end, and the final lowest atomic statistical potential structure is selected.

RESULTS

LCR

Figure 1 shows the local RMSD of various length pieces of the predicted structure of 1aoy, a DNA binding protein, with an overall RMSD of 4.5 Å. For those pieces that are shorter than nine residues in length, at almost every location the RMSD when compared with native is 2 Å or less. Thus, the first protocol we implement is to use constraints on the local pieces to restrict the conformation space searched. This method will restrict the search to local native conformations. The penalties in potential when the local pieces move out of the predefined RMSD range are not large. The energy penalty is 1 kT per angstrom for any piece of local structure that has an RMSD larger than 2 Å. This will give the structure enough flexibility to overcome barriers that separate the initial structure from a possibly better conformation.

First, we compare this protocol with a MD simulation on two proteins, 1cis and 1tit. In each protein, five starting structures with various initial RMSDs, four with RMSD of less than 6.5 Å and one larger, were selected. These starting structures are the top five predictions from our TOUCHSTONE program and are built with atomic details. The MD simulations were performed using the CHARMM program[25] with the GB implicit solvent model.[26] The initial structures

Table III Refinement Results from Reduced Contact Refinement

Name	Length	Class	Initial RMSD (Å)[a]	Best RMSD (Å)[b]	Final RMSD (Å)[c]
1ixa	39	Small	4.5	3.8	4.3
1rpo	61	Small	3.8	2.9	**3.3**
1c5a	66	α	5.7	4.7	**5.2**
1kjs	74	α	4.8	4.6	4.9
1aoy	78	α	4.5	3.3	**4.1**
2af8	86	α	4.4	4.0	4.4
2lfb	100	α	5.8	5.2	**5.4**
256bA	106	α	3.4	3.0	3.5
1hlb	138	α	3.1	2.9	3.2
1pgx	56	α/β	2.4	2.3	2.6
2ptl	60	α/β	2.6	2.3	2.5
1shaA	103	α/β	4.1	3.1	**3.5**
1cewI	108	α/β	5.7	5.2	5.6
1gpt	47	β	3.5	2.9	3.3
1shg	57	β	4.9	4.5	5.1
1csp	64	β	5.0	3.8	**4.3**
1sro	66	β	6.7	4.6	**5.5**
1ah9	71	β	5.4	5.1	*5.9*
1tit	89	β	2.8	2.3	**2.4**
1ksr	100	β	5.6	5.2	5.5
ave			4.4	3.8	4.2

[a] Initial RMSD is the RMSD of the starting structure that was generated with TOUCHSTONE program.

[b] Best RMSD is the best structures generated during the sampling in the refinement. Of course without the knowledge of native structures, we cannot always pick out these best structures.

[c] Final RMSD is the RMSD of the output of the refinement procedure. These are the lowest energy structures selected by the pairwise atomic statistical potential. The cases where the RMSD improves by more than 0.3 Å are shown in bold; those where RMSD deteriorates by more than 0.3 Å are shown in italic.

are minimized and equilibrated for 100 picoseconds with MD, and the final structure is used as the refinement results. The statistical potential refinement is performed as described in the Methods section and the results are compared with MD data in Table I. In these 10 refinement attempts, MD with GB only improves the structure in two cases; in both cases, the improvement is marginal (<0.3 Å in RMSD). These results are consistent with a recent publication.[15] On the other hand, the LCR procedure improves the structure in seven cases, and improves the structures in six of eight cases when the initial structures have an RMSD of less than 6.5 Å. In six cases, the improvement is nontrivial (larger than 0.3 Å). For three structures, the change is no more than 0.3 Å. In only one case did the structure deteriorate from an RMSD of 7.9 to 8.3 Å.

We then tested the LCR method extensively on a total of 24 proteins. The results are listed in Table II. These proteins have different sizes and belong to different secondary structure classes. The initial structures cover a range of 2.4 to 6.6 Å RMSD from native, with one case having an initial structure with an RMSD from native of 10.3 Å. In 14 (eight) cases, the structures improved by more than 0.3 (0.5) Å; especially for 1sro, the improvement is 1.8 Å. For 1aoy, 1rpo, 1ah9, the improvements are 0.7, 0.6, and 0.6 Å, respectively. In only two cases, 1gpt and 1kjs, did the structures get worse, by 0.6 and 0.4 Å, respectively. For the remaining test cases, the refinement results are within 0.3 Å of the initial structures.

Figure 2 shows a successful refinement case, the beta protein 1sro. On the left is the initial predicted structure that has the correct topology when compared with the native structure shown on the right. But it has an overall RMSD from native of 6.6 Å. The problem is a region on the far left side of the protein, where the predicted structure has a piece of helix that doesn't exist in the native structure. In the refined structure shown in the middle, that wrongly predicted helix is significantly bent, which makes that piece of structure much more nativelike. This shows that wrongly predicted secondary structures can be corrected to favor the correct tertiary interactions; here the refinement procedure reduces the RMSD to 4.8 Å. It is worth mentioning that the final structure has an RMSD of 7.7 Å from the initial structure. The refinement procedure samples quite a large

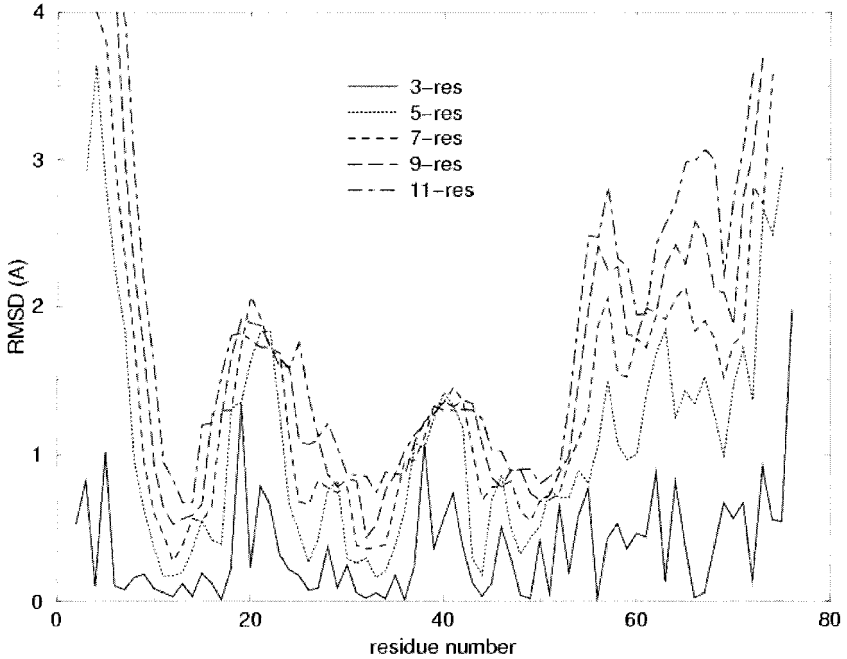

FIGURE 1 Local RMSD between predicted structure and native structure for 1aoy. The global RMSD is 4.6 Å. Each value represents the RMSD from native of a short-piece of structure centered about the residue number shown on the *x*-axis.

range of conformation space even with the local restraint potential. However, in both the initial predicted and the refined structures, the floppy C-termini are different from the native conformation. Further improvement in the refinement procedure will be needed to get the C-terminus correct.

RCR

In the second method, we try to use different constraints than those used in the *ab initio* prediction program. Figure 3 plots the distances of the predicted contacts in the predicted structure 1ksr that has an RMSD of 5.6 Å from native. We can see that for about 25% of the predicted contact pairs, the distances are longer than 10 Å in the initial predicted structure. These contacts are not likely to be satisfied during the refinement. By removing those obviously wrong contacts, and using the rest of the contacts for structure refinement, we can reduce the deleterious effects of wrong contact predictions. An atomic statistical pairwise potential is used to select the final refinement results.

Table III lists the refinement results from the reduced contact refinement procedure on 20 test cases. In eight cases, the structure improved by more than

FIGURE 2 Local constraint refinement result for 1sro. The figure on the left is the initial structure that has an RMSD of 6.6 Å. The middle figure is the refinement result that has an RMSD of 4.8 Å. The figure on the right is the native structure. Protein figures in Figures 2, 4, and 5 were drawn with VMD.[29]

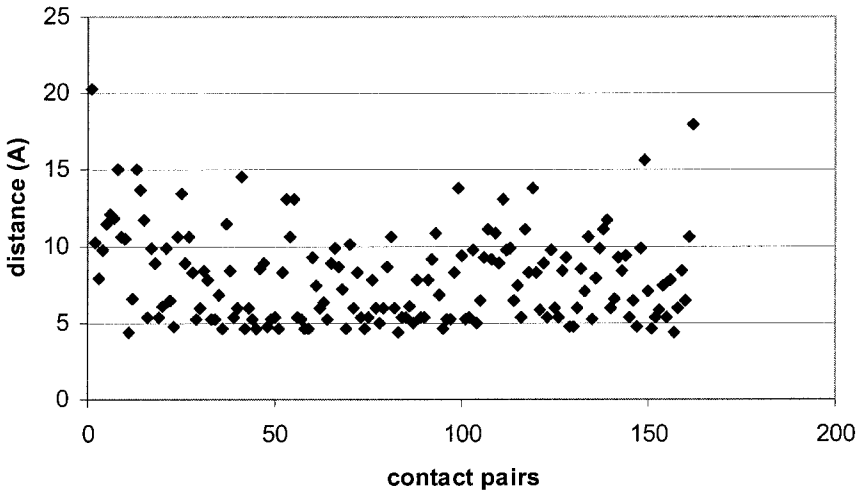

FIGURE 3 Predicted contacts versus the distances in the predicted structures for 1ksr. The RMSD of the predicted structure is 5.6 Å. The *x*-axis is the pair of predicted contacts. Most predicted contacts are already satisfied in the initial structures as the distances between the partners are less than 8 Å. However, there are about 25% of the partners in the predicted contact pair that are more than 10 Å away from each other and may not be satisfied during the refinement.

0.3 Å. In only one case did the structure get worse by more than 0.3 Å. A good refinement example of an α/β protein 1shaA is shown in Figure 4 where the structure improved from 4.1 to 3.5 Å.

Statistical Potential Guided MD

The third method uses an atomic pairwise statistical potential to guide structure refinement. We use a combined MD and atomic statistical potential method. The initial structure is used as a starting point for 20 MD simulations, each 50 picoseconds long. At the end, the 20 final structures are evaluated with an atomic statistical pair potential and the lowest energy one is selected. Then starting from this lowest energy structure, 20 new simulations are performed until the atomic potential converges.

Straightforward simulations have been performed, and the results show that energy drops quickly as the RMSD increases dramatically. We have noticed that the structures collapse and the secondary structures are destroyed during the simulation process. Thus, we constrain the simulation with predicted secondary structures. In each of the 20 simulations, a piece of secondary structure of the protein is selected and moved or rotated by a random distance or random degree. Then the minimization and MD simulations will put the structure in a local minimum.

Ten cases have been tested using this approach. The results are summarized in Table IV. In all cases,

FIGURE 4 Structure refinement for 1shaA using reduced contact refinement. The figure on the left shows the initial structure with an RMSD of 4.1 Å. The middle figure is the refinement result that is 3.5 Å from native structure. The figure on the right shows the native structure.

Table IV Structure Refinement Using the Atomic Pairwise Statistical Potential Guided Molecular Dynamics

PDB Code	Class/Length	Initial RMSD (Å)[a]	Final RMSD (Å)[b]	ΔRMSD (Å)[c]
1ixa	Small/39	4.5	4.9	+0.4
1hp8	α/68	4.9	4.6	−0.3
1pou	α/71	3.8	2.9	−0.9
1ner	α/74	4.1	3.7	−0.4
1bbhA	α/131	6.2	4.8	−1.4
1tfi	β/50	4.6	4.9	+0.3
1shg	β/57	5.1	6.0	+0.9
1tit	β/89	2.5	2.9	+0.4
2fmr	$\alpha\beta$/65	3.7	3.6	−0.1
1cis	$\alpha\beta$/66	4.2	4.8	+0.6
1poh	$\alpha\beta$/85	3.2	3.0	−0.2
ave		4.7	4.6	−0.1

[a] Initial RMSD is the RMSD of the starting structure that was generated with TOUCHSTONE program.

[b] Final RMSD is the RMSD of the output of the refinement procedure. These structures are selected by the pairwise atomic statistical potential.

[c] Changes are the difference between the initial structure and the lowest energy structure obtained from the refinement procedure.

the atomic statistical potential drops significantly. For all-α proteins, modest and occasionally significant improvements have been observed. In four cases, two improved by more than 0.5 Å, the other two improved by more than 0.3 Å. For all-β and mixed α/β proteins, the method didn't work well. An example of a good refinement case for an all-α protein 1bbhA is shown in Figure 5. The initial structure has an RMSD of 6.2

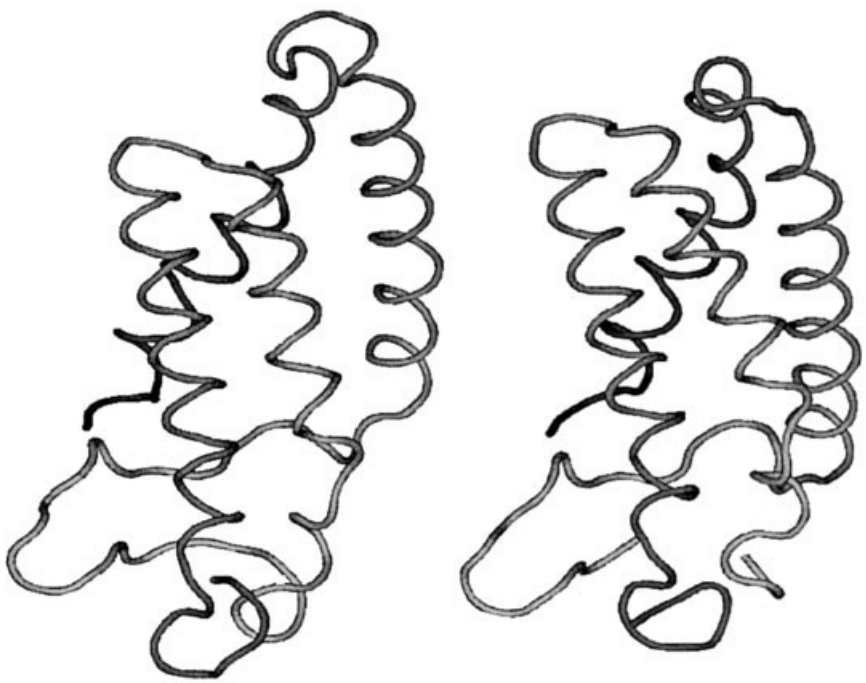

FIGURE 5 Atomic statistical potential guided MD refinement on 1bbhA. The figure on the left is the initial structure that has an RMSD of 6.2 Å, and the figure on the right is the refinement result that has an RMSD of 4.8 Å.

Å. After refinement, the final structure has an RMSD of 4.8 Å. The overall topology of the structures didn't change much during refinement. The improvement is due to the better relative positions of the α-helixes.

Evaluation of the Refinement

For a total of 67 refinement cases, in half of the cases the RMSD improved by more than 0.3 Å, and in only 10% of the cases did the RMSD deteriorate by more than 0.3 Å. Besides RMSD, two measures of the quality of the predicted structure are further used in evaluating the refinement process.

The first measure is the percentage of side chain native contacts (defined as any pair of side chain heavy atoms being within 4.5 Å of each other and found in the native structure) that exists in the predicted structure. When comparing the starting structure with the refined one, the native contact percentage didn't change much. In 60% of the cases, the native contact percentage increased, in 10% of the cases it stayed the same, and in 30% of the cases it decreased. In all cases that the native contact percentage decreased, the difference was less than 5%. In cases that the native contact percentage increases, it can be as high as 13%; for example in 1sro the native contacts increased from 65 to 78%. The native contact percentage improved by an average of 3% in all cases from 59 to 62%.

The second measure is the z-score of the RMSD between the predicted and native structures compared on a sequence-independent basis to obtain the best structure alignment.[30] This z-score measures the probability of finding a structure from random when comparing with the native structures.[27]

In all the pairs of structures aligned, the highest z-scores are reached when at least 80% of the residues are aligned, and the average coverage is 90%. In most refinement evaluations, the z-scores stay the same for initial and refined structures. The average of improvement of the z-score is 0.25 after the refinement. The largest improvement is in 1sro (increased by 0.8) and in 1ksr (increased by 0.6). In all the cases when the z-score decreases, the decrease was never more than 0.3.

DISCUSSION

In this work, we used three methods for structure refinement guided by statistical potentials. In LCR and RCR, we have combined the Monte Carlo search engine with atomic statistical potential selection method to explore near-native conformation space. In these two methods, we have shown that a modest improvement of more than 0.3 Å can be achieved in about half of the cases, while in the other cases the structures don't change by much. And in a few cases, the improvement can be as large as 1 Å. When the atomic pairwise statistical potential is combined with MD, only all-α proteins improved, and the improvement could be large ($>$1 Å) in certain cases.

Obviously our results show the goal of refinement to native structure is still far from successful. But it is the first time a refinement procedure can be used to improve the quality of the structure on a consistent basis. When initial structures are within 6.5 Å in RMSD, for about half of cases, an improvement of more than 0.3 Å can be achieved, while for the other half of the cases the structures do not get worse. For wrongly predicted initial structures (RMSD of 9 or 10 Å and up), the refinement procedure described here won't be able to produce a good structure. Also worth mentioning is that the refinement results for different classes of proteins, all-α, all-β, and mixed α/β, are similar when LCR and RCR are used.

In three test cases, 2ptl, 1pou, and 1sro, a recent publication[15] showed the refinement with MD was not successful; the structures either stayed in the same RMSD or got worse by 0.3 to 0.7 Å. And for 1vif, a MD relaxation also failed to improve the structure. With LCR method, a total of six predicted structures for these four proteins were refined. For two different initial predictions of 1pou, the improvements are 0.2 and 0.5 Å; for both 2ptl and 1vif, the improvements are 0.4 Å; for the two predictions of 1sro, the improvements are 0.3 and 1.8 Å in RMSD.

The reason for the better performance of our method might be due to the fact that the energy landscape in our potential is not as rugged as in a molecular mechanics potential. The use of molecular mechanics potential needs extensive minimization before a reliable evaluation can be made.[22–24,28] Although molecular mechanical potentials have been tested in several works on selecting near-native structures, the number of decoys evaluated is very limited. Thus, its ability to refine a structure has not yet been demonstrated in general.

Currently our refinement procedure is a combination of a residue-based model and an atom-based model or a combination of statistical potential with molecular mechanical potential. The reason we use the combination of a residue and an atom based model is to improve the conformational sampling efficiency.

Comparing LCR and RCR, we found that when using LCR the structures deviate more from the initial structures than when RCR is used. It seems RCR samples less conformational space than LCR. Never-

theless, the cases where improvement is seen do not completely overlap. So we cannot say for sure which method is better. Further testing that combines these two methods is underway.

By using a combined atomic statistical potential with molecular mechanics potential, we have restricted the conformational search to a physically reasonable potential surface. Otherwise, using the atomic statistical potential alone for refinement would destroy the structure. Because secondary structures and hydrogen bonds are inadequately addressed, this shows that the current atomic statistical potential is not complete.

The performance of local restraint and tertiary contact restraint refinement show that the structure improvement with statistical potential is possible. More effort will be spent along this track, such as using an all-atom Monte Carlo search program with a complete set of atomic statistical potentials. However, in principle the statistical potential works the best only for "statistical" proteins. A more reasonable guess is that successful refinement will result from the combination of physics-based and statistical potentials.

This research was supported in part by National Institutes of Health Grant RR12255 of the Division of General Medical Sciences.

REFERENCES

1. Bonneau, R.; Baker, D. Annu Rev Bioph Biom 2001, 30, 173–189.
2. Eyrich, V. A.; Standley, D. M.; Friesner, R. A. J Mol Biol 1999, 288, 725–742.
3. Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. J Mol Biol 2000, 300, 171–185.
4. Pillardy, A.; Czaplewski, C.; Liwo, A.; Lee, J.; Ripoll, D. R.; Kazmierkiewicz, R.; Oldziej, S.; Wedemeyer, W. J.; Gibson, K. D.; Arnautova, Y. A.; Saunders, J.; Ye, Y. J.; Scheraga, H. A. Proc Natl Acad Sci USA 2001, 98, 2329–2333.
5. Kihara, D.; Lu, H.; Kolinski, A.; Skolnick, J. PNAS 2001, 98, 10125–10130.
6. Kihara, D.; Zhang, Y.; Lu, H.; Kolinski, A.; Skolnick, J. PNAS 2002, 5993–5998.
7. Fetrow, J.; Skolnick, J. J Mol Biol 1998, 281, 949–968.
8. Wojciechowski, M.; Skolnick, J. J Comput Chem 2002, 23, 189–197.
9. Kolinski, A.; Jaroszewski, L.; Rotkiewicz, P.; Skolnick, J. J Phys Chem 1998, 102, 4628–4637.
10. Shortle, D.; Simons, K.; Baker, D. Proc Natl Acad Sci USA 1998, 95, 11158–11162.
11. Betancourt, M. R.; Skolnick, J. J Comput Chem 2000, 22, 339–353.
12. Vieth, M.; Kolinski, A.; Brooks, C. L. III; Skolnick, J. DIMACS 1996, 23, 233–236.
13. Simmerling, C.; Lee, M.; Ortiz, A. R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. J Am Chem Soc 1999, 122, 8392–8402.
14. Lee, M. R.; Baker, D.; Kollman, P. J Am Chem Soc 2001, 123, 1040–1046.
15. Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. J Mol Biol 2001, 313, 417–430.
16. Duan, Y.; Kollman, P. Science 1998, 282, 740–744.
17. Skolnick, J.; Kihara, D. Proteins 2001, 42, 319–331.
18. Zhang, Y.; Skolnick, J. J Chem Phys 2001,115, 5027–5032.
19. Lu, H.; Skolnick, J. Proteins 2001, 44, 223–232.
20. Miyazawa, S.; Jernigan, R. Proteins 1999, 36, 357–369.
21. Tobi, T.; Elber, R. Proteins 2000, 41, 40–46.
22. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. Proteins 2002, 48, 404–422.
23. Dominy, B. N.; Brooks III, C. L. J Comput Chem 2001, 23, 147–160.
24. Vorobjev, Y.; Hermans, J. Pro Sci 1999, 36, 407–418.
25. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J Phys Chem B 1998, 102, 3586–3616.
26. Dominy, B. N.; Brooks III, C. L. J Comput Chem 1999, 103, 3765–3773.
27. Betancourt, M. R.; Skolnick, J. Biopolymers 2001, 59, 305–309.
28. Lazaridis, T.; Karplus, M. Curr Opin Struct Biol 2000, 10, 139–145.
29. Humphrey, W.; Dalke, A.; Schulten, K. J Mol Graph 1996, 14, 33–38.
30. Kihara, D.; Skolnick, J. J Mol Biol (in press).