SHORT COMMUNICATION

TASSER-Based Refinement of NMR Structures

Seung Yup Lee,[1] **Yang Zhang,**[2] **and Jeffrey Skolnick**[1*]

[1]*Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York*
[2]*Center for Bioinformatics and Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas*

ABSTRACT The TASSER structure prediction algorithm is employed to investigate whether NMR structures can be moved closer to their corresponding X-ray counterparts by automatic refinement procedures. The benchmark protein dataset includes 61 nonhomologous proteins whose structures have been determined by both NMR and X-ray experiments. Interestingly, by starting from NMR structures, the majority (79%) of TASSER refined models show a structural shift toward their X-ray structures. On average, the TASSER refined models have a root-mean-square-deviation (RMSD) from the X-ray structure of 1.785 Å (1.556 Å) over the entire chain (aligned region), while the average RMSD between NMR and X-ray structures ($RMSD_{NMR\_X-ray}$) is 2.080 Å (1.731 Å). For all proteins having a $RMSD_{NMR\_X-ray}$ >2 Å, the TASSER refined structures show consistent improvement. However, for the 34 proteins with a $RMSD_{NMR\_X-ray}$ <2 Å, there are only 21 cases (60%) where the TASSER model is closer to the X-ray structure than NMR, which may be due to the inherent resolution of TASSER. We also compare the TASSER models with 12 NMR models in the RECOORD database that have been recalculated recently by Nederveen et al. from original NMR restraints using the newest molecular dynamics tools. In 8 of 12 cases, TASSER models show a smaller RMSD to X-ray structures; in 3 of 12 cases, where $RMSD_{NMR\_X-ray}$ <1 Å, RECOORD does better than TASSER. These results suggest that TASSER can be a useful tool to improve the quality of NMR structures. Proteins 2006;63:451–456.
© 2006 Wiley-Liss, Inc.

Key words: NMR structure refinement; X-ray structure; TASSER; Protein structure prediction

INTRODUCTION

The majority of the protein structures in the Protein Data Bank (PDB)[1] are determined by either nuclear magnetic resonance (NMR) spectroscopy or X-ray crystallography.[2,3] Because NMR models are obtained by optimally satisfying experimental restraints,[4] NMR experiments usually provide an ensemble of models that may represent both the thermal motion of the protein chain in solution[5,6] as well as inaccuracies in the force field used in NMR refinement. X-ray crystallography requires that the protein crystallize and that the crystal be of good quality.[7] The protein structure determined by X-ray diffraction results from converting the electron density data into a model, and it is generally of higher resolution than NMR, although crystal packing effects may distort the protein structure.[8–11]

Several studies[12–14] have shown that the structure generation protocols of NMR often use severely simplified nonbonded interactions for the sake of speed, which may lead to structural problems. Using molecular dynamics simulations, Fan and Mark[15] analyzed the relative stability of NMR and X-ray structures and found that X-ray structures are significantly more stable than NMR structures. The comparison of NMR and X-ray structures by Garbuzynskiy et al.[16] revealed that for the same proteins, there are differences in all side chain, backbone, and backbone–side chain contacts as well as in the backbone–backbone hydrogen bonds. In addition, they also found that NMR structures can be altered significantly depending on the algorithms used to construct the model from the experimental restraints. Therefore, although NMR sometimes provides additional structural and dynamic information not available from X-ray structures and NMR restraints are readily implemented into protein structure prediction algorithms,[17,18] there is considerable room for improvement of the quality of NMR structures.[12,14,19–21]

In principle, NMR structures in solution need not necessarily be the same as the X-ray structure determined from a crystalline environment that may restrict structural motion. However, because X-ray structures are highly hydrated, one might envision that such effects are mini-

mal; in such a case, NMR and X-ray structures should be similar. Therefore, it is of interest to explore whether NMR structures on appropriate refinement could move closer to X-ray structures. Of course, to engage in such a refinement study, an algorithm that can actually refine protein structures is essential.

In this work, to address this intriguing question, we perform the structure refinement of NMR PDB structures and compare our final models with X-ray structures. For refinement, we employ our modeling algorithm, TASSER (Threading ASSEmbly Refinement), which showed the ability to recognize the majority of nonevolutionarily related folds in the PDB library as well as for structure refinement in comprehensive benchmarks as well as in CASP6.[22–25] The original version of TASSER generates models by using template alignments from our sequence/structure alignment algorithm, PROSPECTOR_3.[26] For the current situation, we generate templates directly from the NMR structures. The test set includes 61 nonhomologous proteins with size ranging from 46 to 188 residues and having both NMR and X-ray structures available in the PDB. We will also compare our refined models with recently recalculated NMR protein structures in the RECOORD database.

MATERIALS AND METHODS

TASSER consists of template identification, structure assembly, and final model selection. Because detailed descriptions of TASSER are available,[22–24] here we merely provide a brief overview.

Template Identification

For a given target sequence, TASSER usually uses template information generated by our threading program PROSPECTOR_3,[26] an iterative sequence/structure alignment approach. Here, because our objective is to refine NMR structures in the PDB library, we take templates directly from NMR-based PDB files, each of which usually contains more than 10 models. Only the consensus regions of the NMR models will be used as TASSER input templates. To define the consensus regions, we superimpose all models with the first model of the NMR PDB file and calculate the distance, $d(i,m)$, between the i-th C_α atom pairs in the first and m-th models. The averaged distance of residue i over all the models is defined as

$$\overline{d}(i) = \frac{1}{N-1} \sum_{m=2}^{N} d(i,m),$$

where N is the total number of models in the NMR PDB file. The i-th residue will be considered as a consensus residue if $\overline{d}(i)$ is less than a cutoff distance, σ_{cut}. We adjust σ_{cut} so that the consensus region has 90% coverage.

On- and off-Lattice C_α and Side Chain-Based (CAS) Model

In TASSER, a protein is represented by its C_α atoms and side-chain centers of mass (SG), called the CAS model. The aligned regions (here, consensus regions in the NMR models) provide structural fragments whose internal geometry remains unchanged during the simulations. These are off-lattice. The remaining C_α atoms belonging to the unaligned regions (here, inconsistent regions in the NMR models) are confined to an underlying cubic lattice system.

The CAS force field consists of four classes of terms: (1) predicted secondary structure propensities, (2) statistical propensities including local C_α correlations, hydrogen bonding and hydrophobic burial interactions, (3) consensus side-chain contacts taken from the NMR models, and (4) protein specific SG-pair potentials. The force field we use was previously optimized on a set of 100 nonhomologous training proteins (extrinsic to the proteins considered here) that consists of a random collection of NMR and X-ray structures;[22] as such it is not biased to either X-ray or NMR structures. Except for the fact that the templates and consensus contacts are prepared from NMR structures, this work uses the same procedure as in the original TASSER method; therefore, all other propensities are predicted ones over the entire chain.

Structure Assembly

Because template alignments provide no information about the unaligned or loop regions, these portions of the chains are predicted by the ab initio component of TASSER. Zhang and Skolnick[24] used two types of parameters to evaluate the accuracy of TASSER prediction for these unaligned or loop regions: the $RMSD_{local}$ for the local conformation and $RMSD_{global}$ for both the local conformation and global orientation. They observed that the accuracy of unaligned or loop modeling decreases with increasing size; interestingly, the local conformation of the chain is rather well predicted, while the global orientation becomes worse for large size loops. A similar situation was noted here as well. Therefore, to enhance the modeling accuracy of TASSER, we split the unaligned or loop region into two small parts by assigning a center unaligned residue as an aligned one and TASSER is then rerun. By shortening the length of unaligned region, we slightly improve the quality of the final models.

Because template identification just provides coordinates of the aligned regions, an initial full-length model is prepared by connecting the continuous template fragments by a random walk of C_α–C_α lattice bond vectors. The unaligned regions serve as linkages for rigid block movement. There are two kinds of conformational updates: (1) off-lattice movements involving rigid fragment translation and rotations, and (2) lattice confined residues are updated by two- to six-bond movements and multibond sequence shifts. From the initial full-length model, conformational space is sampled using Parallel Hyperbolic Monte Carlo Sampling[27] where 40 replicas are used in TASSER. Of these, the 14 lowest temperature replicas are submitted to the structural clustering program, SPICKER,[28] to select the final models. These models are ranked by structure density. Finally, we analyze the final models by comparing those with X-ray PDB structures.

RESULTS AND DISCUSSIONS
Benchmark NMR/X-ray PDB Data Set

For test proteins, we use a NMR/X-ray PDB data set having both NMR and X-ray structures available in the PDB library, which has been previously used in the comparison study of X-ray and NMR structures by Garbuzynskiy et al.[16] By carefully inspecting the SCOP database 1.63 release,[29] they constructed the data set by 10 criteria; for example, there is no modified residue in chain, all heavy atoms have coordinates, the sequence identity between different chosen proteins does not exceed 30%,[30] the NMR and X-ray structures correspond to the same sequence except perhaps for a few residues at the end of chain, and nonwater ligands are either identical or small (which means they contain fewer than 10 heavy atoms). Here, we add one additional criterion to their original NMR/X-ray PDB test protein set: we exclude protein pairs where the NMR PDB file contains only one model structure for which the conserved structural region cannot be identified. This results in 61 NMR/X-ray pairs with length from 46 to 188 residues. Of these, 16, 17, 24, and 4 are all α-, all β-, αβ-, and small proteins, respectively.

Refinement of NMR PDB Structures

Here, as previously mentioned, TASSER uses a subset of NMR structures that are structurally conserved as the template. We note that the main objective of this work is to examine whether TASSER refined models are closer to X-ray structures; therefore, the RMSD and TM-score[31] (a measure of similarity of global topologies of protein structures ranging from [0,1], with 0.17 the average value of a pair of random structures independent of chain length. A TM-score of 1.0 means that the two structures are identical) values are calculated with respect to the X-ray structures.

Figure 1(a) shows the comparison of RMSD of NMR models and that of the TASSER refined models, which are calculated over the same aligned region (gray circles). Seventy-nine percent of the final models (48 of 61 proteins) show an improvement (smaller RMSD to X-ray structure) over the initial NMR structure-based templates. We also compare the full-length RMSD values of the initial NMR structures with those of TASSER refined models (black triangles). Again, the majority of TASSER refined models have greater similarity to X-ray structures than NMR structures. The final models have an average RMSD of 1.785 Å (1.556 Å) over the entire chain (aligned region), while NMR structures have an average RMSD of 2.080 Å (1.731 Å). In Figure 1(b), we show the comparison of the TM-score of the TASSER models and NMR structures. The average TM-score of the structures increases from 0.828 to 0.846 after TASSER refinement. We note that X-ray structures are slightly more compact than NMR structures. Their average radius of gyration of 12.74 Å is slightly less than that of NMR structures (12.86 Å). The final models have an average radius of gyration of 12.65 Å, indicating that they are slightly more compact than both NMR and X-ray structures.
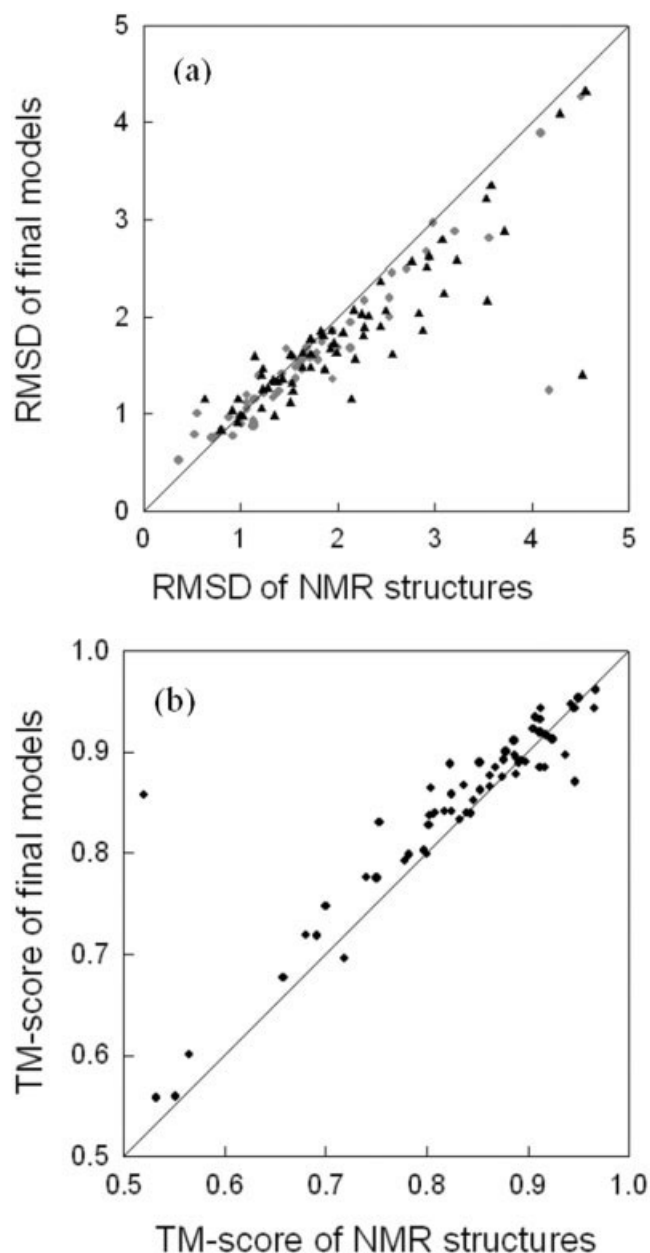

Fig. 1.   Comparison of (a) the RMSD of TASSER final models with the RMSD of the NMR structures over the same aligned region (gray circles) and entire chain (black triangles). (b) Scatter plot of TM-score of TASSER final models versus TM-score of NMR structures. All RMSD and TM-score are calculated with respect to X-ray structures.

A representative example for an NMR/X-ray pair (1acp/1l0h) is shown in Figure 2(a). The initial NMR structure has a RMSD/TM-score of 4.431 Å/0.520 over the entire chain. After refinement, TASSER generates a final model which has a RMSD/TM-score of 1.396 Å/0.857. Obviously, in the majority of cases, the TASSER refined models are closer to the X-ray structures than to the NMR structures.

To assess the performance of TASSER modeling in detail, we show a histogram of the cumulative fraction of better and worse TASSER refined models in Figure 3. There are five NMR/X-ray pairs (1oca/2cpl, 1d3z/
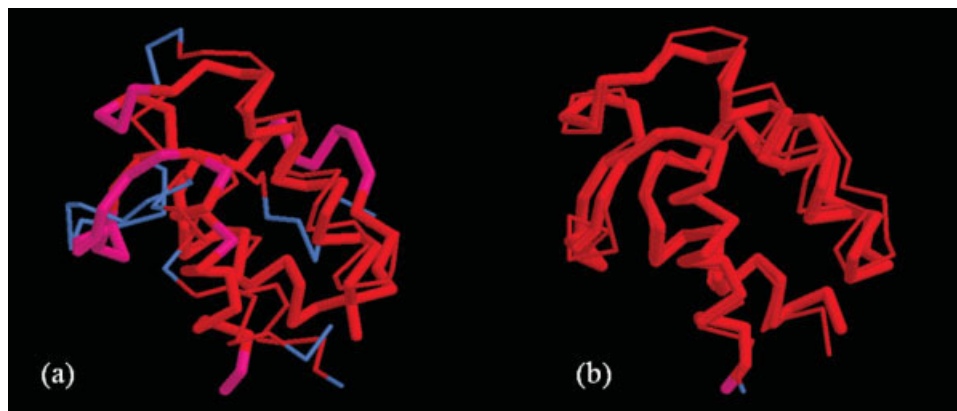
Fig. 2. Representative example showing the improvement of the TASSER final model with respect to the NMR structure. (a) NMR (1acp) structure superimposed to X-ray structure (1l0h) of the same protein with a RMSD/TM-score of 4.431 Å/0.520. (b) TASSER final model superimposed on the X-ray structure with a RMSD/TM-score of 1.396 Å/0.857. The thick lines are the X-ray structure and the thin lines are (a) NMR structure and (b) TASSER final model. Red color identifies those residue pairs of distance <5 Å after TM-score superposition.
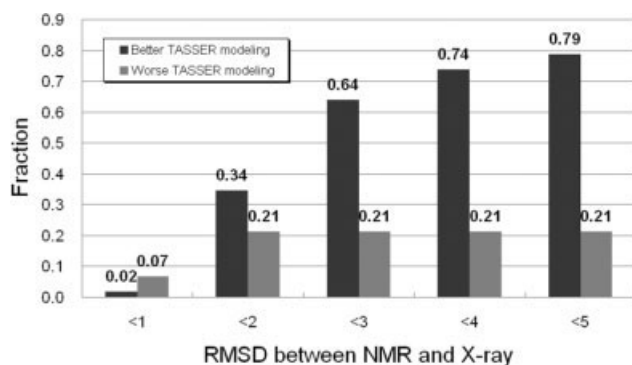

Fig. 3. Histogram of the cumulative fraction of better and worse TASSER modeling cases as compared to the RMSD between NMR and X-ray structures.

1ubi, 3gb1/1pgb, 1r63/1r69, and 1aey/1shg) that have a $RMSD_{NMR\_X-ray}$ <1 Å. In this regime, (which is actually at or below the limit of resolution of TASSER and most experimental results), TASSER moves only one target (1d3z/1ubi) slightly closer to the X-ray structure. However, as the $RMSD_{NMR\_X-ray}$ increases to between 2 to 5 Å, a regime where experimental differences become more meaningful, TASSER refined models show a clear improvement compared with NMR structures. Especially, we note that although TASSER fails to improve some proteins (13 of 61) whose NMR and X-ray structures are very similar ($RMSD_{NMR\_X-ray}$ <2 Å), TASSER successfully refines all models toward X-ray structures for proteins having a $RMSD_{NMR\_X-ray}$ >2 Å in our test set.

Table I presents a summary of TASSER modeling results starting from NMR templates. Except the case for five proteins having $RMSD_{NMR\_X-ray}$ <1 Å, TASSER refined models have a smaller RMSD and larger TM-score than those of NMR structures. All results for each of 61 proteins are available on our homepage, http://www.bioinformatics.buffalo.edu/current_buffalo/TASSER/NMR_X-ray/.

Recently, Nederveen et al.[21] constructed a refined NMR database (RECOORD) by recalculating NMR structures for which coordinates and NMR restraints are available from the Protein Data Bank (http://www.ebi.ac.uk/msd-srv/docs/NMR/recoord/main.html), by using state-of-the-art atomic molecular dynamics tools. In the RECOORD database, we find 12 proteins that are included in our protein set. Table II shows the comparison of TASSER refined models with protein models in the RECOORD database (CNW set from CNS). Each protein model in the RECOORD database has 25 structures, and the calculated RMSD to the X-ray structure is the averaged value over these 25 structures. For the 12 proteins, the protein models in RECOORD also have a smaller average RMSD of 2.358 Å than the RMSD (2.423 Å) of the original NMR structures. However, TASSER refined models are on average closer to the X-ray structures (average RMSD of 2.193 Å) than the models in RECOORD. For 8 of 12 proteins, TASSER refined models are closer than those of RECOORD. There are three NMR/X-ray pairs (1oca/2cpl, 3gb1/1pgb, and 1r63/1r69) having an $RMSD_{NMR\_X-ray}$ <1 Å for which the RECOORD models are closer than our models. In the remaining one case (1k19/1kx8), the TASSER refined model also shows improvement over original NMR structure, although the improvement due to RECOORD is better.

CONCLUSIONS

In this work, we combine spatial restraints taken from NMR structures with the inherent force field of TASSER and examine whether the NMR structures could be moved closer to their X-ray determined counterparts, because X-ray experiments have been shown in general to have higher stability and structural accuracy. We applied this methodology to a benchmark set of 61 NMR/X-ray structure pairs. The majority of TASSER refined models (48 of 61) clearly become closer to their X-ray structures than the starting NMR structures. The TASSER refined models

TABLE I. Comparison of Final Models from TASSER with X-ray Structures

RMSD NMR/X-ray	RMSD to X-ray (ali)[a]		RMSD to X-ray (all)[b]		TM-score to X-ray (all)[c]	
	NMR	Model	NMR	Model	NMR	Model
<1	0.664	0.814	0.860	1.014	0.940	0.915
<2	1.222	1.179	1.428	1.377	0.885	0.887
<3	1.495	1.384	1.797	1.601	0.850	0.861
<4	1.600	1.474	1.959	1.709	0.841	0.853
<5	1.731	1.556	2.080	1.785	0.828	0.846

RMSD to X-ray structures:
[a]NMR and final TASSER models over the same aligned regions;
[b]NMR and TASSER models over the entire chain.
TM-score to X-ray structures:
[c]NMR and TASSER models over the entire chain.

TABLE II. Comparison of Final Models from TASSER with Recalculated NMR Structures in the RECOORD Database

PDB name		RMSD to X-ray (all)		
NMR	X-ray	NMR	TASSER	RECOORD
1bip	1blu	3.090	2.233	2.837
1eq0	1hka	3.230	2.581	3.000
1jnj	1lds	3.532	3.209	3.341
1jor	1ey4	2.321	2.002	2.482
1k19	1kx8	4.555	4.319	4.212
1ssn	1c76	4.288	4.089	4.152
1oca	2cpl	0.920	1.033	0.864
1d3z	1ubi	0.970	0.909	1.241
3gb1	1pgb	0.629	1.146	0.808
1pf1	1fil	1.727	1.608	1.725
1r63	1r69	0.978	1.151	0.821
1cxr	1btn	2.831	2.032	2.814
Average		2.423	2.193	2.358

have an average RMSD of 1.785 Å (1.556 Å) over the entire chain (aligned region) and an average TM-score of 0.846. The corresponding NMR structures have an average RMSD of 2.080 Å (1.731 Å) and a TM-score of 0.828. In 13 of 34 proteins whose NMR and X-ray structures are very close ($RMSD_{NMR\_X-ray}$ <2 Å), TASSER fails to drive the NMR model closer to the X-ray structure, which shows the need for a higher level of representation of the protein in the TASSER force field for refinements at this resolution. However, we note that for all proteins with a $RMSD_{NMR\_X-ray}$ >2 Å in our benchmark set, TASSER provides the consistent improvement compared with NMR structures. We also compare our TASSER refined models with recalculated NMR structures by the newest molecular dynamics tools from original NMR restraints in the RECOORD database. Most of the TASSER refined models are closer to their X-ray structure than those of RECOORD, with the exception of three proteins having a $RMSD_{NMR\_X-ray}$ <1 Å, a region where the accuracy of TASSER modeling is limited by the reduced representation.

After TASSER runs, there are 26 final models whose unaligned loops regions show a relatively large $RMSD_{local/global}$. To improve the accuracy of TASSER modeling for unaligned or loop regions, we shorten the length of these regions and rerun TASSER. This is accomplished by splitting the unaligned region in half and

inserting an aligned residue in the middle of the loop. Because this residue is off-lattice, this increases the efficiency of the conformational search. Interestingly, for 23 of 26 proteins, this slightly improves the RMSD and TM-score of the final models. For example, for the 1myf/1duk NMR/X-ray pair, the RMSD of the final model decreases from 1.686 to 1.266 Å and the TM-score increases from 0.933 to 0.947. The utility of this approach for the more general problem of structure prediction will be explored in future work.

Overall, the work shows the possibility of systematically improving the quality of NMR structures by combining NMR consensus motifs with automated structure modeling procedures. The further application of current methodology to all NMR structures in the PDB library is in progress.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Branden C, Tooze J. Introduction to protein structure. New York: Garland Science; 1999.
3. Fersht AR. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: Freeman; 1999.
4. Guntert P. Structure calculation of biological macromolecules from NMR data. Q Rev Biophys 1998;31:145–237.
5. Doreleijers JF, Rullmann JA, Kaptein R. Quality assessment of NMR structures: a statistical survey. J Mol Biol 1998;281:149–164.
6. Doreleijers JF, Raves ML, Rullmann T, Kaptein R. Completeness of NOEs in protein structure: a statistical analysis of NMR. J Biomol NMR 1999;14:123–132.
7. Drennth J. Principles of protein X-ray crystallography. New York: Springer-Verlag; 1999.
8. Doreleijers JF, Vriend G, Raves ML, Kaptein R. Validation of nuclear magnetic resonance structures of proteins and nucleic acids: hydrogen geometry and nomenclature. Proteins 1999;37:404–416.
9. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins 2001;44:79–96.
10. Spronk CA, Linge JP, Hilbers CW, Vuister GW. Improving the quality of protein structures derived by NMR spectroscopy. J Biomol NMR 2002;22:281–289.
11. Spronk CA, Nabuurs SB, Bonvin AM, Krieger E, Vuister GW, Vriend G. The precision of NMR structure ensembles revisited. J Biomol NMR 2003;25:225–234.
12. Linge JP, Nilges M. Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. J Biomol NMR 1999;13:51–59.
13. Kuszewski J, Clore GM. Sources of and solutions to problems in

the refinement of protein NMR structures against torsion angle potentials of mean force. J Magn Reson 2000;146:249–254.

14. Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AM, Vuister GW, Vriend G, Spronk CA. DRESS: a database of refined solution NMR structures. Proteins 2004;55:483–486.

15. Fan H, Mark AE. Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: a molecular dynamics simulation study. Proteins 2003;53:111–120.

16. Garbuzynskiy SO, Melnik BS, Lobanov MY, Finkelstein AV, Galzitskaya OV. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray and NMR-resolved protein structures? Proteins 2005;60:139–147.

17. Li W, Zhang Y, Kihara D, Huang YJ, Zheng D, Montelione GT, Kolinski A, Skolnick J. TOUCHSTONEX: protein structure prediction with sparse NMR data. Proteins 2003;53:290–306.

18. Li W, Zhang Y, Skolnick J. Application of sparse NMR restraints to large-scale protein structure prediction. Biophy J 2004;87:1241–1248.

19. Kordel J, Pearlman DA, Chazin WJ. Protein solution structure calculation in solution: solvated molecular dynamics refinement of calbindin D9k. J Biomol NMR 1997;10:231–243.

20. Kuszewski J, Gronenborn AM, Clore GM. Improvements and extensions n the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J Magn Reson 1997;125:171–177.

21. Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, Nabuurs SB, Guntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AM. RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins 2005;59:662–672.

22. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594–7599.

23. Zhang Y, Skolnick J. Tertiary structure prediction on a comprehensive benchmark on medium to large size proteins. Biophy J 2004;87:2647–2655.

24. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 2004;102:1029–1034.

25. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61:91–98.

26. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;56:502–518.

27. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.

28. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 2004;25:865–871.

29. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.

31. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.