PROTEINS: Structure, Function, and Genetics 53:290–306 (2003)

TOUCHSTONEX: Protein Structure Prediction With Sparse NMR Data

Wei Li,[1] Yang Zhang,[1] Daisuke Kihara,[1] Yuanpeng Janet Huang,[2] Deyou Zheng,[2] Gaetano T. Montelione,[2] Andrzej Kolinski,[1,3] and Jeffrey Skolnick[1]*

[1]*Center of Excellence in Bioinformatics, University at Buffalo, Buffalo, New York*
[2]*Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey*
[3]*Faculty of Chemistry, Warsaw University, Warsaw, Poland*

ABSTRACT ***TOUCHSTONEX*, a new method for folding proteins that uses a small number of long-range contact restraints derived from NMR experimental NOE (nuclear Overhauser enhancement) data, is described. The method employs a new lattice-based, reduced model of proteins that explicitly represents C_α, C_β, and the sidechain centers of mass. The force field consists of knowledge-based terms to produce protein-like behavior, including various short-range interactions, hydrogen bonding, and one-body, pairwise, and multibody long-range interactions. Contact restraints were incorporated into the force field as an NOE-specific pairwise potential. We evaluated the algorithm using a set of 125 proteins of various secondary structure types and lengths up to 174 residues. Using *N*/8 simulated, long-range sidechain contact restraints, where *N* is the number of residues, 108 proteins were folded to a C_α-root-mean-square deviation (RMSD) from native below 6.5 Å. The average RMSD of the lowest RMSD structures for all 125 proteins (folded and unfolded) was 4.4 Å. The algorithm was also applied to limited experimental NOE data generated for three proteins. Using very few experimental sidechain contact restraints, and a small number of sidechain–main chain and main chain–main chain contact restraints, we folded all three proteins to low-to-medium resolution structures. The algorithm can be applied to the NMR structure determination process or other experimental methods that can provide tertiary restraint information, especially in the early stage of structure determination, when only limited data are available. Proteins 2003; 53:290–306.** © 2003 Wiley-Liss, Inc.

Key words: contact restraints; NOE; protein folding; protein structure prediction; sparse NMR data

INTRODUCTION

In this postgenomic era, the elucidation of the three-dimensional (3D) structures of proteins from their sequences is of great importance. At present, the NMR solution structure determination of proteins comprises a small portion of this process. Traditional NMR structure determination methods[1,2] require a large number of nuclear Overhauser enhancement (NOE) restraints—typically 15–20 NOE restraints per residue—to obtain a high-resolution structure (equivalent to about a 2 Å X-ray structure).[3] Yet, in reality, such a large number of restraints is hard to obtain because of line broadening and spectral crowding, especially for large proteins.[3] Although tremendous advances in both NMR hardware and software have taken place during the past decade, this problem still limits the range of utilization of NMR as a tool for protein structure determination. As an alternative and complementary approach, protein structure prediction with a limited number of distance restraints holds great promise.[4–12] Research has indicated that in ab initio protein structure prediction, a limited number of distance restraints can be sufficient to guide folding to a correct structure that would otherwise be difficult to predict.[9] Such a small number of distance restraints is relatively easy to obtain from NOE data, even in the early stage of NMR-based structure determination. The resulting low-to-medium resolution structures can be used directly for structural and functional analysis.[13,14] Alternatively, they can serve as an initial model for resonance and constraint assignments, which would greatly simplify the overall procedure. Thus, protein structure prediction with sparse NMR data should speed up the process of protein structure determination.

Among the studies of protein folding that use a small number of distance restraints, Smith-Brown et al.[4] have modeled a protein as a chain of glycine residues using a substantial number of tertiary restraints. Connolly et al.[5] used an off-lattice reduced representation of proteins to estimate fold from incomplete and approximate NOE distance data (0.5 long-range restraints per residue). Aszodi et al.[6] used a distance-geometry–based method to assemble protein structures, with experimental tertiary distance restraints supplemented by predicted interresi-

Grant sponsor: NIH; Grant numbers: GM-37408 (to Jeffrey Skolnick) and P50-GM62413 (to Gaetano T. Montelione) of the Division of General Medical Sciences of the National Institutes of Health.

*Correspondence to: Jeffrey Skolnick, UB Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Ste. 300, Buffalo, NY 14203. E-mail: skolnick@buffalo.edu

Received 28 January 2003; Accepted 2 May 2003

due distance restraints extracted from multiple-sequence alignment. 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) below 5 Å from native. Skolnick and Kolinski have used a high-coordination lattice model of protein structure and a knowledge-based force field to fold proteins.[8,9] Nine proteins up to 247 residues in length could be folded to moderate resolution with as few as $N/7$ long-range restraints and some knowledge of the secondary structure. Bowers et al.[12] selected peptide fragments from proteins of known structure based on sequence similarity and consistency with the chemical shift and NOE data, then assembled proteins to high resolution using ~1 NOE restraint per residue.

In this article, we describe the folding of proteins with sparse, long-range contact restraints using an extended version of our protein structure prediction algorithm, *TOUCHSTONE*, named here *TOUCHSTONEX*. Different from our previous algorithms, our new lattice-based, reduced model of proteins explicitly includes the C_α, C_β, and sidechain centers of mass (CABS). The force field includes knowledge-based terms to produce protein-like behavior, including various short-range interactions, hydrogen bonding, and one-body, pairwise, and multibody long-range interactions. Correct contact restraints were incorporated into the force field as an NOE-specific pairwise interaction. The algorithm uses predicted instead of known secondary structure, as was the case with our previous algorithms.[8,9] Here, we also apply our algorithm to the folding of 125 proteins using $N/8$ randomly generated, exact, long-range sidechain contact restraints. For 65 of those proteins, $N/12$ and $N/4$ restraints were also used to evaluate the performance of the algorithm as the number of restraints varied. We also applied this algorithm to three proteins with experimental NOE data to evaluate its performance in real-life applications. For these three proteins, sidechain contact restraints, as well as sidechain–main chain and main chain–main chain contact restraints derived from NOEs, were used.

MATERIALS AND METHODS

Protein Model and Force Field

A newly developed, lattice-based, reduced model of proteins, the CABS model, is used. A detailed description of the model may be found in a separate article.[15] Briefly, each amino acid is represented by up to three united atom groups, namely, the C_α, C_β, and sidechain center of mass. For computational simplicity, the main chain C_α atoms are restricted to a 3D underlying cubic lattice system, with a lattice spacing of 0.87 Å, and 312 allowed bond vectors. The C_α–C_α virtual bond length fluctuates from 3.26 Å to 4.35 Å. The virtual C_α–C_α bond angle is restricted to the experimental range of 65–165°. The positions of three subsequent C_α atoms define the local coordinate system used for the determination of the remaining two interaction centers—the C_β (except glycine) and the center of mass of the sidechain heavy atoms (except glycine and alanine). The parameters for the determination of C_β and sidechain center of mass are extracted from the Protein Data Bank (PDB).[16] A two-rotamer approximation is assumed, depending on the expanded (e.g., sheet) or compact (e.g., helix) main chain conformation. Any protein structure can be fitted to the corresponding lattice model, with an average accuracy of 0.45 Å.

The force field consists of a variety of terms based on the regularities seen in protein structures. They contain both generic (sequence-independent) interactions to produce protein-like structures and sequence-dependent interactions derived from sequence analysis and multiple-sequence alignments. Interactions are divided into two categories: the short-range, secondary structure type of interactions and long-range, tertiary interactions. Explicitly, the short-range interactions include short-range C_α–C_α and sidechain–sidechain correlations (between residue i and residues $i + 2$, $i + 3$, $i + 4$, and $i + 5$), and local protein-like conformational stiffness with a bias toward predicted secondary structure. Among the long-range interactions are long-range orientation-dependent pairwise interactions that include both generic and sequence-dependent terms. The one-body center-symmetric burial interactions represent the general propensity of amino acids to be buried inside the protein or exposed to solvent. The environmental profile describes the contact environment of amino acids. Debye–Huckel electrostatic interactions are allowed among the four charged amino acids (Asp^-, Glu^-, Lys^+, and Arg^+). The explicit cooperative hydrogen-bond interactions can be short- or long-range, depending on the involved secondary structure. Finally, the contact order and contact number terms enforce biases toward the expected contact order (the number of residues along the chain between contact residues)[17] and expected length-dependent contact number. A detailed description of the force field can be found in other publications.[8,9,15,18,19]

In addition to the aforementioned terms, the scoring function also has specific penalty terms to incorporate predicted restraints from our threading algorithm *PROSPECTOR*.[20] These restraints include predicted local C_α distance restraints (less than 6 residues along the sequence) and predicted nonsequential sidechain contact restraints. See articles by Kolinski et al.,[21] Kihara et al.,[22] and Zhang et al.[15] for a detailed description of these terms.

The overall scoring function is the combination of all of the energy terms described above. The relative weights of these nonindependent energy terms are determined by maximization of the correlation of energy and the decoy–native similarity.[15]

Because the force field imposes strong conformational biases to predicted secondary structure, reasonably high-accuracy secondary structure prediction is extremely important for the success of tertiary structure prediction. According to our test results from the comparison of three contemporary secondary structure prediction algorithms, PHD,[23] SAM-T99,[24] and PSIPRED,[25] the "overlap" of PSIPRED and SAM-T99 prediction with a cutoff equal to 0.5 has the highest accuracy[15] and is therefore our method of choice.

Generation of Long-Range NOE-Like Contact Restraints

Contact restraints are either simulated or extracted from experimental NOE data. The simulated restraints are randomly selected from sidechain contacts in the native protein structure. Two sidechains that have at least one pair of their heavy atoms within 4.5 Å are considered to be in contact. These simulated restraints are also termed "exact restraints" in contrast to the predicted restraints mentioned earlier. For NMR experiments, the proton NOE data are first divided into three groups: between sidechain and sidechain atoms, between sidechain and main chain atoms (H_α, H_N), and between main chain and main chain atoms. The atomic-level NOE data are then converted into sidechain contact restraints, sidechain–main chain contact restraints, and main chain–main chain contact restraints between residues. For both simulated and NOE-derived contact restraints, only contact partners at least 5 residues apart along the protein chain are considered.

Implementation of NOE-Specific Pairwise Interactions

Long-range contact restraints are incorporated into the scoring function as NOE-specific penalties. In the case of sidechain contact restraints, the penalties are as follows:

$$E_{NOE} = \begin{cases} +\varepsilon + \varepsilon_a \cdot (d - 6.3) & \text{for } 6.3 < d < 7 \text{ (in lattice units)} \\ +2\varepsilon + \varepsilon_a \cdot (d - 6.3) & \text{for } d > 7 \text{ (in lattice units)} \end{cases} \quad (1)$$

Two terms in Eq. (1) hierarchically penalize violation. The first term penalizes the violation of distances between sidechain centers (d) with penalty (ϵ). The second term further penalizes the violation of d according to the extent of the violation scaled by ϵ_a. The algorithm allows a small amount of violation. Only the part of E_{NOE} that exceeds a threshold ($\epsilon_{threshold}$) enters into the scoring function. $\epsilon_{threshold}$ is set numerically equal to the number of restraints. This enables the protein to undergo large conformational changes and jump out of local minimum during the simulation through a partial violation of the restraints. The value of $\epsilon_a = 0.5\,\epsilon$, with $\epsilon = 8$. This weight is 8 times the weight of the predicted contact restraints.

In the case of sidechain–main chain and main chain–main chain contact restraints, similar penalties are used:

$$E_{NOE} = +2\varepsilon + \varepsilon_a \cdot (d - 7) \quad \text{for } d > 7 \text{ (in lattice units)} \quad (2)$$

Here, d is the distance between sidechain center and C_α for sidechain–main chain restraints, and between C_α atoms for main chain–main chain restraints. The distance cutoff is set to 7 in lattice units instead of 6.3. In the second term of Eq. (2), a small amount of violation is also allowed. The threshold ($\epsilon_{threshold}$) is set numerically equal to the number of restraints. ϵ and ϵ_a are scaled the same as for the sidechain restraints.

Another kind of restraint involves the main chain hydrogens. Backbone hydrogen bonding can provide fully independent and supporting evidence for regular secondary structure.[26] Long-range (at least 5 residues apart), regular-pattern backbone hydrogen bonding comes exclusively from β-sheets. The pattern of hydrogen bonding can be used to differentiate parallel and antiparallel β-sheets. In our algorithm, these main chain hydrogen-bonding restraints are incorporated as follows: If the distance d between the two C_α atoms is less than 6.7 in lattice units, and the backbone orientation associated with the C_α atoms is appropriate for antiparallel and parallel β-sheets, respectively, then

$$E_{NOE} = -2\varepsilon \quad (3)$$

To judge the backbone orientation, we define two vectors on each C_α. One is the cross product of the two successive C_α–C_α virtual bond vectors connected to the C_α atom; the other is the bisector of them. For parallel β-sheets, the angle between the two bisector vectors associated with the two C_α atoms is within −60–60°, and the angle between the two cross product vectors is within −65–65°. For antiparallel β-sheets, the angle between the two bisector vectors is within −60–60°, and the angle between the two cross product vectors lies within 115–245°. ϵ is scaled the same as for the sidechain restraints.

Monte Carlo Sampling Scheme and Folding Protocol

The conformational sampling scheme uses a replica exchange Monte Carlo (REMC) method.[27,28] Initial, arbitrary random coil conformations of 40 replicas are created and assigned different temperatures between an initial maximum and an initial minimum. The initial maximum temperature is equal to the final maximum temperature. The initial minimum temperature is scaled according to the final minimum temperature. The temperature difference between the initial minimum and maximum is one third of the temperature difference between the final minimum and maximum. The final maximum and minimum temperatures are scaled according to protein length,[15] the number of the restraints, and weight of the restraints; they are proportional to the 0.7 power of the product of the number and the weight of the restraints. The final maximum temperature ranges from 17 to 110 (in dimensionless units). The final minimum temperature ranges from 1.1 to 1.8 (in dimensionless units). A conformational update of the protein lattice chain includes local 2–6 bond moves, a small displacement of a larger portion of the chain, and chain end moves.[15,18,19] Global conformational swaps also occur between replicas at different temperatures. Between two consecutive global swaps, there are $200 \times N$ (with N being the number of residues) local movement attempts. The simulation starts at high temperature; then, the minimum temperature is gradually lowered in 20 steps to the final minimum temperature during the first 400 global swaps. Finally, a 5 times longer (2000 global swaps) isothermal run at the lowest temperature is performed. A total of 2400 global swaps are performed for each replica.

Usually, the calculation takes 24–48 h of central processing unit (CPU) time on a 1.26-GHz Pentium III processor for proteins less than 200 residues.

Clustering and Computing the Average Structures

About 16,000 structures selected from the 8 lowest temperature replicas are clustered.[29] The clusters are ranked according to the average energy of the structures in the cluster. For each cluster, a centroid is determined by optimally aligning the structures and computing their average. The centroids from all of the clusters are then compared with the native protein structure, and their C_α coordinate RMSDs from native are calculated. If there is at least one cluster centroid with a global RMSD to native less than 6.5 Å, the protein is considered successfully folded. A more practical criterion involves only the top 5 lowest energy clusters instead of all clusters. If within the top 5 lowest energy clusters, there is at least 1 cluster centroid with a RMSD to native less than 6.5 Å, then the protein is considered to have been successfully folded. Both criteria are used. The lowest RMSD cluster centroid (from native) is considered to be the best structure.

Sets of Proteins

Two test sets were examined. The first, set I, consisted of 65 proteins of different types (4 small proteins with little secondary structure, 21 α-proteins, 20 β-proteins, 20 α/β-proteins) and lengths (ranging from 39 to 146 residues), and was the same as that previously used.[22] The second data set, set II, consisted of 60 proteins carefully selected from the PDB for benchmark purposes, including 20 α-proteins whose lengths ranged from 36 to 156 residues, 20 β-proteins whose lengths ranged from 36 to 153 residues, and 20 α/β-proteins whose lengths ranged from 64 to 174 residues.[15] This set consisted of more large proteins than the first set. For both sets, simulated sidechain contact restraints were used.

Recent advances in software development for interpreting NMR data make it possible to acquire useful structural constraints efficiently in a remarkably short time. For instance, by using the programs AutoAssign[30] and AutoStructure,[31–33] backbone resonance assignments and initial structural constraints can be generated within 1 or 2 days with the use of good-quality NMR spectral peak lists.[34,35] To explore the value of these NMR data in structure prediction, we tested 3 sets of experimental contact restraints that can generally be derived from NMR spectra in the initial stage of NMR structure determination, including experimental NMR data obtained for the Z-domain of staphylococcal protein A (58 residues), the C-terminal BRCA-1–like domain from *T. thermophilus* DNA ligase BRCT (92 residues), and the human melanoma inhibitory activity (MIA) protein (108 residues). The NMR data of protein MIA were kindly provided by the NMR group at GeneFormatics, Inc. (San Diego, CA).

Generation of Experimental Distance and Hydrogen-Bonding Constraints for Z-Domain and BRCT

The program AutoAssign[30] was used for automated analysis of backbone resonance assignments of Z-domain and BRCT. The input for AutoAssign analysis of BRCT included peak lists from 2D ^{1}H-^{15}N HSQC and 3D CBCANH, CBCAcoNH, HNCA, haCAcoNH, HAcaNH, HAcacoNH, and HNCO experiments, recorded as described elsewhere.[36,37] Results obtained from the automated analysis were extended and, in some cases, corrected by manual analysis of these data together with 3D hCCcoNH-TOCSY, HcccoNH-TOCSY, and HCCH-COSY experiments. We made sidechain aromatic, guanido, and amide ^{1}H, ^{13}C and ^{15}N resonance assignments using homonuclear 2D TOCSY, 2D ^{1}H-^{13}C HSQC, and ^{13}C-edited NOESY data. For Z-domain, data collection and analysis were carried out as described in an article by Zheng et al.,[35] with a uniformly ^{13}C,^{15}N,^{2}H-enriched sample with ^{1}H-^{13}C–labeled methyl groups. The following ^{2}H-decoupled triple resonance experiments were used as input for AutoAssign in determining backbone ^{15}N, H^N, $^{13}C^\alpha$, $^{13}C'$, sidechain H^N, and methyl sidechain ^{1}H-^{13}C resonance assignments: HNCO, HNcaCO, HNCA, HNcoCA, HNCACB, HNcoCACB, HcccoNH-TOCSY, and hCcccoNH-TOCSY.

The initial analysis of NOESY peak lists and generation of conformational constraints for the Z-domain and BRCT were carried out in a fully automated manner with the program AutoStructure.[31–33] AutoStructure is a rule-based, expert system that automatically determines protein structures from NMR data. It generates a reliable initial protein fold using intelligent analysis methods based on spectrum-specific properties and the identification of self-consistent NOE contact patterns, without the use of any 3D structure model. In particular, the software identifies secondary structures, including alignments and hydrogen-bonding constraints between β-strands, based on a combined pattern analysis of secondary structure-specific NOE contacts, chemical shift, scalar coupling constant, and slow amide proton exchange data. The experimental NMR input data for Z-domain and BRCT used for AutoStructure analysis included the resonance assignment lists, NOESY peak lists derived from the 3D ^{15}N- and ^{13}C–edited NOESY data, $^3J(H^N\text{-}H^\alpha)$ scalar coupling constants (not for Z domain), and slow amide ^{1}H/^{2}H exchange data. Coupling constants $^3J(H^N\text{-}H^\alpha)$ were obtained from 2D HSQC-J spectra.[38] We determined amide hydrogen-exchange rates by lyophilizing the protein from H_2O, dissolving the protein in D_2O, and acquiring a series of 2D ^{1}H-^{15}N HSQC spectra at 20 min, 45 min, 70 min, 100 min, and 100 h. A small subset of these distance and hydrogen-bond constraints were then selected to generate long-range, NOE-like contact restraints for structure predictions by *TOUCHSTONEX*, with use of the criteria described above.

TABLE I. Prediction Results for the 65 Proteins in Set I With *N*/8-Simulated Exact Long-Range Sidechain Contact Restraints

ID	N[a]	N_{exact}[b]	$N_{pred\ cont}$[c]	$N_{pred\ dist}$[d]	Predicted Only[e]	Exact Only[f]	Exact + Predicted[g]	$N_{exact\ sat}$[h]	$N_{pred\ sat}$[i]
Small									
1ixa	39	5	64	15	5.7 (1/9)	5.3 (2/15)	4.3 (7/9)	5	59
1fc2C	44	5	19	37	3.2 (1/5)	2.9 (1/6)	2.8 (1/7)	5	15
6pti	57	7	92	56	4.8 (1/7)	6.7 (3/7)	4.6 (1/7)	7	79
1rpo	61	8	23	560	1.4 (1/4)	2.2 (1/6)	2.0 (1/6)	8	12
α									
1bw6A	56	7	80	100	4.7 (1/9)	4.1 (1/5)	4.1 (1/6)	5	77
2ezh	65	8	21	147	3.0 (1/5)	2.5 (1/6)	2.7 (1/10)	8	13
1c5a	66	8	26	177	4.9 (1/3)	3.8 (1/7)	3.8 (1/7)	8	16
1hp8	68	9	15	193	4.0 (2/6)	2.4 (1/10)	2.3 (1/7)	9	12
2bby	69	9	104	55	4.2 (1/7)	2.4 (1/8)	3.3 (1/7)	9	84
1ftz	70	9	91	180	2.0 (1/9)	3.2 (1/4)	2.0 (1/4)	8	83
1pou	71	9	102	91	3.6 (1/7)	2.8 (1/8)	2.4 (1/7)	9	75
1lea	72	9	136	153	4.2 (1/8)	3.7 (1/7)	4.1 (1/8)	9	120
1kjs	74	9	97	202	8.4 (4/4)	3.4 (1/11)	3.8 (1/7)	9	42
1ner	74	9	85	48	3.0 (2/7)	4.9 (3/7)	2.7 (2/4)	8	60
1aoy	78	10	46	145	5.8 (1/4)	6.5 (1/4)	6.2 (1/6)	10	7
1nkl	78	10	12	213	3.8 (1/7)	1.3 (1/11)	1.3 (1/10)	10	41
1a32	85	11	44	401	8.8 (2/2)	3.2 (3/3)	2.9 (1/3)	11	26
1ngr	85	11	165	109	2.1 (1/10)	2.1 (1/11)	1.7 (1/9)	11	135
2af8	86	11	122	94	4.1 (1/9)	4.4 (1/6)	4.4 (1/7)	10	92
2ezk	93	12	28	130	9.2 (1/6)	5.2 (2/3)	5.1 (1/2)	12	14
21fb	100	13	78	130	9.0 (5/5)	3.9 (1/5)	4.2 (5/7)	12	52
256bA	106	13	133	390	3.3 (1/9)	3.0 (1/9)	2.9 (1/8)	13	123
1hmdA	113	14	50	366	4.4 (3/7)	2.8 (1/6)	2.7 (2/5)	14	29
1hlb	138	17	31	534	5.4 (5/8)	2.9 (1/8)	2.6 (1/7)	16	20
1mba	146	18	266	822	3.5 (1/7)	2.9 (1/9)	2.6 (1/7)	17	224
β									
1tfi	50	6	67	9	8.2 (19/21)	5.9 (1/3)	7.5 (12/14)	3	36
1bq9A	53	7	48	9	6.8 (2/11)	6.5 (7/11)	5.5 (3/13)	7	43
1nxb	53	7	90	40	4.9 (2/4)	7.3 (1/10)	4.4 (2/4)	6	85
1shg	57	7	93	56	7.4 (3/6)	6.3 (2/6)	7.2 (1/3)	7	79
1vif	60	8	58	38	5.5 (3/14)	3.1 (1/6)	3.5 (2/4)	8	35
1fas	61	8	157	73	3.7 (1/6)	4.9 (1/9)	3.6 (1/8)	8	139
1csp	64	8	88	51	3.5 (1/8)	4.8 (2/8)	3.1 (1/7)	7	72
1sro	66	8	74	87	6.7 (1/5)	4.1 (2/7)	5.8 (1/4)	8	56
1pse	69	9	52	27	6.9 (3/10)	8.4 (1/6)	5.9 (1/8)	9	48
1ah9	71	9	113	41	3.6 (1/5)	4.7 (1/8)	3.4 (1/6)	7	94
1iyv	79	10	124	40	4.4 (1/5)	5.3 (4/7)	3.7 (1/12)	10	106
1rip	81	10	106	131	4.7 (1/8)	5.0 (4/4)	4.4 (1/4)	9	92
1tit	89	11	243	206	1.7 (1/8)	2.2 (1/3)	1.5 (1/9)	11	209
1wiu	93	12	279	123	3.2 (1/5)	5.3 (1/6)	2.8 (1/6)	12	233
2pcy	99	12	297	122	4.0 (1/7)	4.7 (3/11)	3.1 (1/8)	11	237
1ksr	100	13	86	51	7.1 (3/11)	4.7 (1/3)	4.5 (1/7)	13	69
1tlk	103	13	224	191	2.9 (1/9)	4.4 (1/5)	2.5 (1/8)	13	195
1thx	108	14	193	136	2.3 (1/8)	3.0 (1/8)	2.6 (1/5)	14	164
4fgf	121	15	68	71	7.6 (1/12)	7.3 (1/13)	4.2 (1/8)	15	56
2azaA	129	16	241	102	5.4 (1/10)	4.9 (1/9)	3.7 (1/8)	15	184
α/β									
1gpt	47	6	82	45	3.6 (1/7)	5.2 (1/9)	3.2 (1/11)	6	71
2fdn	55	7	146	5	2.3 (1/9)	2.5 (1/15)	1.8 (1/8)	7	133
1pgx	56	7	43	43	5.0 (3/8)	3.0 (2/6)	4.3 (1/6)	6	32
2ptl	60	8	21	59	2.9 (1/7)	1.9 (1/8)	1.9 (1/10)	8	14

TABLE I. (Continued)

ID	N[a]	N_{exact}[b]	$N_{pred\ cont}$[c]	$N_{pred\ dist}$[d]	Predicted Only[e]	Exact Only[f]	Exact + Predicted[g]	$N_{exact\ sat}$[h]	$N_{pred\ sat}$[i]
2fmr	65	8	106	89	4.6 (1/7)	6.2 (1/9)	4.6 (1/11)	6	92
1cis	66	8	37	71	5.8 (1/6)	4.8 (1/7)	4.2 (1/10)	8	30
1ctf	68	9	108	51	6.0 (2/5)	3.1 (1/4)	4.4 (3/4)	9	68
1stu	68	9	59	172	3.6 (1/3)	6.7 (2/3)	4.3 (1/5)	8	49
1ubi	76	10	123	112	4.2 (1/9)	2.7 (1/6)	3.1 (1/11)	10	97
1vcc	77	10	30	48	6.2 (2/20)	2.5 (1/7)	3.2 (1/4)	9	22
1poh	85	11	138	120	9.8 (2/3)	3.0 (1/5)	4.4 (2/4)	11	91
1ife	91	11	151	180	7.9 (6/6)	2.8 (1/6)	6.0 (2/4)	11	79
2sarA	96	12	116	24	8.0 (1/4)	6.0 (2/8)	4.4 (1/5)	12	91
1stfI	98	12	25	74	7.8 (1/13)	7.4 (1/5)	6.7 (1/6)	12	15
1tsg	98	12	82	110	7.9 (2/10)	10.6 (1/6)	8.3 (2/6)	12	62
1shaA	103	13	73	80	8.9 (2/10)	5.4 (2/7)	5.1 (2/8)	11	61
1erv	105	13	275	167	2.0 (1/8)	2.5 (1/7)	2.3 (1/8)	13	238
5fd1	106	13	103	185	9.6 (2/4)	7.6 (1/3)	8.9 (1/2)	12	87
1cewI	108	14	160	127	5.2 (1/5)	6.4 (1/4)	5.1 (1/8)	14	137
1pdo	121	15	67	122	4.8 (1/2)	4.6 (1/3)	4.2 (2/6)	15	55
Average					5.11 (1.9/7.4)	4.40 (1.4/7.0)	3.91 (1.5/6.9)		
RMSD < 6.5					47 (47)	56 (55)	60 (59)		
RMSD < 6.0					45 (45)	51 (51)	58 (57)		
RMSD < 5.0					38 (38)	43 (43)	52 (51)		
RMSD < 4.0					24 (24)	30 (30)	34 (34)		
RMSD < 3.0					11 (11)	20 (20)	21 (21)		

[a]N: the number of protein residues.
[b]N_{exact}: the number of simulated, exact long-range sidechain contact restraints.
[c]N_{pred_cont}: the number of predicted sidechain contact restraints.
[d]N_{pred_dist}: the number of predicted local distance restraints.
[e]Predicted only: prediction with use of predicted sidechain contact restraints and predicted local distance restraints.
[f]Exact only: prediction with use of exact sidechain contact restraints and predicted local distance restraints.
[g]Exact + predicted: prediction with use of exact sidechain contact restraints, predicted sidechain contact restraints, and predicted local distance restraints.
[h]N_{exact_sat}: the number of exact sidechain contact restraints satisfied in the best cluster centroids from prediction with use of exact sidechain contact restraints, predicted sidechain contact restraints, and predicted local distance restraints.
[i]N_{pred_sat}: the number of predicted sidechain contact restraints satisfied in the best cluster centroids from prediction with use of exact sidechain contact restraints, predicted sidechain contact restraints, and predicted local distance restraints.
In columns 6, 7, and 8, the data are shown as the RMSD of the best cluster centroid (rank of the best cluster centroid/total number of clusters). The best cluster centroid is the one that has the lowest RMSD from native.
The last five rows list the number of proteins folded with an RMSD from native below 6.5 Å, 6.0 Å, 5.0 Å, 4.0 Å, and 3.0 Å, respectively. The data are represented as the number of proteins folded to a given RMSD threshold in all clusters (the number of proteins folded to a given RMSD threshold in top five lowest energy clusters).
RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.

RESULTS AND DISCUSSION

All structure prediction results, including PDB format predicted structures, can be found on our group's website: http://www.bioinformatics.buffalo.edu/new_buffalo/people/wli7/touchstonex.

Structure Prediction of 65 Benchmark Proteins With Use of *N*/8 Simulated Restraints

The 65 proteins in set I are listed in Table I, together with their types and lengths. The simulated, exact, long-range sidechain contact restraints were randomly generated, as described in the Materials and Methods section. The number of exact sidechain contact restraints for each protein is listed in Table I, together with the number of predicted sidechain contact restraints and predicted local distance restraints. As shown in the Column 7 of Table I, with the use of only *N*/8 exact sidechain contact restraints, with predicted local distance restraints, but without any predicted contact restraints, 56 proteins are foldable (i.e., at least one cluster centroid with an RMSD of less than 6.5 Å from native was obtained). Fifty-five of these 56 centroids are from the top 5 lowest energy clusters. The accuracy/RMSD of the prediction shows no apparent dependence on protein length or secondary structure type. In other words, the prediction is good for both large proteins and β-proteins. The average RMSD of the best cluster centroid (lowest RMSD from native) of all 65 proteins (folded and unfolded) is 4.4 Å. The average rank of the best RMSD cluster centroids is 1.4, with an average of 7.0 for the total number of clusters.

Column 8 in Table I lists the prediction results with use of the simulated exact contact restraints plus the predicted contact restraints, together with the predicted local distance restraints. Here, the results are improved even

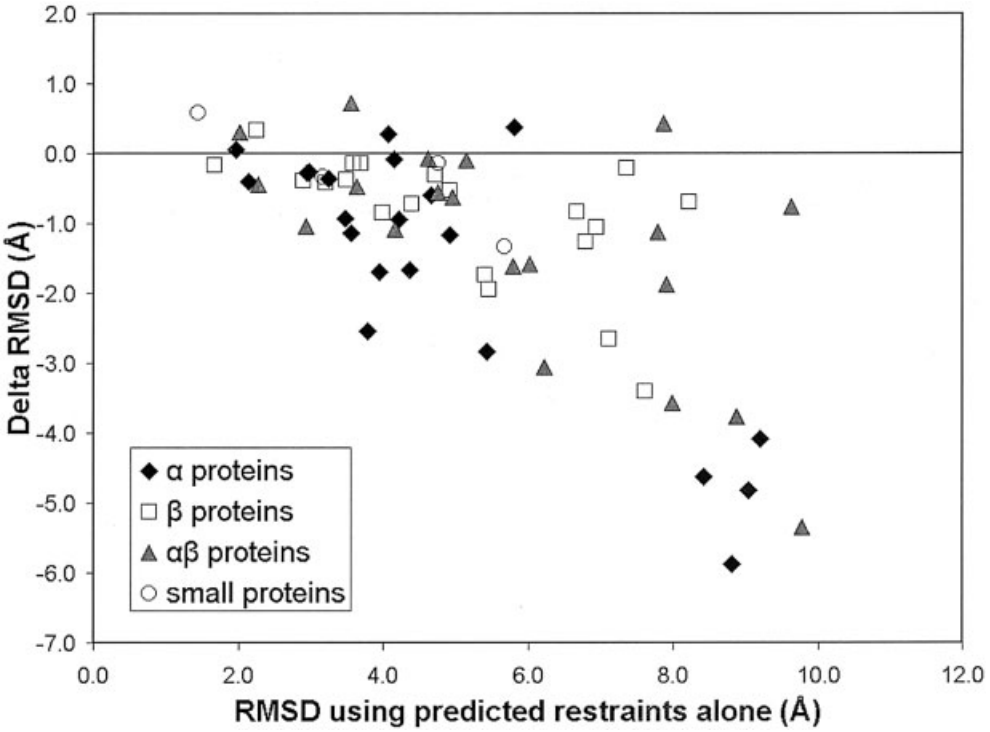

Fig. 1. RMSD improvement with the use of *N*/8 exact, long-range sidechain contact restraints and predicted restraints as a function of the RMSD with the use of only predicted restraints for 65 proteins in set I. RMSD improvement (Delta RMSD) is the difference in the RMSD of the lowest RMSD cluster centroid from native between the prediction with predicted restraints and that with both exact and predicted restraints.

TABLE II. Summary of Prediction Results for 65 Proteins in Set I With Use of *N*/12-, *N*/8-, *N*/4-Simulated, Exact Long-Range Sidechain Contact Restraints

N_{exact}[a]	*N*/12[b,c]	*N*/8[b,c]	*N*/4[b,c]
Average	4.17 (1.3/7.1)	3.85 ± 0.08 (1.3 ± 0.2/6.7 ± 0.2)	3.20 (1.2/6.7)
RMSD < 6.5	59 (59)	61 ± 1 (61 ± 2)	65 (65)
RMSD < 6.0	56 (56)	57 ± 1 (57 ± 1)	64 (64)
RMSD < 5.0	45 (45)	52 ± 1 (51 ± 1)	59 (59)
RMSD < 4.0	33 (33)	38 ± 4 (38 ± 4)	49 (49)
RMSD < 3.0	21 (21)	21 ± 2 (21 ± 2)	32 (32)

[a]N_{exact}: the number of simulated, exact long-range sidechain contact restraints.
[b]*N*: the number of protein residues.
[c]Predictions with *N*/12 and *N*/8 exact contact restraints used both exact and predicted contact restraints and predicted local distance restraints, whereas the prediction with *N*/4 exact contact restraints used exact contact restraints and predicted local distance restraints alone.
Results with *N*/8 exact contact restraints (the third column) are from predictions that used three sets of randomly generated restraints. The mean (± standard deviation) is shown.
In the second row, the data are shown as the average RMSD of the best cluster centroids (average rank of the best cluster centroids/average total number of clusters) for 65 proteins in set I. The best cluster centroid is the one that has the lowest RMSD from native.
The third through seventh rows list the numbers of proteins folded with an RMSD from native below 6.5 Å, 6.0 Å, 5.0 Å, 4.0 Å, and 3.0 Å, respectively. The data are represented as the number of proteins folded to a given RMSD threshold in all clusters (the number of proteins folded to a given RMSD threshold in top five lowest energy clusters).
RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.

further: 60 proteins were folded to RMSD from native below 6.5 Å, 59 of which are from the top 5 lowest energy clusters. Also, significant improvement is seen in the number of folded structures with RMSD from native below 6.0, 5.0, and 4.0 Å, which increases from 51 to 58, from 43 to 52, and from 30 to 34, respectively. The average RMSD of the best cluster centroids of all 65 proteins also decreases to 3.91 Å compared to 4.4 Å previously. The average ranking of the best RMSD cluster centroid is 1.5 from an average of 6.9 for the total number of clusters.

As a control, column 6 in Table I shows results with the use of only predicted contact and distance restraints. Without exact restraints, our ab initio folding algorithm, *TOUCHSTONE*,[22] can fold 47 proteins, all within the top 5 lowest energy clusters. The average RMSD of the best cluster centroids of all 65 proteins is 5.11 Å. Compared to folding with the use of only predicted restraints, the prediction with only *N*/8 exact contact restraints (without predicted contact restraints) does a better job even though the number of predicted restraints is, on average, 1.3 times the number of the residues. When we added *N*/8 exact contact restraints to the predicted restraints, a 1.2 Å improvement in the average RMSD of the best cluster centroids was achieved. Figure 1 shows the individual RMSD improvement of the prediction with use of both exact and predicted restraints versus the RMSD of the prediction with predicted restraints alone for different types of proteins. There is a strong correlation between them (i.e., greater improvement for high RMSD structures than for low RMSD structures). However, the RMSD improvement shows no apparent dependence on protein type and only a very weak dependence on protein length.

We can also see here the complementary effect of the exact and predicted restraints. In 41 out of 65 cases, the prediction with both exact and predicted contact restraints performed better than the prediction with either alone. The two kinds of restraints are synergistic.

We also show in columns 9 and 10 of Table I the numbers of satisfied exact and predicted contact restraints in the best cluster centroids predicted with the use of both exact and predicted restraints. As expected, most of the exact restraints are satisfied. On average, 77% of the predicted contact restraints are satisfied.

To see how much the prediction results depend on a particular selection of the exact contact restraints, we randomly generated three sets of *N*/8 exact restraints and made three predictions. The average results and standard deviations are listed in Table II, column 3. All three sets achieve a similar performance. The mean of the average RMSD of best cluster centroids for all 65 proteins is 3.85 Å. The standard deviation of the average RMSD is only 0.08 Å. The mean of the number of proteins folded to below 6.5 Å from native is 61. The standard deviation of the number of these folded proteins is 1. The mean rank of the best cluster centroids is 1.3. The standard deviation of the rank is 0.2. The mean total number of clusters is 6.7. The standard deviation of the total number of clusters is 0.2. Clearly, the use of different random sets of restraints did not make much of a difference in our results.

Structure Prediction of 65 Benchmark Proteins as a Function of the Number of Restraints

Table II summarizes the results with the use of *N*/12, *N*/8, and *N*/4 exact sidechain contact restraints. The results with *N*/8 exact sidechain contact restraints are averaged over three sets of randomly generated restraints. When the number of restraints is small (e.g., *N*/12 and *N*/8), the prediction is made with the use of both exact and predicted contact restraints. When the number of restraints is large (e.g., *N*/4), the prediction is made with the use of only exact contact restraints. In such cases, the number of exact restraints should be sufficient to achieve successful folding. The predicted restraints sometimes only add noise, because they are not totally accurate. As expected, the more exact the restraints, the more accurate the results. The average RMSD of the lowest RMSD (from native) cluster centroids for all 65 proteins decreases with an increasing number of restraints—from 4.2 Å to 3.9 Å to 3.0 Å as the number of restraint increases from *N*/12 to *N*/8 to *N*/4. Also, the number of proteins folded to less than 6.5 Å—6.0, 5.0, 4.0, and 3.0 Å—from native gradually increases. This is especially significant for the number of proteins folded below 4.0 Å from native, which increases from 33 to 38 to 49 as the number of restraints grows from *N*/12 to *N*/8 to *N*/4. When we used *N*/4 exact restraints, all 65 proteins in this set were folded to an RMSD from native less than 6.5 Å within the top 5 lowest energy clusters. The average of the total number of clusters and the rank of the best cluster centroid also have a tendency to decrease with the increase of the number of restraints. The average of the total number of clusters is 7.1, 6.7, and 6.7 for *N*/12, *N*/8, and *N*/4 exact restraints, respectively. The average rank of the best cluster centroids is 1.3, 1.3, and 1.2, respectively.

We have also observed that the results from the prediction with the use of *N*/12 exact restraints do not differ very much from those with *N*/8 exact restraints for this set of 65 proteins. The average RMSD of the lowest RMSD clusters for all 65 proteins differs only by 0.32 Å. The number of proteins folded to below 6.5 Å from native differs only by two. This indicates that our folding algorithm can make a reasonably good prediction even with as few as *N*/12 exact restraints.

Structure Prediction of 60 Benchmark Proteins With Use of *N*/8 Simulated Restraints

For the 60 proteins in set II (a harder set), the number of the exact sidechain contact restraints, the predicted sidechain contact restraints and local distance restraints for each protein, and the prediction results are shown in Table III, columns 3 through 8. Due to the larger size of the proteins and the lower quality of the predicted restraints, the second set of 60 proteins is much more challenging to fold than the first set. This can be seen from the folding simulation with use of only predicted restraints. Only 28 out of 60 proteins (47%) were folded to an RMSD from native less than 6.5 Å for set II, whereas 47 out of 65 (72%) proteins were folded for set I. The average RMSD of the best cluster centroids for all 60 proteins in set II is 7.08 Å, whereas the average of the 65 proteins in set I is 5.11 Å.

Nevertheless, when we use only *N*/8 exact contact restraints (without predicted contact restraints), 47 proteins are folded to an RMSD from native of less than 6.5 Å, 46 of which are from the top 5 lowest energy clusters. The average RMSD of the best cluster centroids for all 60 proteins is 5.22 Å. The average of the total number of

TABLE III. Prediction Results for the 60 Proteins in Set II Using *N*/8-Simulated, Exact Long-Range Sidechain Contact Restraints

ID	N[a]	N_{exact}[b]	N_{pred_cont}[c]	N_{pred_dist}[d]	Predicted Only[e]	Exact Only[f]	Exact + Predicted[g]
α							
1ppt	36	5	31	68	7.5 (3/3)	3.2 (1/3)	2.2 (3/3)
2erl	40	5	14	57	5.8 (4/7)	6.6 (5/5)	5.4 (1/4)
1eq7A	56	2	6	517	2.6 (3/4)	1.6 (1/2)	1.6 (2/3)
1nkd	59	7	13	785	4.1 (2/2)	2.6 (1/3)	2.5 (1/3)
1i2tA	61	8	52	188	3.2 (2/3)	1.6 (1/7)	2.4 (1/8)
1isuA	62	8	110	131	2.8 (1/9)	6.0 (1/12)	2.1 (1/12)
1ail	70	9	53	154	7.4 (2/4)	2.0 (1/7)	2.8 (1/9)
1utg	70	9	51	269	8.7 (6/6)	6.2 (1/5)	5.4 (2/8)
1hbkA	89	11	29	432	14.3 (1/2)	4.7 (1/3)	5.1 (1/5)
1cy5A	92	12	152	317	1.7 (1/10)	2.4 (1/8)	1.9 (1/8)
1fk5A	93	12	132	89	5.1 (2/4)	3.4 (1/9)	3.7 (1/9)
1bkrA	108	14	131	225	2.1 (1/15)	2.3 (1/6)	2.0 (1/11)
1e6iA	110	14	8	1275	13.1 (1/4)	10.5 (1/6)	10.4 (6/9)
2a0b	118	15	124	227	11.2 (3/5)	2.6 (1/8)	5.9 (2/9)
1fazA	122	15	5	276	11.2 (8/8)	4.7 (1/9)	5.0 (1/9)
1 cpq	129	16	17	1002	7.0 (1/4)	5.4 (1/6)	5.1 (1/9)
1eca	136	17	201	450	3.5 (1/11)	3.0 (1/8)	2.8 (1/8)
2hbg	147	18	161	455	1.7 (1/10)	2.6 (1/10)	2.0 (1/12)
1sra	151	19	118	136	9.4 (5/10)	3.7 (1/8)	4.6 (1/5)
1bd8	156	20	228	439	3.2 (1/17)	7.5 (1/6)	2.6 (1/12)
β							
1dxgA	36	5	41	54	6.3 (1/4)	5.0 (1/6)	3.3 (2/4)
1apf	49	6	43	38	5.1 (4/7)	4.4 (2/8)	3.7 (2/6)
2cdx	60	8	58	127	3.6 (1/7)	4.7 (2/6)	3.8 (1/7)
1aiw	62	8	13	85	8.0 (8/18)	6.3 (7/13)	6.5 (1/8)
3ebx	62	8	74	149	2.2 (1/7)	2.4 (1/4)	1.9 (1/7)
1f94A	63	8	75	47	4.1 (1/10)	6.2 (1/14)	4.0 (1/3)
1msi	66	8	27	107	4.4 (1/22)	4.0 (1/8)	3.8 (1/8)
1hoe	74	9	45	108	8.6 (1/9)	5.8 (1/6)	5.0 (2/6)
1ezgA	82	10	52	152	9.6 (8/17)	11.0 (4/13)	9.9 (2/11)
1wkt	88	11	95	124	11.0 (15/16)	5.0 (2/9)	10.2 (8/9)
1fna	91	11	122	133	3.1 (1/7)	2.7 (1/4)	2.9 (1/10)
1who	94	12	145	69	5.5 (1/7)	3.8 (1/7)	3.4 (1/13)
1 tul	102	13	78	144	9.6 (3/11)	3.5 (1/5)	3.4 (1/3)
1sfp	111	14	75	120	7.1 (1/6)	3.7 (1/8)	2.7 (1/9)
2mcm	112	14	71	153	10.0 (1/13)	5.8 (2/9)	5.2 (1/7)
1b2pA	119	15	93	211	11.8 (6/28)	8.3 (4/20)	9.0 (3/13)
1bfg	126	16	168	211	3.5 (1/9)	8.4 (1/11)	2.3 (1/8)
1c3mA	145	18	72	167	10.8 (1/15)	7.6 (1/10)	6.3 (3/12)
1qj8A	148	19	88	262	11.2 (6/9)	7.7 (2/6)	12.3 (1/5)
2i1b	153	19	97	166	6.9 (15/18)	4.9 (2/9)	5.6 (2/8)
α/β							
1c8cA	64	8	86	215	7.9 (8/8)	7.1 (5/5)	5.6 (3/6)
1i27A	73	9	72	80	5.6 (4/10)	6.0 (1/5)	4.4 (1/7)
1kp6A	79	10	19	120	9.8 (1/13)	6.4 (1/10)	8.7 (2/6)
1opd	85	11	100	224	10.6 (3/4)	1.7 (1/6)	2.8 (1/5)
1npsA	88	11	100	127	3.6 (1/8)	4.0 (1/9)	2.9 (1/9)
1bm8	99	12	46	135	8.1 (8/13)	5.9 (1/9)	5.3 (1/3)
1t1dA	100	13	101	156	3.7 (1/12)	3.9 (1/6)	2.5 (1/12)
1lkkA	105	13	118	269	4.0 (1/10)	4.7 (2/7)	3.5 (1/12)
1bkf	107	13	96	195	11.0 (1/10)	5.8 (2/7)	5.8 (1/7)
1gnuA	117	15	69	120	9.8 (6/7)	6.1 (2/5)	6.7 (2/7)
1dhn	121	15	135	366	3.3 (1/9)	5.1 (1/9)	3.1 (1/10)
2sak	121	15	40	170	9.2 (3/27)	4.3 (1/8)	6.3 (1/3)

TABLE III. (Continued)

ID	N[a]	N_{exact}[b]	N_{pred_cont}[c]	N_{pred_dist}[d]	Predicted Only[e]	Exact Only[f]	Exact + Predicted[g]
1lid	131	16	146	420	2.5 (1/13)	3.6 (1/14)	2.8 (1/10)
1qqhA	144	18	105	128	14.6 (3/10)	11.6 (4/12)	12.0 (1/8)
1f4pA	147	18	176	103	2.8 (1/14)	3.7 (1/3)	2.6 (1/10)
1nbcA	155	19	118	248	5.7 (1/8)	5.2 (2/6)	4.2 (1/7)
1qstA	160	20	216	223	7.7 (2/8)	5.8 (1/8)	4.2 (1/12)
1fw9A	164	21	29	324	12.7 (11/21)	7.7 (1/4)	9.5 (1/8)
1koe	172	22	162	219	14.3 (9/12)	8.1 (3/8)	12.1 (1/9)
1amm	174	22	205	289	10.28 (2/5)	13.2 (8/14)	9.6 (3/7)
Average					7.08 (3.2/10.0)	5.22 (1.7/7.7)	4.92 (1.5/7.8)
RMSD < 6.5					28 (28)	47 (46)	48 (48)
RMSD < 6.0					27 (27)	40 (40)	46 (46)
RMSD < 5.0					21 (21)	32 (32)	35 (35)
RMSD < 4.0					18 (18)	22 (22)	30 (30)
RMSD < 3.0					8 (8)	12 (12)	20 (20)

[a]N: the number of protein residues.
[b]N_{exact}: the number of simulated exact long-range sidechain contact restraints.
[c]N_{pred_cont}: the number of predicted sidechain contact restraints.
[d]N_{pred_dist}: the number of predicted local distance restraints.
[e]Predicted only: prediction with use of predicted sidechain contact restraints and predicted local distance restraints.
[f]Exact only: prediction with use of exact sidechain contact restraints and predicted local distance restraints.
[g]Exact + predicted: prediction with use of exact sidechain contact restraints, predicted sidechain contact restraints, and predicted local distance restraints.
In the sixth through eighth columns, the data are shown as the RMSD of the best cluster centroid (rank of the best cluster centroid/total number of clusters). The best cluster centroid is the one that has the lowest RMSD from native.
The last five rows list the number of proteins folded with an RMSD from native below 6.5 Å, 6.0 Å, 5.0 Å, 4.0 Å, and 3.0 Å, respectively. The data are shown as the number of proteins folded to a given RMSD threshold in all clusters (the number of proteins folded to a given RMSD threshold in top five lowest energy clusters);
RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.

cluster centroids is 7.7, and the average rank of the best cluster centroids is 1.7.

When both exact and predicted contact restraints were used, 48 proteins were folded to an RMSD from native less than 6.5 Å within the top 5 lowest energy clusters. More significant improvement is seen in the number of proteins folded to an RMSD from native below 6.0, 5.0, 4.0, and 3.0 Å. The average RMSD of the best cluster centroids of 60 proteins is 4.92 Å, which is a 2.2 Å improvement over the prediction without exact restraints. The individual RMSD improvement of the prediction with the use of both exact and predicted restraints for each protein compared to the RMSD of the prediction with predicted restraints alone is plotted in Figure 2 according to protein type. Again, the RMSD improvement has a strong correlation to the RMSD of the prediction without exact restraints, with no correlation to protein length and only a very weak correlation to protein type (α- and α/β-proteins are slightly better improved than β-proteins). The average of the total number of clusters is 7.8, and the average rank of the best cluster centroids is 1.5.

Compared to set I, the difference between the prediction with both exact and predicted contact restraints, and the prediction with only predicted contact restraints for set II is greater; this indicates that the overall quality of the predicted restraints for this set is not as good as that for the first set. This is also indicated by the weaker complementary effect of the exact and predicted contact restraints. In 34 out of 60 cases, the prediction with both exact and predicted contact restraints performed better than the prediction with either exact or predicted contact restraints alone. However, this result also demonstrates that a small number of exact restraints can be very critical.

It is worthwhile to look into the details of the largest proteins in this set. For the 11 proteins that are larger than the largest protein in set I (1f4pA, α/β, 147 residues; 2hbg_, α, 147 residues; 1qj8A, β, 148 residues; 1sra_, α, 151 residues; 2i1b_, β, 153 residues; 1nbcA, α/β, 155 residues; 1bd8_, α, 156 residues; 1qstA, α/β, 160 residues; 1fw9A, α/β, 164 residues; 1koe_, α/β, 172 residues; 1amm_, α/β, 174 residues), 4 (1f4pA, 2hbg_, 1nbcA, 1bd8_) were folded to an RMSD from native less than 6.5 Å with the use of predicted restraints alone, and an additional 3 (1sra_, 2i1b_, 1qstA) were folded with the use of both predicted and N/8 exact restraints. There are still 4 proteins (1qj8A, 1fw9A, 1koe_, and 1amm_) that could not be folded with N/8 exact restraints. These are large β- or α/β-proteins having many secondary structure elements, especially 1amm_ which consist of two domains; hence, they need more restraints to fold. When we used N/4 exact restraints, all were folded.

Four proteins from this set could not be folded, even with the use of N/4 exact contact restraints. These were not necessarily the largest proteins; rather, they possess special topologies. 1b2pA (β, 119 residues) is part of a homodimer. It is a 3-β-sheet orthogonal prism made up of 12

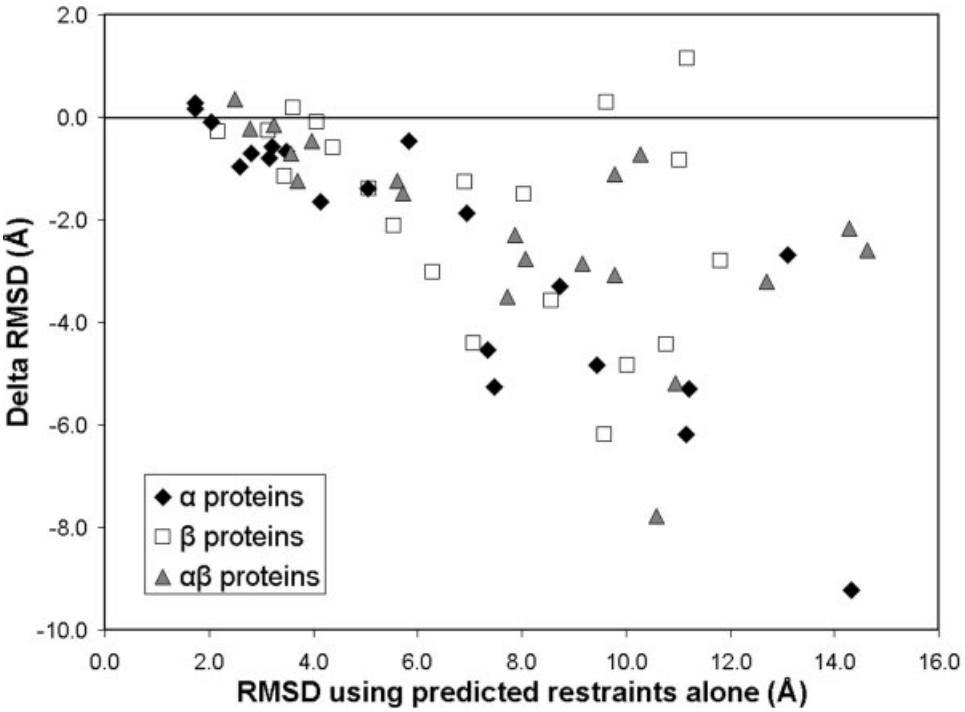

Fig. 2. RMSD improvement with the use of *N*/8 exact long-range sidechain contact restraints and predicted restraints as a function of the RMSD with the use of only predicted restraints for 60 proteins in set II. RMSD improvement (Delta RMSD) is the difference in the RMSD of the lowest RMSD cluster centroid from native between the prediction with predicted restraints and that with both exact and predicted restraints.

strands. 1e6iA (α, 110 residues) forms a complex with a small peptide ligand. It is a 7-helix up–down bundle. 1ezgA (β, 82 residues) is also part of a homodimer and is a solenoid made up of 6 β-strands. 1qqhA (α/β, 144 residues) is a very elongated structure made up of bifurcated and coiled β-sheets (9 strands) and 3 helices.

In the test sets discussed above, we used randomly generated, exact sidechain contact restraints. These restraints usually distribute over the protein randomly. Yet, in reality, NOE data do not necessarily distribute over the protein. They can be limited to part of the protein. To address this problem, we randomly selected 18 proteins from set II and considered the case of exact contact restraints from a randomly selected region of the protein sidechain contact map. With *N*/8 exact contact restraints, *TOUCHSTONEX* improved the overall RMSD of the best cluster centroids by an average of 1.2 Å (from an average RMSD of 7.2 Å decreasing to 6.0 Å). The RMSD of the protein regions was improved by an average of 1.9 Å (from an average RMSD of 6.6 Å decreasing to 4.7 Å). This shows that our folding algorithm does not require that the restraints be distributed over the protein. Any kind of exact restraints are helpful.

In these test sets, we used exact restraints (100% accurate). A practical issue in the NMR structure determination process is that sometimes bad restraints are encountered. In the implementation of restraints in our force field, we allowed a small number of violations of the exact restraints. Our algorithm can tolerate a small number of bad restraints. However, the quality of the exact restraints indeed is very important for our folding algorithm. Too many bad restraints can ruin the results.

Comparison of Prediction Results for Eight Proteins with Previously Published Results from Our Group

Our group has previously published two articles on folding proteins with the use of a small number of exact tertiary restraints.[8,9] In the first, we used the CAPLUS model of proteins, and in the second, the SICHO model of proteins. In those articles, 9 proteins were folded to moderate resolution with as few as *N*/7 long-range sidechain contact restraints. To see how much better we could do with our current algorithm, we refolded 8 of these 9 proteins with the same set of restraints we used previously (the data for protein 4fab are missing in the previous article). For the purpose of direct comparison, in the folding, we did not use any predicted contact restraints.

The results are clearly better with our new algorithm. As shown in Table IV, using the same set of restraints, we folded all proteins with an RMSD that is ~0.5–1.7 Å closer to native. In the previous articles, the algorithms used the native secondary structure as input. Here, we used predicted secondary structure. Nevertheless, our new algorithm still greatly outperformed our previous algorithms.

The robustness of our new method, *TOUCHSTONEX*, comes from the interplay of the protein model, the force field, and the conformational search engine. Due to the differences in these aspects between *TOUCHSTONEX*

TABLE IV. Comparison of Prediction Results for Eight Proteins With Previously Published Results

ID	N[a]	Type	N_{exact}[b]	RMSD[c] (CABS)	RMSD[d] (SICHO)	RMSD[e] (CAPLUS)
2gb1	56	α/β	8	2.8	3.4	3.3
6pti	57	α/β	12	4.4	—	10.0
			12	4.8	—	6.2
1ctf	68	α/β	10	2.6	3.2	4.2
1pcy	99	β	46	2.4	3.8	3.5
			25	3.3	4.9	5.4
2trxA	108	α/β	30	2.2	3.1	3.4
3fxn	138	α/β	35	2.2	4.1	3.9
1mba	146	α	20	2.7	4.3	5.9
1timA	247	α/β	62	5.0	5.1	—
			50	5.2	6.0	—
			36	5.3	6.7	—

[a]N: the number of residues of the protein.
[b]N_{exact}: the number of simulated, exact, long-range sidechain contact restraints.
[c]CABS: our new algorithm. The results are from the lowest RMSD cluster centroid from native.
[d,e]SICHO[9] and CAPLUS[8]: our previously published algorithms.
RMSD: coordinate root-mean-square deviation for the C_α atoms in angstrom units.

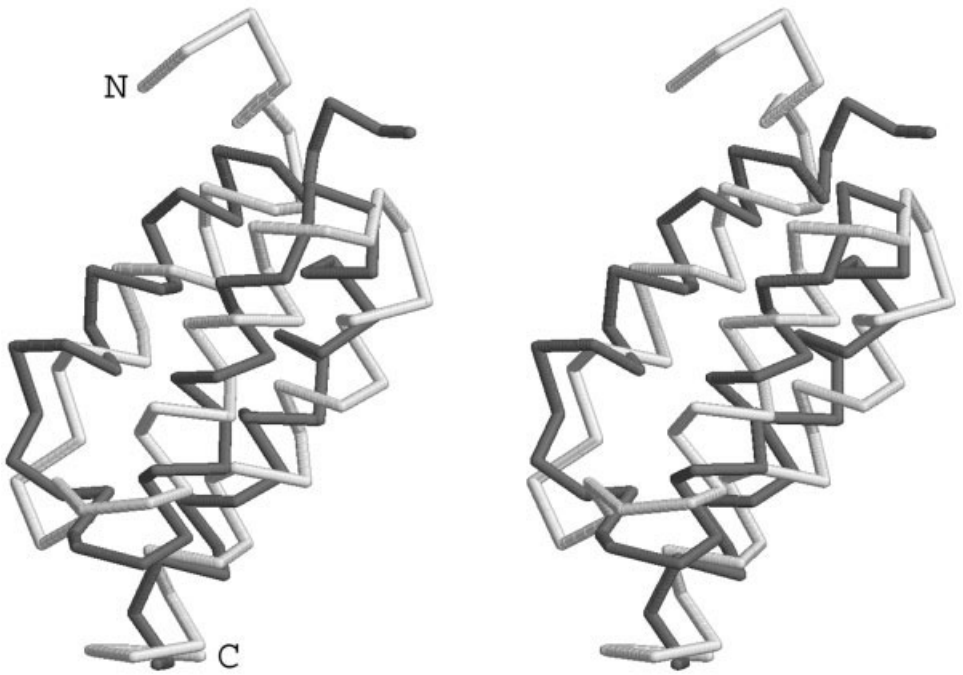

Fig. 3. The C_α model of the Z-domain of staphylococcal protein A from experimental restraints superimposed on the first PDB NMR structure of 2SPZ (native). The C_α backbone of the predicted model is indicated by the dark line, and the C_α backbone of 2SPZ, by the light line.

and other approaches reviewed in our Introduction, it is difficult to identify uniquely the specific basis for the differences in performance. However, our impression is that the optimized force field is a key factor to the success of *TOUCHSTONEX*.

Structure Prediction of Three Proteins With the Use of Experimental Restraints

In this section, we describe the application of our algorithm in three cases in which experimental restraints were available. These restraints are slightly different from the simulated restraints used in the above test sets. They include not only sidechain contact restraints but also sidechain–main chain and main chain–main chain contact restraints. The main chain assignment can usually be performed during the early stage of NMR structure determination, before the sidechain assignment is performed.[39] Therefore, utilizing main chain restraints is very important for rapid protein structure determination by NMR.

TABLE V. Summary of Data for the Z-Domain of Staphylococcal Protein A

A. Restraints for the Z-Domain Obtained From Experimental NOE Data

Type of Restraints		Number of Restraints	Residue 1 and Residue 2
Experimental restraints	Sidechain contact restraints	4	16 34
			21 55
			22 55
			26 55
	Sidechain–main chain contact restraints	3	55 21
			22 55
			55 27
Predicted sidechain contact restraints		71	
Predicted local distance restraints		142	

B. Prediction Results for the Z-Domain as Compared to the Native Structure 2SPZ

Cluster	Number of structures	RMSD[a]	DRMSD[b]	Topology
1	9931	8.28	2.52	Mirror
2	792	2.35	1.87	Native
3	8	12.88	11.27	Misfold

[a]RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.
[b]dRMSD: distance root-mean-square deviation for C_α atoms in angstrom units.

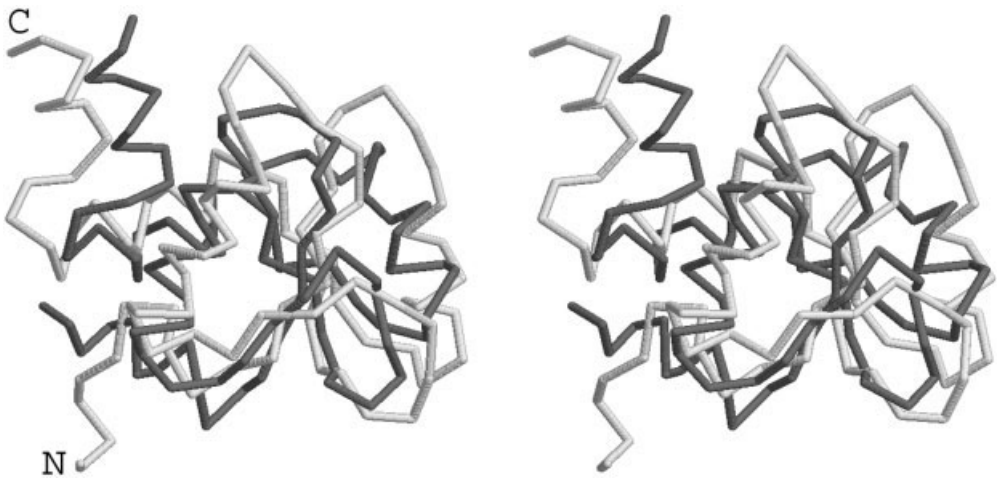

Fig. 4. The C_α model of BRCT from experimental restraints superimposed on the first PDB NMR structure of 1L7B (native). The C_α backbone of the predicted model is shown by the dark line, and the C_α backbone of 1L7B, by the light line.

Structure prediction of the Z-domain of staphylococcal protein A

The first protein is the Z-domain of staphylococcal protein A. The high-resolution NMR structure of this protein has been solved and deposited in the PDB (PDB code 2SPZ).[40] This protein is a 3-helix bundle with 58 residues (Fig. 3). For the purpose of rapid NMR structure determination with only H^N and methyl assignments, we generated a list of distance constraints between H^N and methyl proton resonances by AutoStructure. A subset of long-range constraints was selected and used here for *TOUCHSTONEX* analysis.

We filtered the selected experimental constraints according to the procedure described in the Materials and Methods section. The input contact restraints are listed in Table V(A), together with the number of predicted local distance restraints and sidechain contact restraints. There are 4 experimental long-range sidechain contact restraints and 3 long-range sidechain–main chain contact restraints. The folding algorithm generated 3 clusters in total. Table V(B) lists the number of structures in each cluster, the topology of the cluster centroids, and their C_α coordinate RMSD and distance RMSD (dRMSD) to native 2SPZ. The lowest energy cluster centroid is the mirror image of the native fold. The second lowest energy cluster centroid has the correct native fold. These 2 clusters are also the most populated ones, containing 9931 and 792 structures, respectively. The third cluster is misfolded and also much less populated (containing only 8 structures). The C_α model of the second cluster centroid superimposed on the first PDB NMR model of 2SPZ is shown in Figure 3.

TABLE VI. Summary of Data for BRCT

A. Restraints for BRCT Obtained From Experimental NOE Data

Type of Restraints		Number of Restraints	Residue 1 and Residue 2
Experimental restraints	Sidechain contact restraints	6	8 49
			11 69
			15 27
			26 79
			49 78
			79 86
	Sidechain–main chain contact restraints	10	77 8
			12 45
			27 14
			19 51
			49 30
			60 39
			60 47
			48 68
			84 49
			85 70
	Main chain–main chain contact restraints	6	13 47
			14 48
			15 49
			48 68
			49 69
			50 70
	Main chain hydrogen-bond restraints	2	14 48
			49 69
Predicted sidechain contact restraints		294	
Predicted local distance restraints		214	

B. Prediction Results for BRCT Compared to the Native Structure 1L7B

Cluster	Number of Structures	RMSD[a]	DRMSD[b]	Topology
1	12,815	6.13	4.38	Native
2	91	4.49	3.53	Native
3	23	11.39	5.08	Misfold
4	4	10.75	6.09	Misfold
5	8	11.96	5.92	Misfold
6	22	9.47	5.98	Misfold
7	4	12.29	6.17	Misfold
8	4	10.51	7.01	Misfold

[a]RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.
[b]dRMSD: distance root-mean-square deviation for C_α atoms in angstrom units.

The two structures are very similar. Neglecting the flexible ends, the RMSD to native for residues 8–55 is 2.4 Å. For comparison, the RMSD of the model generated without experimental restraints is 3.5 Å from native 2SPZ.

Structure prediction of BRCT

The second protein considered is BRCT. The high-resolution NMR structure of this protein has been determined, with the program AutoStructure, and deposited in the PDB (PDB code 1L7B).[37] This protein is larger than 2SPZ and has 92 residues with a two-layer $\alpha + \beta$ fold (Fig. 4). The list of distance and hydrogen-bonding constraints was first generated by the initial stage of AutoStructure analysis. The long-range constraints from peaks of weak–medium to strong intensity were then selected and filtered for *TOUCHSTONEX* analysis, including 6 long-range sidechain–sidechain contact restraints, 10 long-range sidechain–main chain contact restraints, 6 long-range main chain–main chain contact restraints, and 2 long-range main chain hydrogen-bonding restraints. The restraints are listed in Table VI(A).

The restrained folding generated 8 clusters. The overall results are listed in Table VI(B). The lowest energy cluster is the dominant cluster. Its centroid has the overall topology of 1L7B, with an RMSD of 6.1 Å for residues 6–91 (neglecting the flexible ends). The lowest RMSD centroid came from the second lowest energy cluster, which has 91 structures. The RMSD to native is 4.5 Å. The C_α model of the second cluster centroid superimposed on the first PDB NMR model of 1L7B is shown in Figure 4. The two proteins are very similar except for their flexible ends. The remaining clusters are all misfolded and much less populated. For comparison, the folding algorithm could not fold this protein without

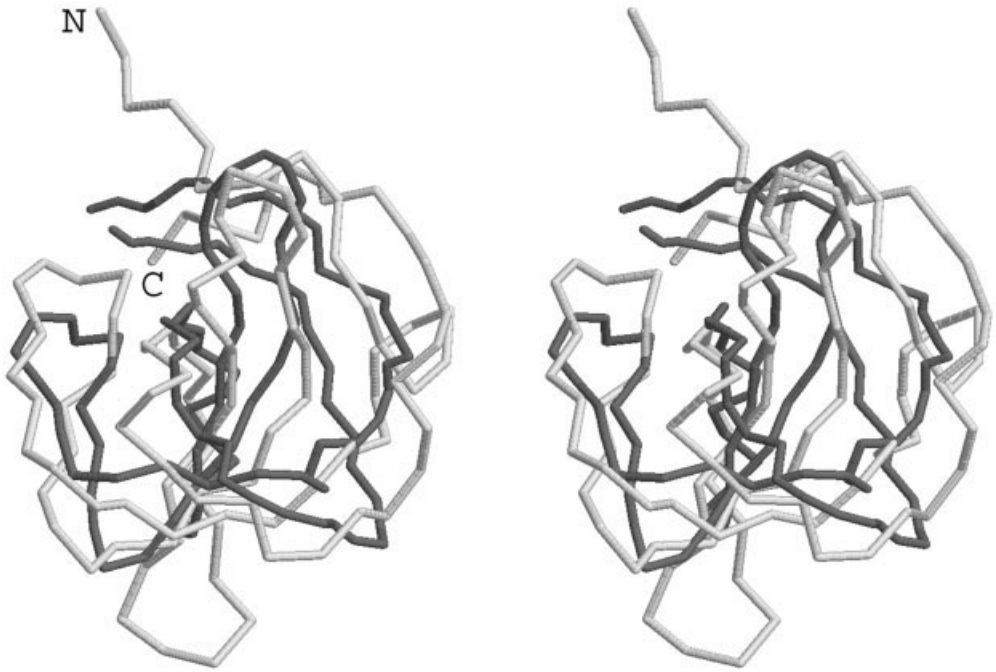

Fig. 5. The C_α model of the MIA protein from experimental restraints superimposed with the PDB X-ray structure of 1I1J chain A (native). The C_α backbone of the predicted model is indicated by the dark line, and the C_α backbone of 1I1J chain A, by the light line.

any experimental restraints. The lowest RMSD cluster centroid is 6.8 Å from native.

Structure prediction of MIA

The third protein considered is the 108-residue MIA protein. The X-ray structure of this protein (106 residues, without the first and last residues) has been solved and deposited in the PDB (PDB code: 1I1J[41]). 1I1J consists of two extremely similar chains, A and B, with a backbone RMSD of 0.2 Å. Both are mainly β-proteins and have a partly open β-barrel topology made up of 7 antiparallel β-strands (Fig. 5).

The NOE data for MIA were provided by Dr. Peter Domaille of GeneFormatics, Inc. (San Diego, CA). The data came from the early stage of NMR structure determination and contain very few unambiguous restraints. After filtering, we obtained only 2 long-range sidechain contact restraints, and both involve 1 residue—residue 61. In addition, we obtained 2 long-range sidechain–main chain contact restraints and 12 long-range main chain–main chain contact restraints. These restraints are listed in Table VII(A). In the absence of any experimental restraints, the protein could barely be folded with the use of only predicted restraints. Excluding the dangling first 4 residues, the lowest RMSD between the cluster centroid and 1I1J chain A is 6.5 Å. When experimental restraints were added, the lowest RMSD between the cluster centroid and 1I1J chain A improved to 5.6 Å. The C_α model of this lowest RMSD cluster centroid and X-ray structure of 1I1J are superimposed in Figure 5. The model reproduced the overall topology, even though the secondary structure prediction was poor; it even reproduced the 30s–40s loop region with high fidelity. This lowest RMSD cluster is the lowest energy cluster in a total of 6 clusters generated. Table VII(B) lists the overall results of all 6 clusters. Among them, the first 2 lowest energy cluster centroids have the native fold and are also the most populated, including 3505 and 6863 structures, respectively. The second cluster centroid was also folded to a slightly lower resolution structure, with an RMSD from the 1I1J chain A of 6.2 Å. The third and fourth clusters were also significantly populated, including 2220 and 1391 structures, respectively. The fourth cluster centroid was the mirror image of the native fold. The remaining 2 clusters were misfolded and included very few structures (12 and 4 structures, respectively).

For these three examples, only sparse NOE data were used, especially for MIA, in which the data came from the very early stage of the NMR structure determination. From this limited NOE data, only very few sidechain contact restraints are generated, far fewer than $N/8$ (for Z-domain, $N/15$; for BRCT, $N/15$; for MIA, $N/54$). Aside from the sidechain contact restraints, a small number of sidechain–main chain and main chain–main chain contact restraints (for Z-domain, $N/19$; for BRCT, $N/5$; for MIA, $N/8$) were also generated. Using a total of $N/8$, $N/4$, and $N/7$ restraints, our folding algorithm was able to improve the RMSD of the Z-domain, BRCT, and MIA significantly (1.1 Å, 2.3 Å, and 0.9 Å, respectively).

TABLE VII. Summary of Data for the Melanoma Inhibitory Activity (MIA) Protein

A. Restraints for MIA Obtained From Experimental NOE Data

Type of Restraints		Number of Restraints	Residue 1 and Residue 2
Experimental restraints	Sidechain contact restraints	2	61 52
			61 82
	Sidechain–main chain contact restraints	2	61 53
			41 31
	Main chain–main chain contact restraints	12	63 79
			12 51
			61 79
			62 79
			57 8
			30 43
			53 61
			66 47
			28 85
			50 63
			27 46
			33 40
Predicted sidechain contact restraints		176	
Predicted local distance restraints		316	

B. Folding Results for MIA Compared to the Native Structure 111J Chain A

Cluster	Number of Structures	RMSD[a]	DRMSD[b]	Topology
1	3505	5.60	4.80	Native
2	6863	6.24	4.38	Native
3	2220	9.72	5.64	Misfold
4	1391	10.37	4.83	Mirror
5	12	10.24	6.82	Misfold
6	4	13.95	7.41	Misfold

[a]RMSD: coordinate root-mean-square deviation for C_α atoms in angstrom units.
[b]dRMSD: distance root-mean-square deviation for C_α atoms in angstrom units.

CONCLUSIONS

In this article, we have demonstrated the utility of using a small number of long-range contact restraints in protein structure prediction. Our folding algorithm employs a lattice-based, reduced protein model that explicitly includes C_α, C_β, and sidechain centers of mass. Contact restraints are incorporated into the scoring function as an NOE-specific pairwise contact potential. With use of *N*/8-simulated exact, long-range sidechain contact restraints (with *N* being the number of residues), the accuracy of the assembled structures has been improved relative to the prediction without any exact restraints, as verified by the test case of 125 proteins of various secondary structure types and lengths up to 174 residues. Of these 125 proteins, 108 were folded (with a lowest RMSD cluster centroid below 6.5 Å from native)—33 more than that of the prediction without the use of any exact restraints. The average RMSD of the lowest RMSD cluster centroids from native for all 125 proteins (folded and unfolded) is 4.4 Å, which is 1.6 Å less than that of the prediction without the use of any exact restraints. Moreover, for the 65 proteins in the first set, 59 could be folded with the use of as few as *N*/12 exact restraints compared to 47 without any exact restraints. Indeed, a small number of exact restraints can guide the folding process to reach the native fold and expand the range of manageable proteins.

We also explored the application of our folding algorithm to three proteins with limited experimental NOE data. Using very few experimental sidechain contact restraints (*N*/54 to *N*/15) and a small number of sidechain–main chain and main chain–main chain contact restraints (*N*/19 to *N*/5), all three proteins were folded to low-to-medium resolution structures. The application of *TOUCHSTONEX* to the NMR structure determination process is very promising, especially in the early stages, when only limited data are available. Although we have only considered the case of NMR-derived data here, the algorithm is not limited to NMR, and other experiments that can provide structural restraints are equally valuable.

ACKNOWLEDGMENTS

We gratefully acknowledge Dr. Peter Domaille and GeneFormatics, Inc., for providing us with the NMR data for MIA.

REFERENCES

1. Havel TF, Wüthrich K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of internuclear H1-H1 proximities in solution. Bull Math Biol 1984;46:673–698.
2. Brünger AT. X-PLOR Version 3.1: A system for X-ray crystallography and NMR. New Haven, CT: Yale University Press; 1992.
3. Xu Y, Xu D, Crawford OH, Einstein JR. A computational method

for NMR-constrained protein threading. J Comput Biol 2000;7:449–467.
4. Smith-Brown MJ, Kominos D, Levy RM. Global folding of proteins using a limited number of distance constraints. Protein Eng 1993;6:605–614.
5. Connolly PC, Stern AS, Hoch JC. Estimating protein fold from incomplete and approximate NMR data. J Am Chem Soc 1994;116:2675–2676.
6. Aszodi A, Gradwell MJ, Taylor WR. Global fold determination from a small number of distance restraints. J Mol Biol 1995;251:308–326.
7. Kolinski A, Skolnick J. Lattice models of protein folding, dynamics and thermodynamics. Austin, TX: R.G. Landes Co.; 1996.
8. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: A method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
9. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: An efficient Monte Carlo model. Proteins 1998;32:475–494.
10. Standley DM, Eyrich VA, Felts AK, Friesner RA, McDermott AE. A branch and bound algorithm for protein structure refinement from sparse NMR data sets. J Mol Biol 1999;285:1691–1710.
11. Debe DA, Carlson MJ, Sadanobu J, Chan SI, Goddard WA. Protein fold determination from sparse distance restraints: The restrained generic protein direct Monte Carlo method. J Phys Chem B 1999;103:3001–3008.
12. Bowers PM, Strauss CE, Baker D. De novo protein structure determination using sparse NMR data. J Biomol NMR 2000;18:311–318.
13. Skolnick J, Fetrow JS, Kolinski A. Structural genomics and its importance for gene function analysis. Nat Biotechnol 2000;18:283–287.
14. Skolnick J, Fetrow JS. From genes to protein structure and function: Novel applications of computational approaches in the genomic era. Trends Biotechnol 2000;18:34–39.
15. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J 2003;85.
16. 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.
17. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
18. Kolinski A, Jaroszewski L, Rotkiewicz P, Skolnick J. An efficient Monte Carlo model of protein chains: Modeling the short-range correlations between side group centers of mass. J Phys Chem 1998;102:4628–4637.
19. Zhang Y, Kihara D, Skolnick J. Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding. Proteins 2002;48:192–201.
20. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–331.
21. Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 2001;44:133–149.
22. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 2001;98:10125–10130.
23. Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comput Appl Biosci 1994;10:53–60.
24. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
25. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
26. Wüthrich K. NMR of proteins and nucleic acids. New York: Wiley; 1986.
27. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin glasses. Phys Rev Lett 1986;57:2607–2609.
28. Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 1997;281:140–150.
29. Betancourt M, Skolnick J. Finding the needle in a haystack: Educing protein native folds from ambiguous ab initio folding predictions. J Comput Chem 2001;22:339–353.
30. Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT. Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 1997;269:592–610.
31. Huang YJ. Automated determination of protein structures from NMR data by iterative analysis of self-consistent contact patterns [Ph.D. thesis]. New Brunswick, NJ: Rutgers University; 2001.
32. Greenfield NJ, Huang YJ, Palm T, Swapna GV, Monleon D, Montelione GT, Hitchcock-DeGregori SE. Solution NMR structure and folding dynamics of the N terminus of a rat non-muscle alpha-tropomyosin in an engineered chimeric protein. J Mol Biol 2001;312:833–847.
33. Huang YP, Swapna GVT, Rajan PK, Ke H, Xia B, Shukla K, Inouye M, Montelione GT. Solution NMR structure of ribosome binding factor A (RbfA): A cold-shock adaptation protein from *Escherichia coli*. J Mol Biol 2003;327(2):521–536.
34. Monleón D, Colson K, Moseley HNB, Anklin C, Oswald R, Szyperski T, Montelione GT. High throughput analysis of protein assignments and secondary structure: Rapid data collection using triple resonance CryoProbesTM, parallel processing on LINUX-based processor architectures and automated analysis using AutoAssign software. J Struct Funct Genomics 2002;2:93–101.
35. Zheng D, Huang YJ, Moseley HNB, Xiao R, Aramini J, Swapna GVT, Montelione GT. Automated protein fold determination using a minimal NMR constraint strategy. Protein Sci 2003. Submitted for publication.
36. Montelione GT, Rios BC, Swapna GVT, Zimmerman DE. NMR pulse sequences and computational approaches for automated analysis of sequence-specific backbone resonance assignments of proteins. In: Berliner L, Krishna NR, editors. Modern techniques in protein NMR. Kluwer Academic: New York, 1999. p 81–130.
37. Sahota G, Dixon BL, Huang YJ, Goldsmith D, Aramini J, Bhattacharya A, Monleón D, Swapna GVT, Yin C, Xiao R, Anderson S, Honig B, Montelione GT, Tejero R. Solution NMR structure of the BRCT domain from *T. thermophilus* DNA ligase. Proteins 2003. Submitted for publication.
38. Moy FJ, Li YC, Rauenbuehler P, Winkler ME, Scheraga HA, Montelione GT. Solution structure of human type-alpha transforming growth factor determined by heteronuclear NMR spectroscopy and refined by energy minimization with restraints. Biochemistry 1993;32:7334–7353.
39. Ferentz AE, Wagner G. NMR spectroscopy: A multifaceted approach to macromolecular structure. Q Rev Biophys 2000;33:29–65.
40. Tashiro M, Tejero R, Zimmerman DE, Celda B, Nilsson B, Montelione GT. High-resolution solution NMR structure of the Z domain of staphylococcal protein A. J Mol Biol 1997;272:573–590.
41. Lougheed JC, Holton JM, Alber T, Bazan JF, Handel TM. Structure of melanoma inhibitory activity protein, a member of a recently identified family of secreted proteins. Proc Natl Acad Sci U S A 2001;98:5515–5520.