Application of Sparse NMR Restraints to Large-Scale Protein Structure Prediction

Wei Li, Yang Zhang, and Jeffrey Skolnick
Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York 14203

ABSTRACT The protein structure prediction algorithm *TOUCHSTONEX* that uses sparse distance restraints derived from NMR nuclear Overhauser enhancement (NOE) data to predict protein structures at low-to-medium resolution was evaluated as follows: First, a representative benchmark set of the Protein Data Bank library consisting of 1365 proteins up to 200 residues was employed. Using *N*/8 simulated long-range restraints, where *N* is the number of residues, 1023 (75%) proteins were folded to a C_α root-mean-square deviation (RMSD) from native $<$6.5 Å in one of the top five models. The average RMSD of the models for all 1365 proteins is 5.0 Å. Using *N*/4 simulated restraints, 1206 (88%) proteins were folded to a RMSD $<$6.5 Å and the average RMSD improved to 4.1 Å. Then, 69 proteins with experimental NMR data were used. Using long-range NOE-derived restraints, 47 proteins were folded to a RMSD $<$6.5 Å with *N*/8 restraints and 61 proteins were folded to a RMSD $<$6.5 Å with *N*/4 restraints. Thus, TOUCHSTONEX can be a tool for NMR-based rapid structure determination, as well as used in other experimental methods that can provide tertiary restraint information.

INTRODUCTION

The prediction of the three-dimensional structure of proteins from their primary sequences has increased in importance as additional genomes have been sequenced (Baker and Sali, 2001; Skolnick et al., 2000), but the application of pure ab initio approaches to protein structures has been limited to quite small proteins (Liwo et al., 1999; Simons et al., 2001; Zhang et al., 2003). However, it has been found that in ab initio protein structure prediction, sparse distance restraints can be sufficient to guide folding to a correct structure, which otherwise would be difficult to obtain (Kolinski and Skolnick, 1998; Skolnick et al., 1997). Several articles have been published on this subject (Aszodi et al., 1995; Bowers et al., 2000; Connolly et al., 1994; Kolinski and Skolnick, 1998; Li et al., 2003; Skolnick et al., 1997; Smith-Brown et al., 1993). For example, Smith-Brown et al. (1993) modeled a protein as a chain of glycine residues using a substantial number of tertiary restraints. Connolly et al. (1994) used an off-lattice reduced representation of proteins to estimate the tertiary structure from incomplete and approximate nuclear Overhauser enhancement (NOE) distance data (0.5 long-range restraints per residue). Aszodi et al. (1995) used a distance geometry-based method to assemble protein structure using experimental tertiary distance restraints supplemented by predicted interresidue distance restraints extracted from multiple sequence alignments. On average, more than *N*/4 restraints, where *N* is the number of residues, were required to obtain structures with a root-mean-square deviation (RMSD) $<$5 Å from native. Skolnick and Kolinski used a high-coordination lattice-reduced model of protein structure and a knowledge-based force field (Kolinski and Skolnick, 1998; Skolnick et al., 1997). Nine proteins up to 247 residues in length were folded to moderate resolution with as few as *N*/7 long-range restraints and some knowledge of the secondary structure. Bowers et al. selected peptide fragments from proteins of known structure based on sequence similarity and consistency with the chemical shift and NOE data, and then assembled proteins to high resolution using $\sim$1 NOE restraint per residue (Bowers et al., 2000). Most recently, Li et. al. developed an algorithm, TOUCHSTONEX, which folded 86% of proteins to moderate resolution with *N*/8 long-range restraints using a test set of 125 proteins up to 174 residues in length (Li et al., 2003).

One of the most commonly used sets of distance restraints come from NOE data generated from NMR experiments and serve as the key element in NMR structure determination. Although traditional NMR structure determination methods require a large number of NOE restraints to define a high-resolution structure, sparse NOE data are relatively easy to obtain even in the early stage of NMR structure determination process. As demonstrated in our recent publication (Li et al., 2003), TOUCHSTONEX incorporates a limited number of distance restraints into the force field as an NOE-specific pairwise interaction. The algorithm was evaluated using 125 proteins of various secondary structure types and lengths up to 174 residues. Using as few as *N*/8 long-range contact restraints randomly selected from the native protein structure, where *N* is the number of residues, 108 proteins (86%) were folded to $<$6.5 Å from the native protein structures within the top five lowest energy clusters. One-hundred three (82%), 86 (69%), 64 (51%), 41 (33%), and 9 (7%) proteins were folded to $<$6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. The average C_α-RMSD of the

Submitted April 20, 2004, and accepted for publication May 17, 2004.

Address reprint requests to Jeffrey Skolnick, Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Buffalo, NY 14203. E-mail: skolnick@buffalo.edu.

0006-3495/04/08/1241/08 $2.00

doi: 10.1529/biophysj.104.044750

lowest RMSD cluster centroids for all 125 proteins (folded and unfolded) is 4.4 Å. Moreover, three proteins with limited experimental NOE data—Z-domain of staphylococcal protein A (58 residues) (Tashiro et al., 1997), the C-terminal BRCA-1-like domain from *Thermus thermophilus* DNA ligase BRCT (92 residues) (G. Sahota, S. Goldsmith-Fischman, B. Dixon, Y. J. Yuang, J. Aramini, C. Yin, R. Xiao, A. Bhattacharya, D. Monléon, G. V. T. Swapna, S. Anderson, B. Honig, A. N. A Monteiro, G. T. Montelione, and T. Tejero, unpublished data), and the human melanoma inhibitory activity protein MIA (108 residues) (Lougheed et al., 2001)—were folded to low-to-medium resolution structures.

The evaluation of the algorithm based on the 125-protein test set in the previous article was not comprehensive. One reason is that the protein set covers only a very limited number of the topologies adopted by proteins. Considering that there are ~500 fold families in the CATH protein structure classification database, the 125 proteins can by no means cover all the representative topologies in the current Protein Data Bank (PDB). Another reason is that due to the small size of the 125-protein set, there is a danger of overtuning the parameters. Before applying the algorithm on a genomic scale, a much larger and truly representative testing set should be used. In this article, we further evaluate TOUCHSTONEX on a representative PDB protein set of proteins up to 200 residues in length. The set consists of 1365 proteins that cover the whole PDB at the level of sequence identities <35%. By testing the algorithm on this protein set, a better understanding of how the algorithm will perform in large-scale applications will be provided. We predicted the structures of these 1365 proteins using randomly selected long-range contact restraints from the native protein structure. We then focused on 69 proteins in this set that have experimental NMR data. These proteins have more complicated topologies than the three proteins with NMR data examined in the previous article (Li et al., 2003). We predicted the structures of these proteins using NOEs mainly involving the main-chain atoms and sometimes also NOEs involving side-chain methyl groups. These NOEs tend to be assigned early in the NMR structure determination process. The results will give a real-life performance test of the algorithm.

METHODS

Protein model, force field, and implementation of NOE-specific pairwise interaction

The protein model, the force field, and the implementation of NOE-specific pairwise interaction have been described in detail in our previous TOUCHSTONEX article (Li et al., 2003) as well as in another article (Zhang et al., 2003). Here, we only give a brief description. The CABS model, which is a lattice-based reduced protein model, represents each amino acid by up to three united atom groups—the C_α, C_β, and side-chain center of mass. NOE-derived contact restraints are incorporated into the force field as a square-well penalty. Appropriate to the type of NOE restraint, a penalty is added between the side-chain centers of mass, between the side-chain center of mass and the C_α, or between C_αs as appropriate. The overall force field also consists of other knowledge-based terms to produce protein-like behavior, including various short-range interactions, hydrogen bonding, one-body, pairwise, and multibody long-range interactions. Besides the NOE-specific penalty, the force field has another penalty term that incorporates predicted contact restraints (Kihara et al., 2001) from the threading algorithm PROSPECTOR_3 (Skolnick et al., 2004).

Protein set

The protein set consists of 1365 representative proteins selected from the PDB. There are two selection criteria: 1), the size of the protein must be from 41 to 200 residues and 2), their pairwise sequence identity must be <35%. The final 1365 proteins include 392 α-proteins, 429 β-proteins, 536 α/β-proteins, and an additional eight proteins with little regular secondary structure. This protein set is the same as the one used in our recent articles (Skolnick et al., 2004; Zhang and Skolnick, 2004a), but excludes the proteins that have fewer than $N/4$ long-range side-chain contacts, where N is the number of the residues (see the following section for the generation of restraints).

From this protein set, there are 69 proteins (5 α-proteins, 30 β-proteins, and 34 α/β-proteins) that have experimental NMR data in the PDB (ftp://ftp.rcsb.org/pub/pdb/data/structures/divided/nmr_restraints/); these were used to test the algorithm with experimental NMR restraints.

Generation of long-range contact restraints

For the set of 1365 proteins, simulated contact restraints are randomly selected from side-chain contacts in the native protein structure. Two side chains that have at least one pair of heavy atoms within 4.5 Å are considered to be in contact. The simulated restraints are also termed ''the correct restraints'' in contrast to the predicted and sometimes inexact restraints generated by PROSPECTOR_3.

For the 69 proteins with experimental NMR data, NOE-derived contact restraints were used. The proton NOE data selected are mainly between main-chain atoms (H_α, H_N), because these NOEs tend to be recognized first during the NMR structure determination process. NOEs between side-chain methyl groups are also selected sometimes (e.g., for α-proteins), because these NOEs are also relatively easy to identify in the early stages of the NMR structure determination process. The atomic level NOE data are then converted into contact restraints between residues.

For both simulated and NOE-derived contact restraints, only long-range restraints (contact partners at least five residues apart along the protein chain) were used.

Conformational updates and Monte Carlo sampling scheme

Conformational updates invoke five kinds of C_α-chain movements: the basic 2- and 3-bond movements, 4-, 5-, and 6-bond movements, 6–12 bond transitions, multibond sequence shifts, and chain end movements (Zhang et al., 2003). The conformational sampling scheme uses a newly developed parallel hyperbolic sampling method (Zhang et al., 2002) that differs from the regular replica exchange sampling method by flattening the local high-energy barriers by a nonlinear transformation to alleviate the problem of ''ergodicity breaking''. The folding protocol consists of an annealing part followed by an isothermal run (Li et al., 2003). Usually a calculation takes up to 48 h of CPU time on a 1.26-GHz Pentium III processor for a protein of no more than 200 residues.

Structure clustering, ranking, and evaluations

Twenty-four thousand structures selected from various temperature replicas are clustered. The clusters are ranked according to the cluster density. For each cluster, optimally aligning the structures and computing their average coordinates determine a centroid. The centroids are compared with the native protein structure, and their C_α coordinate root-mean-square deviations (C_α-RMSD) from native are calculated. A protein is considered to have been successfully folded if there is at least one cluster centroid with a C_α-RMSD from native $<$6.5 Å in the top five lowest energy clusters. The lowest C_α-RMSD cluster centroid is considered to be the best structure.

Different from our previous TOUCHSTONEX article (Li et al., 2003), which used a clustering algorithm developed by Betancourt et al. (Betancourt and Skolnick, 2001), a newly developed clustering algorithm, SPICKER (Zhang and Skolnick, 2004b) was used here to cluster the structures. We found that for the 1365 protein set, SPICKER could on average identify 10–13% more proteins with a C_α-RMSD $<$6.5 Å from native in the top five lowest energy clusters than the previous clustering algorithm. The average RMSD of the best structure in the top five lowest energy clusters was between 0.8 and 0.9 Å better.

When comparing the cluster centroids of the proteins with NMR data with the native structures, the nonflexible part of the structures is considered. The nonflexible part of the native protein structure is determined either from the conserved part of the overlapped models (when there are several NMR models) or from the temperature factors (when there are temperature factors for NMR models).

RESULTS AND DISCUSSION

Structure prediction of 1365 benchmark proteins using simulated restraints

Structure prediction results for the 1365 benchmark protein set can be found on web site: http://www.bioinformatics.buffalo.edu/touchstonex/benchmark1365.

A summary of the results from the structure prediction of the 1365 benchmark proteins is shown in Table 1. Column four lists the prediction results using *N*/8 simulated correct (but randomly chosen) long-range contact restraints as well as the predicted contact restraints generated from the threading algorithm PROSPECTOR_3. One-thousand twenty-three (75%) proteins out of the 1365 protein set were foldable, i.e., there is at least one cluster whose centroid C_α-RMSD from native is $<$6.5 Å in the top five lowest energy clusters. Nine-hundred sixty (70%), 817 (60%), 627 (46%), 333 (24%), and 42 (3%) proteins were folded to a RMSD from native $<$6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. The average RMSD of the lowest RMSD (best) cluster centroids in the top five clusters is 5.0 Å for all 1365 proteins. The best cluster ranks 2.4 on average.

Compared with the results in the previous TOUCHSTONEX article (Li et al., 2003), the results shown here are somewhat worse; for *N*/8 restraints the average RMSD is 4.4 Å, and 86% are foldable. The reason is obvious: the protein set used in the previous article was much smaller and the proteins were not as large and complicated. The 1365 proteins in this set cover various types of proteins in PDB amenable for ab initio folding ($\leq$200 residues) with pairwise sequence identity $<$35%. The results for this large protein set are more objective and realistic.

When more restraints are used, as expected, there is a significant improvement in the overall results. Table 1, column six, shows the prediction results using *N*/4 correct long-range contact restraints together with the predicted contact restraints. One-thousand six (88%) proteins were folded to a RMSD from native $<$6.5 Å in the top five lowest energy clusters, which is 183 (13%) more target proteins folded than when only *N*/8 correct restraints were used. One-thousand one-hundred fifty-nine (85%), 1034 (76%), 827 (61%), 474 (35%), and 71 (5%) proteins were folded to a RMSD from native $<$6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å respectively, which are also much higher than those using *N*/8 correct restraints. The average RMSD of the best cluster centroids in the top five clusters is 4.1 Å, which is 0.9 Å

TABLE 1 Summary of the structure prediction results of 1365 benchmark proteins using simulated restraints

	Predicted restraints		*N*/8 simulated restraints*		*N*/8 simulated restraints and predicted restraints*		*N*/4 simulated restraints*		*N*/4 simulated restraints and predicted restraints*	
	Best RMSD†	Best rank‡	Best RMSD†	Best rank‡	Best RMSD†	Best rank‡	Best RMSD†	Best rank‡	Best RMSD†	Best rank‡
Average of 1365	6.72	2.5	5.87	2.1	5.03	2.4	4.37	2.2	4.11	2.2
RMSD $<$ 6.5§	754 (55.2%)		904 (66.2%)		1023 (74.9%)		1199 (87.8%)		1206 (88.4%)	
RMSD $<$ 6.0§	699 (51.2%)		807 (59.1%)		960 (70.3%)		1132 (82.9%)		1159 (84.9%)	
RMSD $<$ 5.0§	580 (42.5%)		582 (42.6%)		817 (59.9%)		976 (71.5%)		1034 (75.8%)	
RMSD $<$ 4.0§	425 (31.1%)		323 (23.7%)		627 (45.9%)		698 (51.1%)		827 (60.6%)	
RMSD $<$ 3.0§	229 (16.8%)		127 (9.3%)		333 (24.4%)		317 (23.2%)		474 (34.7%)	
RMSD $<$ 2.0§	32 (2.3%)		10 (0.7%)		42 (3.1%)		56 (4.1%)		71 (5.2%)	
RMSD $<$ 1.0§	0 (0.0%)		1 (0.1%)		1 (0.1%)		1 (0.1%)		1 (0.1%)	

*N, number of residues.
†Best RMSD, RMSD of the best (lowest RMSD) cluster centroid.
‡Best rank, rank of the best (lowest RMSD) cluster.
§The number of proteins predicted to various RMSD resolution.
RMSD, coordinate root-mean-square deviation for C_α atoms in Å.

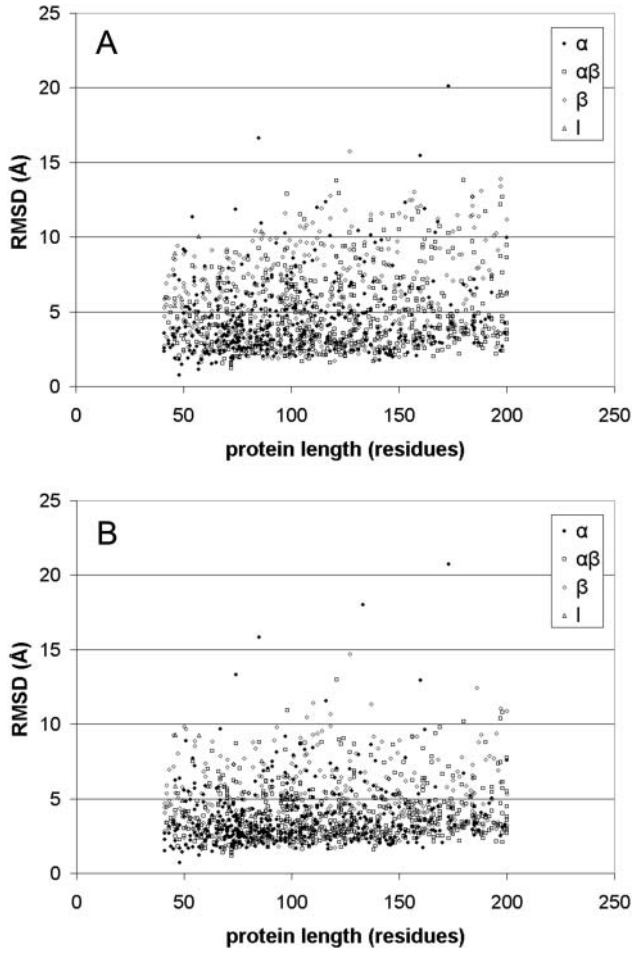

FIGURE 1 RMSD of the best (lowest RMSD) cluster centroid from structure prediction of 1365 benchmark proteins using simulated restraints as a function of protein length for four types of proteins—α-proteins, β-proteins, α/β-proteins, and l protein (little secondary structure). (*A*) *N*/8 simulated restraints; (*B*) *N*/4 simulated restraints. *N* is the number of residues.

lower compared to that using *N*/8 correct restraint. The best cluster on average ranks 2.2, which is lower than that using *N*/8 correct restraints with rank 2.4.

Fig. 1, *A* and *B*, shows the RMSD of the best cluster centroid obtained using *N*/8 and *N*/4 restraints as a function of protein size for four types of proteins in this set—α-proteins, β-proteins, α/β-proteins, and proteins with little secondary structure. The detailed distribution of RMSD for this protein set can be seen clearly. There is no obvious correlation between best RMSD and protein length or type. Good predicted structures can be obtained even for large proteins and difficult β-proteins.

As a control, column two lists the prediction results using only the predicted contact restraints from PROSPECTOR_3. Seven-hundred fifty-four (55%) proteins were folded to a RMSD from native $<$6.5 Å in the top five lowest energy clusters. The average RMSD of the best cluster centroids in the top five clusters using only predicted restraints is 6.7 Å. The best cluster on average ranks 2.5. By comparing with the results when additional *N*/8 (column four) and *N*/4 (column six) correct restraints were used, a general trend can be seen. The more the correct restraints, the more proteins are folded for a given RMSD range, the lower the average RMSD is and the better the best cluster ranks. Fig. 2 shows the RMSD improvement using *N*/8 or *N*/4 correct restraints together with the predicted restraints over the RMSD using only the predicted restraints for each protein. A very strong correlation can be seen, i.e., there is a larger improvement for higher RMSD structures whereas there is a smaller improvement for lower RMSD structures.

Because the predicted restraints generated by the threading algorithm PROSPECTOR_3 are not perfect (on average 46% correct), when the number of correct restraints is large enough predicted restraints should not be used. An important fact we observed here is that the predicted restraints are necessary for better results for the predictions using *N*/8 correct restraints as well as using *N*/4 correct restraints. As can be seen from Table 1, the results using *N*/8 correct restraints without any predicted restraints (column three) are

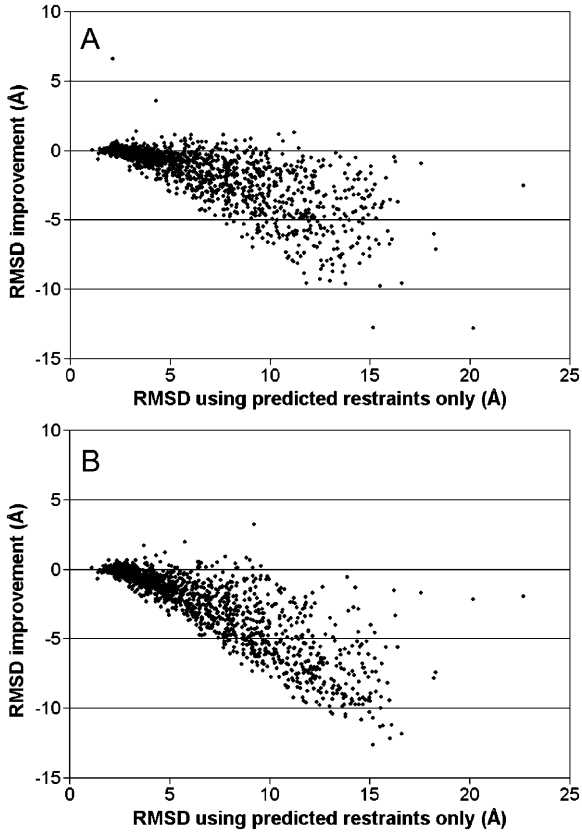

FIGURE 2 RMSD improvement of 1365 benchmark proteins using simulated restraints as a function of RMSD of the best (lowest RMSD) cluster centroid from structure prediction using only predicted restraints. The RMSD improvement is the RMSD difference of the best (lowest RMSD) cluster centroid using both simulated restraints and predicted restraints and the prediction using only predicted restraints. (*A*) *N*/8 simulated restraints; (*B*) *N*/4 simulated restraints. *N* is the number of residues.

TABLE 2 Structure prediction results of 69 proteins using experimental restraints

			Without experimental restraints		*N*/8 experimental restraints*				*N*/4 experimental restraints*			
ID	*N**	Type†	Best RMSD‡	Best rank§	N_{main_chain}¶	N_{methyl}‖	Best RMSD‡	Best rank§	N_{main_chain}¶	N_{methyl}‖	Best RMSD‡	Best rank§
1a1tA	55	α	6.42	3	7	0	6.56	5	14	0	5.79	5
1kmaA	55	α	7.42	5	7	0	7.3	5	14	0	4.63	1
1bno_	87	α	3.71	1	1	10	3.51	2	1	21	2.88	2
1fadA	95	α	3.22	2	1	11	3.12	1	1	23	3.04	1
1f16A	192	α	4.96	2	1	23	3.38	5	1	47	3.67	3
1l3yA	41	β	5.58	5	5	0	4.8	3	10	0	5.13	2
1qdp_	42	β	6.21	3	5	0	6.95	1	11	0	5.1	5
1i2uA	44	β	4.49	1	6	0	3.46	4	11	0	2.28	4
1zaq_	44	β	6.12	3	6	0	6.21	4	11	0	5.84	2
1g9pA	45	β	10.36	1	6	0	9.09	5	11	0	6.01	5
1e8pA	46	β	7.33	1	6	0	3.77	5	12	0	3.81	1
1hx2A	60	β	5.33	3	8	0	5.21	1	15	0	4.12	5
1jgkA	66	β	2.4	1	8	0	2.6	2	17	0	2.8	1
1dx8A	70	β	9.04	4	9	0	6.67	1	18	0	6.88	1
1fgp_	70	β	9.08	1	9	0	6.95	1	18	0	4.61	2
1g47A	70	β	5.53	1	9	0	4.59	4	18	0	4.88	3
1ghj_	79	β	2.06	3	10	0	2	1	20	0	2.14	1
1iyu_	79	β	2.42	3	10	0	2.52	2	20	0	2.42	3
1f53A	84	β	8.04	4	11	0	8.01	5	21	0	5.49	2
1couA	85	β	9.72	1	11	0	7.91	1	21	0	7.27	1
1g4fA	86	β	10.55	1	11	0	6.9	1	22	0	5.62	1
1g6eA	87	β	10.67	5	11	0	5.21	4	22	0	3.71	1
1ewwA	90	β	12.11	5	11	0	8.49	5	23	0	7.7	2
1j8kA	94	β	2.13	1	12	0	2.03	1	24	0	2.1	5
1nct_	98	β	2.62	1	12	0	2.86	1	25	0	2.67	2
2ezm_	101	β	11.64	2	13	0	9.34	4	25	0	6.45	3
1c8pA	102	β	2.78	2	13	0	2.77	5	26	0	2.89	3
1jt8A	102	β	7.51	1	13	0	6.87	4	26	0	6.99	3
1d2bA	126	β	10.32	1	16	0	6.72	1	32	0	4.86	1
1k8hA	133	β	12.72	5	17	0	10.48	4	33	0	6.38	3
1k0sA	151	β	6.57	1	19	0	5.68	2	38	0	4.65	5
1xnaA	151	β	14.66	1	19	0	9.22	1	38	0	5.2	1
1cx1A	153	β	9.27	1	19	0	4.15	4	38	0	4.04	3
1clh_	166	β	3.37	1	21	0	3.25	1	42	0	3.2	5
1cz4A	185	β	4.61	1	23	0	4.5	1	46	0	4.11	5
1jkzA	46	$\alpha\beta$	3.3	1	6	0	3.25	1	12	0	3.7	1
1ncs_	47	$\alpha\beta$	3.84	5	6	0	3.34	4	12	0	2.89	1
1tih_	53	$\alpha\beta$	5.6	1	7	0	5.34	5	13	0	4.66	3
1dax_	64	$\alpha\beta$	2.58	1	8	0	2.5	1	16	0	2.6	1
1g25A	65	$\alpha\beta$	4.15	1	8	0	3.18	1	16	0	3.03	1
1f0zA	66	$\alpha\beta$	4.82	1	8	0	4.04	2	17	0	3.55	1
1ha6A	70	$\alpha\beta$	5.72	5	9	0	5.69	3	18	0	5.2	5
1afi_	72	$\alpha\beta$	1.65	1	9	0	1.58	1	18	0	1.61	1
1bo0_	76	$\alpha\beta$	6.48	4	10	0	5.37	2	17	2	4.52	2
1dcjA	81	$\alpha\beta$	2.72	1	10	0	2.6	1	20	0	2.57	4
1ip9A	85	$\alpha\beta$	5.3	3	11	0	5.02	3	21	0	3.48	3
1khmA	89	$\alpha\beta$	4.07	4	11	0	3.69	3	22	0	4.62	5
1hqi_	90	$\alpha\beta$	6.74	1	11	0	6.17	2	23	0	5.82	1
1mnl_	91	$\alpha\beta$	8.52	2	11	0	6.91	1	23	0	5.75	3
1jh3A	99	$\alpha\beta$	7.03	1	12	0	4.49	2	25	0	4.13	4
1g10A	102	$\alpha\beta$	5.7	2	13	0	4.73	2	26	0	4.46	5
1jrmA	104	$\alpha\beta$	8.86	4	13	0	7.26	1	26	0	5.22	2
1ghtA	105	$\alpha\beta$	6.1	2	13	0	4.26	2	26	0	3.25	5
1eiwA	111	$\alpha\beta$	5.85	1	14	0	3.2	1	24	4	2.62	4
1ji8A	111	$\alpha\beta$	10.82	2	12	2	10.56	3	12	16	4.04	1
1qndA	123	$\alpha\beta$	7.79	5	15	0	5.19	1	31	0	5.12	1
1dc7A	124	$\alpha\beta$	2.64	1	16	0	2.7	1	31	0	2.61	1
1eo1A	124	$\alpha\beta$	10.37	5	16	0	7.87	5	31	0	6.58	1
1hpwA	129	$\alpha\beta$	7.89	5	16	0	7.82	5	32	0	6.75	4

(*Continued*)

TABLE 2 (*Continued*)

			Without experimental restraints		*N*/8 experimental restraints*				*N*/4 experimental restraints*			
ID	*N**	Type†	Best RMSD‡	Best rank§	N_{main_chain}¶	N_{methyl}‖	Best RMSD‡	Best rank§	N_{main_chain}¶	N_{methyl}‖	Best RMSD‡	Best rank§
1mut_	129	$\alpha\beta$	4.35	1	16	0	4.41	1	32	0	4.33	1
1gd5A	130	$\alpha\beta$	4.02	2	16	0	3.9	1	33	0	3.6	1
1tbd_	134	$\alpha\beta$	12.71	1	17	0	10.98	1	34	0	10.31	5
1c05A	159	$\alpha\beta$	12.08	3	20	0	8.56	1	34	6	7.67	1
1bxdA	161	$\alpha\beta$	2.81	2	20	0	3.09	2	40	0	2.77	5
1ao8_	162	$\alpha\beta$	2.77	3	20	0	2.92	4	41	0	2.96	1
1f3yA	165	$\alpha\beta$	4.96	5	21	0	4.97	3	41	0	3.69	3
1ak6_	174	$\alpha\beta$	5.55	4	22	0	4.95	1	44	0	4.26	1
1dgqA	188	$\alpha\beta$	2.47	1	24	0	2.38	1	47	0	2.87	1
1ds9A	198	$\alpha\beta$	7.21	1	25	0	5.26	2	50	0	5.6	1
Average of 69			6.35	2.3			5.24	2.4			4.44	2.5
RMSD < 6.5**			41				47				61	
RMSD < 6.0**			36				45				58	
RMSD < 5.0**			27				36				45	
RMSD < 4.0**			18				25				29	
RMSD < 3.0**			13				12				17	
RMSD < 2.0**			1				1				1	

**N*, number of residues.
†Type, protein secondary structure type.
‡Best RMSD, RMSD of the best (lowest RMSD) cluster centroid.
§Best rank, rank of the best (lowest RMSD) cluster.
¶N_{main_chain}, number of main-chain–main-chain experimental restraints.
‖N_{methyl}, number of side-chain methyl–side-chain methyl experimental restraints;
**The number of proteins predicted to various RMSD resolution.
RMSD, coordinate root-mean-square deviation for C_α atoms in Å.

much worse than the results when predicted restraints are also used (column four). One-hundred nineteen (9%), 153 (11%), 235 (17%), 304 (22%), 206 (15%), and 32 (2%) fewer proteins were folded to a RMSD from native <6.5 Å, 6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. The average RMSD of the best cluster centroid is 0.8 Å larger. This is even true when using *N*/4 correct restraints (column five versus column six), although not as significant as the *N*/8 case. Seven (1%), 27 (2%), 58 (4%), 129 (9%), 157 (12%), and 15 (1%) fewer proteins were folded to a RMSD from native <6.5 Å, 6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. The average RMSD of the best cluster centroids is 0.3 Å larger. As the number of correct restraints increases, the dependence of the results on the predicted restraints becomes weaker. However, the predicted restraints are necessary to use even when there are as many as *N*/4 correct restraints. The predicted restraints contribute to better structural detail to refine structures to the 3.0–4.0-Å RMSD range.

Structure prediction of 69 proteins using experimental restraints

For the 69 proteins with experimental NMR data in the PDB, we generated a set of *N*/8 contact restraints and a set of *N*/4 contact restraints from the complete NOE data. These sets of restraints come from mostly main-chain–main-chain NOE data and sometimes also side-chain methyl–side-chain methyl NOE data. The numbers of main-chain restraints and side-chain methyl restraints used for each protein are listed in Table 2. The overall structure prediction results using these sets of restraints, together with the results using no experimental restraints are also shown in Table 2. Structure prediction results for the 69 proteins using NMR data can be found on our web site at http://www.bioinformatics.buffalo.edu/touchstonex/nmr_folding.

Overall, from these 69 proteins, 41 proteins were folded to a RMSD from native <6.5 Å in the top five lowest energy clusters without using any experimental restraints (Table 2, column two). The average RMSD of the best cluster centroids from the top five clusters for all the proteins is 6.35 Å. On adding *N*/8 experimental restraints (Table 2, column three), the average RMSD of the best cluster centroids from the top five clusters improve to 5.2 Å. Forty-seven proteins were folded to <6.5 Å from native in the top five clusters. 45, 36, 25, 12, and 1 proteins were folded to <6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. When *N*/4 experimental restraints are used (Table 2, column four), the average RMSD further improved to 4.4 Å. Sixty-one proteins were folded to <6.5 Å from native in the top five clusters. Proteins (58, 45, 29, 17, and 1) were folded to <6.0 Å, 5.0 Å, 4.0 Å, 3.0 Å, and 2.0 Å, respectively. Fig. 3 shows the RMSD improvement using *N*/

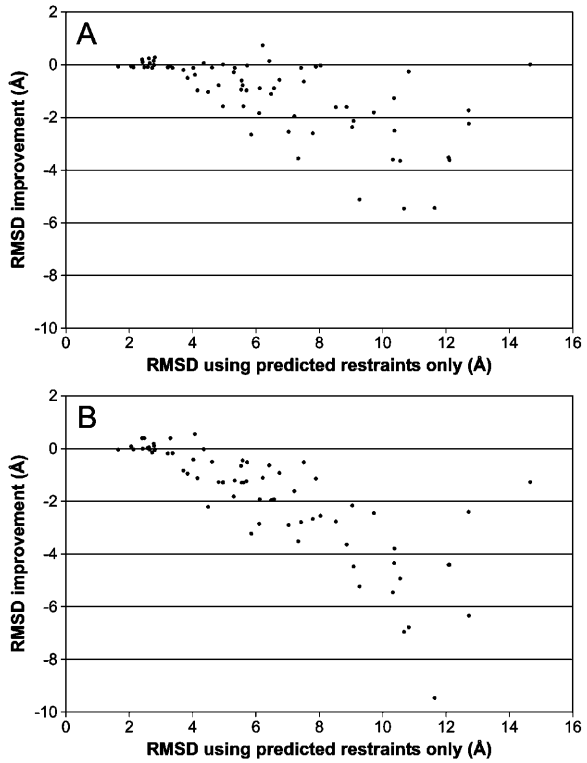

FIGURE 3 RMSD improvement of 69 proteins using experimental restraints as a function of RMSD of the best (lowest RMSD) cluster centroid from structure prediction using only predicted restraints. The RMSD improvement is the RMSD difference of the best (lowest RMSD) cluster centroid using both experimental restraints and predicted restraints and the prediction using only predicted restraints. (*A*) *N*/8 experimental restraints; (*B*) *N*/4 experimental restraints. *N* is the number of residues.

8 or *N*/4 experimental restraints versus the RMSD using no experimental restraints for the 69 proteins. The trend is very similar to Fig. 2 with simulated restraints, i.e., a larger improvement for higher RMSD structures and a smaller improvement for lower RMSD structures.

The structure prediction results shown above using experimental NOE-derived restraints are generally comparable to the results using simulated restraints for these 69 proteins. Using *N*/8 simulated restraints, the average RMSD of the best cluster centroid from the top five lowest energy clusters for the 69 proteins is 5.0 Å, which is only 0.2 Å lower than that using the *N*/8 experimental restraints. Using *N*/4 simulated restraints, the average RMSD for the 69 proteins is 4.2 Å, which is also 0.2 Å lower than that using *N*/4 experimental restraints. A major reason for the slightly worse results with experimental restraints is the type of restraints used. For most proteins, the experimental restraints are mainly main-chain restraints. The main-chain restraints are mainly observed in β-sheet structures, and thus are important for β-sheet structure prediction. However, they usually are not observed often in helical structures. Therefore, sometimes the helical part of the structure cannot be predicted well. For example, the protein 1c05A was folded to 3.5 Å using *N*/4 simulated restraints and was not foldable using *N*/4 experimental restraints. 1c05A is a 159-residue elongated RNA binding protein consisting of two distinct subdomains; one is all helical and the other includes a β-sheet (Markus et al., 1999). In our predicted model of 1c05A using *N*/4 experimental NOE-derived restraints, although the RMSD of the whole protein is 7.7 Å, the RMSD of the β-sheet-containing subdomain is low (3.2 Å). The high overall RMSD mainly comes from the helical subdomain, which has a RMSD of 8.9 Å. The experimental restraints used consist of 34 main-chain restraints and only six side-chain methyl restraints. The helical subdomain only has five restraints, which are not enough to define a reasonably good structure. Another example is protein 1tbd_, it was folded to 4.4 Å using *N*/4 simulated restraints and was not foldable using *N*/4 experimental restraints. This protein is a 134-residue $\alpha\beta$-sandwich-type DNA-binding protein with a central five-stranded antiparallel β-sheet flanked by two helices on both sides of the β-sheet (Luo et al., 1996). The β-sheet was predicted fairly well, but the arrangement of the helices was predicted incorrectly. The experimental restraints consist of 34 main-chain restraints and no side-chain methyl restraints. Contrary to the situation when experimental restraints are used, the simulated restraints, however, do not have this problem.

CONCLUSIONS

We have tested the sparse distance restraint-assisted structure prediction algorithm, TOUCHSTONEX, on a large, representative PDB benchmark set of 1365 proteins. Using *N*/8 simulated correct long-range contact restraints, where *N* is the number of residues, 1023 (75%) proteins were folded to a RMSD from native $<$6.5 Å in the top five lowest energy clusters. Of those, 627 (46%) proteins were folded to a RMSD from native $<$4.0 Å. The average RMSD of the lowest RMSD cluster centroid structures in the top five lowest energy clusters for all 1365 proteins is 5.0 Å. When the number of the correct restraints was increased to *N*/4, 1206 (88%) proteins were folded to a RMSD from native $<$6.5 Å in the top five lowest energy clusters, 827 (61%) were folded to a RMSD $<$4.0 Å. The average RMSD of the lowest RMSD structures was further improved to 4.1 Å. These results show significant improvement compared to the prediction without using any experimental restraints, where 754 (55%) proteins were folded to $<$6.5 Å from native and the average RMSD is 6.7 Å. However, the results also show that the predicted restraints play an important role even when the number of correct restraints is as large as *N*/4.

We further tested TOUCHSTONEX by predicting structures for 69 proteins with experimental NMR data from the PDB. Using *N*/8 long-range experimental restraints, 47 proteins were folded to a RMSD from native $<$6.5 Å in the top five lowest energy clusters. Proteins (45, 36, 25, and 12) were folded to a RMSD from native $<$6.0 Å, 5.0 Å, 4.0 Å,

and 3.0 Å, respectively. The average RMSD of the lowest RMSD cluster centroids in the top five lowest energy clusters is 5.2 Å. When $N/4$ experimental restraints were used, 61 proteins were folded to a RMSD from native <6.5 Å. Proteins (58, 45, 29, and 17) were folded to a RMSD from native <6.0 Å, 5.0 Å, 4.0 Å, and 3.0 Å, respectively. The average RMSD is 4.4 Å. For these 69 proteins, the results using experimental restraints are generally comparable to the results using simulated restraints.

To summarize, the results shown in this article provide an objective and realistic evaluation of TOUCHSTONEX. The algorithm proved to be an efficient method to predict protein structures of medium-to-low resolution using sparse distance restraints, such as but not limited to NOE data from NMR experiments. The resulting medium-to-low resolution structures can be used directly for structural and functional analyses, or they can serve as an initial model for further refinement. Because the 1365-protein test set is comprehensive and representative for the whole PDB for structures up to 200 residues in length, and experimental NOE-derived as well as simulated restraints were used, it is expected that the algorithm will perform comparably well in real-life application. We hope that the algorithm can be an alternative and complimentary tool for NMR-based structure determination in the early stage when only limited NOE data are available, and thus contribute to the acceleration of structural genomics projects. We also hope that the algorithm will be applied to other experimental methods that can provide tertiary restraint information. At the same time, there are continuing efforts in our group to improve the protein-folding algorithm. Using more advanced algorithms, the results of protein structure prediction with sparse NMR restraints are expected to improve.

This research was supported by NIH grant GM-37408 of the Division of General Medical Sciences of the National Institutes of Health.

REFERENCES

Aszodi, A., M. J. Gradwell, and W. R. Taylor. 1995. Global fold determination from a small number of distance restraints. *J. Mol. Biol.* 251:308–326.

Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science*. 294:93–96.

Betancourt, M., and J. Skolnick. 2001. Finding the needle in a haystack: educing protein native folds from ambiguous ab initio folding predictions. *J. Comput. Chem.* 22:339–353.

Bowers, P. M., C. E. Strauss, and D. Baker. 2000. De novo protein structure determination using sparse NMR data. *J. Biomol. NMR*. 18:311–318.

Connolly, P. C., A. S. Stern, and J. C. Hoch. 1994. Estimating protein fold from incomplete and approximate NMR data. *J. Am. Chem. Soc.* 116:2675–2676.

Kihara, D., H. Lu, A. Kolinski, and J. Skolnick. 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA*. 98:10125–10130.

Kolinski, A., and J. Skolnick. 1998. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. *Proteins*. 32:475–494.

Li, W., Y. Zhang, D. Kihara, Y. J. Huang, D. Zheng, G. T. Montelione, A. Kolinski, and J. Skolnick. 2003. TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins Struct. Funct. Genet.* 53:290–306.

Liwo, A., J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga. 1999. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA*. 96:5482–5485.

Lougheed, J. C., J. M. Holton, T. Alber, J. F. Bazan, and T. M. Handel. 2001. Structure of melanoma inhibitory activity protein, a member of a recently identified family of secreted proteins. *Proc. Natl. Acad. Sci. USA*. 98:5515–5520.

Luo, X., D. G. Sanford, P. A. Bullock, and W. W. Bachovchin. 1996. Solution structure of the origin DNA-binding domain of SV40 T-antigen. *Nat. Struct. Biol.* 3:1034–1039.

Markus, M. A., R. B. Gerstner, D. E. Draper, and D. A. Torchia. 1999. Refining the overall structure and subdomain orientation of ribosomal protein S4 delta41 with dipolar couplings measured by NMR in uniaxial liquid crystalline phases. *J. Mol. Biol.* 292:375–387.

Simons, K. T., C. Strauss, and D. Baker. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* 306:1191–1199.

Skolnick, J., J. S. Fetrow, and A. Kolinski. 2000. Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* 18:283–287.

Skolnick, J., D. Kihara, and Y. Zhang. 2004. Development and testing of the PROSPECTOR 3.0 threading algorithm. *Proteins*. 56:502–518.

Skolnick, J., A. Kolinski, and A. R. Ortiz. 1997. MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* 265:217–241.

Smith-Brown, M. J., D. Kominos, and R. M. Levy. 1993. Global folding of proteins using a limited number of distance constraints. *Protein Eng.* 6:605–614.

Tashiro, M., R. Tejero, D. E. Zimmerman, B. Celda, B. Nilsson, and G. T. Montelione. 1997. High-resolution solution NMR structure of the Z domain of staphylococcal protein A. *J. Mol. Biol.* 272:573–590.

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

Zhang, Y., A. Kolinski, and J. Skolnick. 2003. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophys. J.* 85:1145–1164.

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

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