

REFINEMENT OF NMR-DETERMINED PROTEIN STRUCTURES WITH DATABASE DERIVED DISTANCE CONSTRAINTS

Feng Cui

*Program on Bioinformatics and Computational Biology, Iowa State University
Ames, Iowa 50011, USA
fengcui@iastate.edu*

Robert Jernigan

*Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University
Ames, Iowa 50011, USA
jernigan@iastate.edu*

Zhijun Wu*

*Department of Mathematics and Program on Bioinformatics and Computational Biology, Iowa State University
Ames, Iowa 50011, USA
zhijun@iastate.edu*

Received (Day Month Year)

Revised (Day Month Year)

Accepted (Day Month Year)

The protein structures determined by NMR (Nuclear Magnetic Resonance Spectroscopy) are not as detailed and accurate as those by X-ray crystallography and are often underdetermined due to the inadequate distance data available from NMR experiments. The uses of NMR-determined structures in such important applications as homology modeling and rational drug design have thus been severely limited. Here we show that with the increasing numbers of high quality protein structures being determined, a computational approach to enhancing the accuracy of the NMR-determined structures becomes possible by deriving additional distance constraints from the distributions of the distances in databases of known protein structures. We show through a survey on 462 NMR structures that, in fact, many inter-atomic distances in these structures deviate considerably from their database distributions and based on the refinement results on 10 selected NMR structures that these structures can actually be improved significantly when a selected set of distances are constrained within their high probability ranges in their database distributions.

Keywords: NMR protein structure refinement; Protein structure database; Structural bioinformatics.

1. Introduction

The structures determined by NMR (Nuclear Magnetic Resonance Spectroscopy) are not as detailed and accurate as those by X-ray crystallography due to the inadequate distance data available from NMR experiments.¹⁻³ The uses of NMR-determined structures in such

*To whom the correspondence should send.

important applications as homology modeling and rational drug design have thus been severely limited.

The distance data can only be obtained from NMR for specific atoms and in most cases, hydrogen atoms and be estimated approximately with a set of lower and upper bounds. As a result, an ensemble of structures, instead of a single unique one, usually is determined for a protein. While the variation of the structures in the ensemble is often considered as a reflection of the flexibility of the structures in solution, it could be misleading since the variation can also occur from structural under-determination.

In order to increase the accuracy of NMR structures, more distance data has been sought by using various techniques. Experimental approaches such as dipolar coupling have been developed.⁴⁻⁷ Theoretical approaches include techniques to obtain additional conformational constraints from databases of known protein structures such as to derive constraints on dihedral angles based on their distributions in known X-ray structures in structural databases.⁸⁻⁹

With the increasing number of high-resolution protein structures being determined, many structural properties such as secondary structure motifs, native contact patterns, and hydrophobic core formations, have been revealed from their statistical distributions in known protein structures.¹⁰ The inter-atomic distances are also subject to certain statistical distributions, depending on the types of the distances. Such distributions have been employed for constructing various statistical potentials for contact determination, inverse folding, structure alignment, and X-ray structure refinement.¹¹⁻¹⁶

In this work, the distributions of inter-atomic distances in known protein structures and in particular, in known X-ray structures, are studied and used to extract additional distance constraints for NMR structure refinement. In order to estimate the distributions, a large set of high-resolution protein structures from the Protein Data Bank¹⁷ have been utilized. The distances for selected pairs of atoms across one or two residues along the protein backbones (called cross-residue inter-atomic distances) are sampled to obtain the probability distributions of the distances.

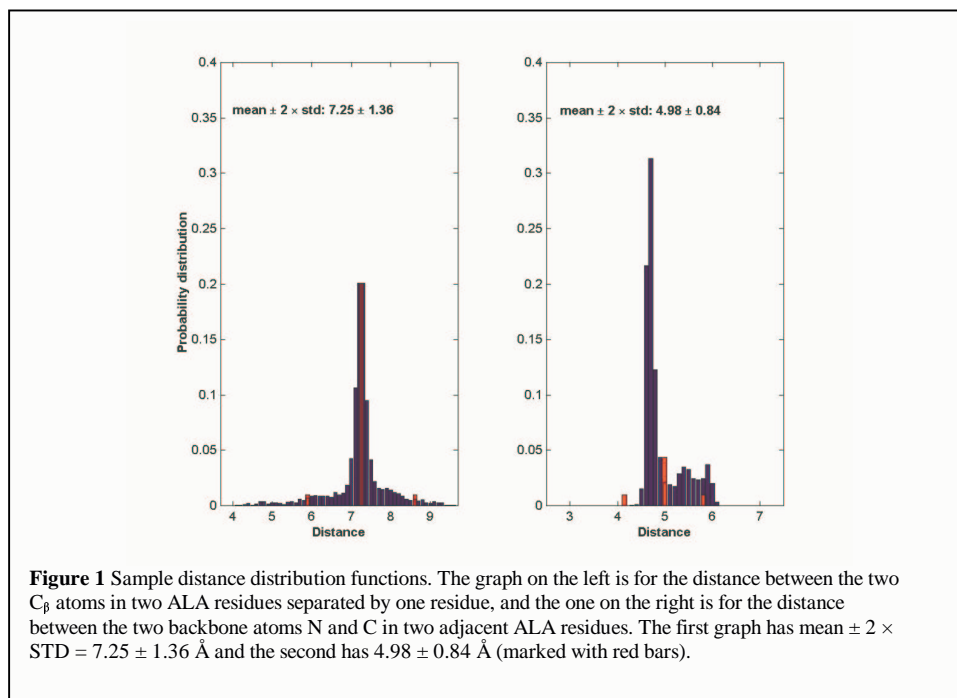
The distribution functions are then used to evaluate a set of NMR structures. The cross-residue inter-atomic distances in each of the structures are compared with their corresponding distribution functions, and the deviations of the distances from their average distributions (means) are recorded. The results show that many cross-residue inter-atomic distances in the structures deviate significantly from their average distributions. More specifically, in each structure, on average, about 22% of the residue pairs that are separated by at most one residue along the protein backbone are found to have cross-residue inter-atomic distances deviating from their means by more than two standard deviations. While the inter-atomic distances in a particular NMR structure do not have to agree with their distributions in known protein structures completely, the large number of cases having large deviations of the distances from the means suggest that many of the distances may be incorrectly formed due to the lack of proper constraints for the distances in the NMR data.

In order to reduce the errors in the distances and hence improve the NMR structures, the distribution functions for selected cross-residue inter-atomic distances are used to extract probable ranges for the distances. The obtained distance constraints (called database distance constraints) are then applied to refining a set of NMR structures, using the modeling software CNS (Crystallography and NMR System) developed by Brünger

and co-workers.¹⁸ The structures are refined through combining the original NMR distance constraints with additional database distance constraints. The refined structures are evaluated in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, the RMSD (root-mean-square-deviation) values of the ensembles of structures, the RMSD values of the structures compared with their X-ray crystal structures (for available ones), as well as the remaining distance errors in the structures. The results show that with additional database distance constraints, the numbers of improperly formed inter-atomic distances in the refined structures decrease significantly, while the RMSD values of the ensembles of structures are reduced and the acceptance rates of the structures are more than doubled, suggesting that protein structures can indeed be determined more accurately and efficiently by combining the distance constraints obtained from NMR experiments with additional distance constraints extracted from known protein structures in structural databases.

2. Distributions of Distances

To estimate the distributions of cross-residue inter-atomic distances of proteins in known protein structures, 2150 X-ray crystal structures with resolution of 2.0 Å or higher and sequence similarity of 90% or less were downloaded from the Protein Data Bank. The distances are specified together with the types of the atom pairs, the types of the residue



pairs, and the sequential separations. More specifically, let D be the distance between two atoms, $A1$ and $A2$ the types of the two atoms, $R1$ and $R2$ the types of the two residues the two atoms are associated with, respectively, and S the number of residues separating $R1$

and R2. Then, the distribution of the distance D between atoms A1 in R1 and A2 in R2 where R1 and R2 are separated by S residues can be represented by using a probability distribution function $P[A1,A2,R1,R2,S](D)$. In this study, only five different types of atoms are considered: the amide N, C_α , and the carbonyl C and O along the backbone and the carbon C_β in the side chain. Residue types include all twenty different amino acid types. The separation S is either one or zero. So in total there are $5 * 5 * 20 * 20 * 2 = 20,000$ possible distance distributions considered. For each set of A1, A2, R1, R2, and S, all corresponding distances in the downloaded crystal structures are computed. The distances are collected into bins of uniform distance intervals $[D_i, D_{i+1}]$, where $D_i = 0.1 * i \text{ \AA}$, $i = 0, 1, \dots, n$. The distribution function $P[A1,A2,R1,R2,S](D)$ for any D in $[D_i, D_{i+1}]$ is then defined to be the number of distances in $[D_i, D_{i+1}]$ normalized by the total occurrences of distances in all intervals. Two sample graphs for $P[A1,A2,R1,R2,S](D)$ are illustrated in Figure 1, one with $A1 = C_\beta$, $A2 = C_\beta$, $R1 = \text{ALA}$, $R2 = \text{ALA}$, and $S = 1$, and another with $A1 = \text{N}$, $A2 = \text{C}$, $R1 = \text{ALA}$, $R2 = \text{ALA}$, and $S = 0$. The distribution graphs for short-range distances have non-uniform patterns in general. This is primarily due to the fact that large portions of protein segments form regular secondary structures, i.e., α -helices or β -sheets, where short-range distances have fixed ranges for fixed types of atoms. The graphs often show two peaks as well, corresponding to the distributions of the related distances in α -helices or β -sheets, respectively. In any case, the distributions can be characterized by their means and standard deviations.

3. Distances in NMR Structures

The inter-atomic distances for 462 averaged and energy-minimized NMR structures downloaded from the Protein Data Bank are examined and compared with their distribution functions as defined and calculated above. The results show that many of these distances have deviations larger than two standard deviations. For example, the distribution of the distance between C_β in ALA and the carbonyl C in ASP separated by one residue is found to have a mean around 7.1 \AA and standard deviation equal to 1.05 \AA , while the distance between such a pair of atoms across the 20th and 22nd residues in the NMR structure 2GB1 is 4.6293 \AA , which is 0.3707 \AA smaller than the mean minus two standard deviations. More example cases of distance deviations in 2GB1 are given in

Table 1 Deviations of distances in NMR-determined structures^{*}

#R	R1	A1	#R	R2	A2	M	$2 \times \sigma$	D
19	GLU	N	20	ALA	C	5.0	0.8	5.94
20	ALA	CB	22	ASP	C	7.1	2.1	4.63
20	ALA	CB	22	ASP	CA	6.7	1.5	5.09
20	ALA	CB	22	ASP	N	5.6	1.3	4.24
20	ALA	CB	22	ASP	O	7.8	2.5	3.51
21	VAL	N	22	ASP	O	5.9	1.0	4.28
21	VAL	CB	23	ALA	CB	7.2	1.6	9.37
21	VAL	CB	23	ALA	CA	6.7	1.1	8.19
21	VAL	CB	23	ALA	N	5.7	0.9	6.95
22	ASP	CB	23	ALA	C	5.4	0.6	4.69

^{*}Shown in the table are sample atomic pairs (A1 and A2) across some of the residues (R1 and R2) in NMR structure 2GB1 with distances (D) deviating more than twice their standard deviations (σ) from their average distributions (μ) in known protein structures.

Table 1. In fact, in each of 462 NMR structures, similar deviations are found in 2% to 44%, or in an average of 21.98% of the residue pairs that are separated by one or zero residue along the protein backbone. The deviations are not only found among backbone atoms (N, O, C, C_α), but also between backbone (N, O, C, C_α) and side-chain atoms (C_β). In most cases, the residues having such distance deviations are located on exposed parts of the proteins, which is consistent with the fact that the surface residues are usually of high mobility and more difficult to determine by NMR.²

4. Refining NMR Structures

The large deviations of inter-atomic distances in NMR structures from their average distributions in known protein structures are clear indications of modeling errors in NMR structures that are probably due to the lack of proper constraints on the corresponding distances in the NMR data. One possible way to reduce the errors is to confine the distances to their most probable ranges according to their distributions in known protein structures. To test such an approach, the distribution functions for selected cross-residue inter-atomic distances are used to generate a set of bound constraints for the distances, with the lower and upper bounds equal to the mean values of the distances minus and plus twice the standard deviations, respectively. The generated distance bounds are then taken as additional distance constraints to refine a set of NMR structures, including five structures for 1EPH, 1GB1, 1IGL, 2IGG, 2SOB and five for 1CEY, 1CRP, 1E8L, 1ITL, 1PFL. The last five are selected because they have X-ray structures available. The original NMR experimental constraints for the structures are downloaded from BioMagResBank.¹⁹ The structures are refined using the default torsion angle dynamic simulated annealing protocol implemented in CNS.¹⁸ The results obtained with and without additional database distance constraints are examined on the deviations of selected inter-atomic distances from their average distributions, and compared and assessed in terms of several criteria used in NMR modeling, including the acceptance rates of the structures, the RMSD values of the ensembles of structures, and the RMSD values of the structures compared with their X-ray structures (for available ones).

CNS can be used to refine either X-ray or NMR structures. The part for NMR structure refinement contains four steps: connectivity calculation, template generation, annealing, and acceptance test. Connectivity calculation takes the protein sequence as the input and produces a connectivity file for the backbone of the protein. Template generation uses the connectivity file to construct an extended structure (or a group of extended structures) for the protein as the initial structures for annealing. The annealing process has two options, one with simple simulated annealing and another with distance geometry simulated annealing. The latter embeds the structure in 3D by satisfying the distance constraints before doing simulated annealing. The last step, acceptance test, evaluates the structures with a group of acceptance criteria including the satisfaction of various experimental constraints and stereochemistry requirements. In our calculations, we have used the simple simulated annealing option with the database-derived distance constraints provided in the same format as the NOE distance constraints. A structure (or a group of structures) is calculated by minimizing the violations of the experimental and database-derived constraints and the CNS-built in energy potentials.¹⁸

4.1 Correction on distance violations

As summarized in Table 2, after being refined with additional database distance constraints, the numbers of incorrectly distributed cross-residue inter-atomic distances in the structures are clearly reduced. For example, in structure 1GB1, there are 15 residue pairs with 28 cross-residue inter-atomic distances deviating from their average distributions by more than twice the standard deviations, but after being refined with additional database distance constraints, the numbers drop to 11 residue pairs and 14 cross-residue inter-atomic distances. There are exceptional cases when the deviations are

Table 2 Incorrect cross-residue inter-atomic distances*

Protein	#R	DA [†]	NOE [‡]	NMR	NMR+DB
1EPH	53	24	6.7	58/25	48/24
1GB1	56	93	16.5	28/15	14/11
1IGL	67	11	7.8	83/30	65/28
2IGG	64	39	7.4	75/31	29/20
2SOB	103	49	8	143/57	74/41

*Numbers of incorrect distances (outside ± 2 standard deviations) versus numbers of affected residue pairs for structures refined with (NMR+DB) and without (NMR) database distance constraints; [†]DA – dihedral angle constraints; [‡]NOE – NOE distance constraints per residue.

increased after the refinement for some marginal distances, but most of them are decreased even if they may still be larger than two standard deviations.

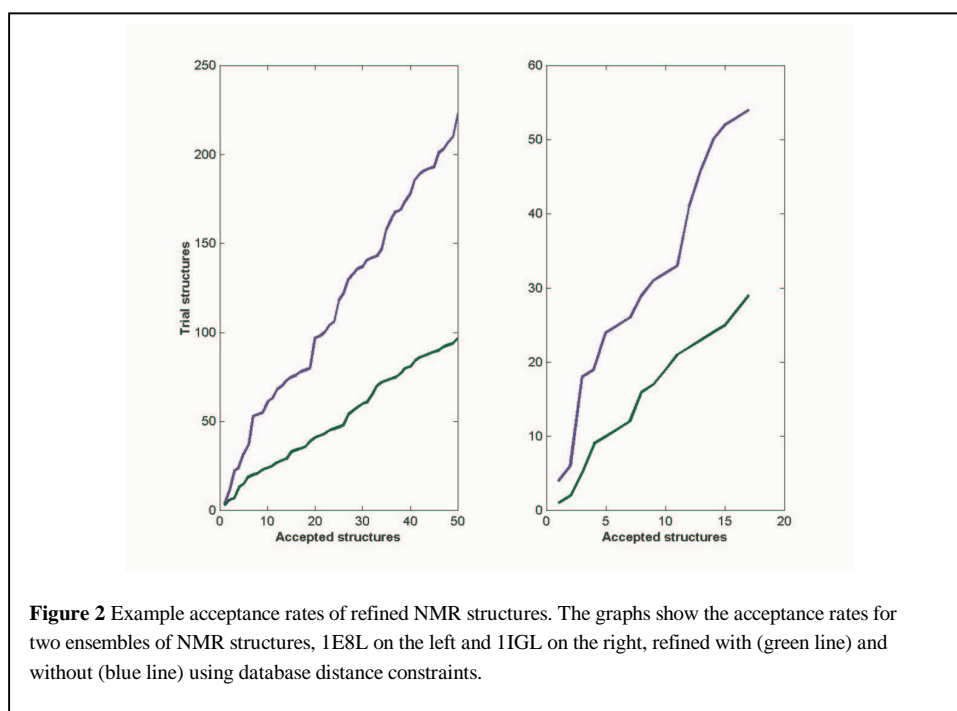
The reason we choose two standard deviations is that there are high percentage of distances in NMR structures having more than two standard deviations as shown in our survey. By constraining these distances within two standard deviations we expect to be able to change the structures more effectively than allowing even larger deviations. On the other hand, using one standard deviation would be too restricted.

Even two standard deviations may exclude some potentially legitimate distances. However, based on our survey, for most of the distance types, the probability to have more than two standard deviations is less than 2% in their database distributions. It is relatively safe to confine the distances in their two standard deviations.

A more natural way to impose the distance constraints may be to implement a potential energy function with the probability distributions of the distances as suggested in Sippl 1990. The distances can then be determined based on the joint probability of the distances or in other words, the energy of the system. This is another line of our investigation, which we will report later elsewhere.

4.2 Acceptance rates

Given an ensemble of accepted NMR structures, the acceptance rate for the ensemble of structures is defined as the number of accepted structures divided by the total number of trial structures including the “rejected” ones. Here, the default acceptance criteria in CNS, including the bond lengths, bond angles, NOE distances, and dihedral angles restraints, are used.¹⁸ A trial structure is accepted if all these requirements are satisfied. With additional database distance constraints, the acceptance rates of the refined NMR structures become much higher than those of the structures obtained with only original NMR distance constraints. As shown in Figure 2, for protein 1E8L, only 97 structures need to be determined to obtain 50 accepted structures when additional database distance



constraints are used, while 223 structures are required if without them. The acceptance rate for protein 1E8L is increased from about 0.25 to more than 0.50. For protein 1IGL, only 29 structures need to be determined to obtain 17 accepted structures if additional database distance constraints are used, while 67 structures are required otherwise. The acceptance rate is increased from about 0.30 to more than 0.60. These increases in efficiency indicate that additional database distance constraints not only help to correct the distance errors in the NMR structures but also improve the performance of the modeling program for obtaining acceptable ensembles of structures.

4.3 RMSD of structural ensembles

The precision of an ensemble of structures determined by NMR usually is measured by the RMSD values of the structures in the ensemble compared with the average structure of the ensemble, and in particular, by the mean and standard deviation of these values.¹⁸

The precision may be overestimated since the ensemble of structures determined by current modeling software may not necessarily contain the whole range of structures determined by the given distance constraints.³ Nevertheless, as shown in Table 3, the means and standard deviations of the RMSD values for the listed ensembles of structures all become smaller after the structures are refined with additional database distance constraints. Note that the RMSD values are calculated in terms of either just backbone atoms or all non-hydrogen atoms. The results are consistent in both calculations.

Table 3 RMSD values of the ensembles of refined NMR structures

Protein	Residue	Data	Means \pm Standard Deviations [*]	
			Backbone [†]	Non-H [‡]
1EPH	53	NMR	2.04 \pm 0.61	2.94 \pm 0.70
		NMR + DB	1.78 \pm 0.40	2.76 \pm 0.54
1GB1	56	NMR	0.45 \pm 0.12	1.04 \pm 0.18
		NMR + DB	0.38 \pm 0.09	0.91 \pm 0.16
1HGL	67	NMR	4.50 \pm 1.52	5.49 \pm 1.55
		NMR + DB	3.81 \pm 1.24	4.70 \pm 1.43
2IGG	64	NMR	2.62 \pm 0.85	3.29 \pm 0.83
		NMR + DB	2.16 \pm 0.90	2.87 \pm 0.85
2SOB	103	NMR	7.25 \pm 1.60	8.06 \pm 1.67
		NMR + DB	5.54 \pm 1.77	6.41 \pm 1.77

^{*}The means and standard deviations of the RMSD values of the structure ensembles refined with and without database distance constraints; [†]RMSD values in terms of backbone atoms; [‡]RMSD values in terms of all non-hydrogen atoms.

4.4 Comparison with crystal structures

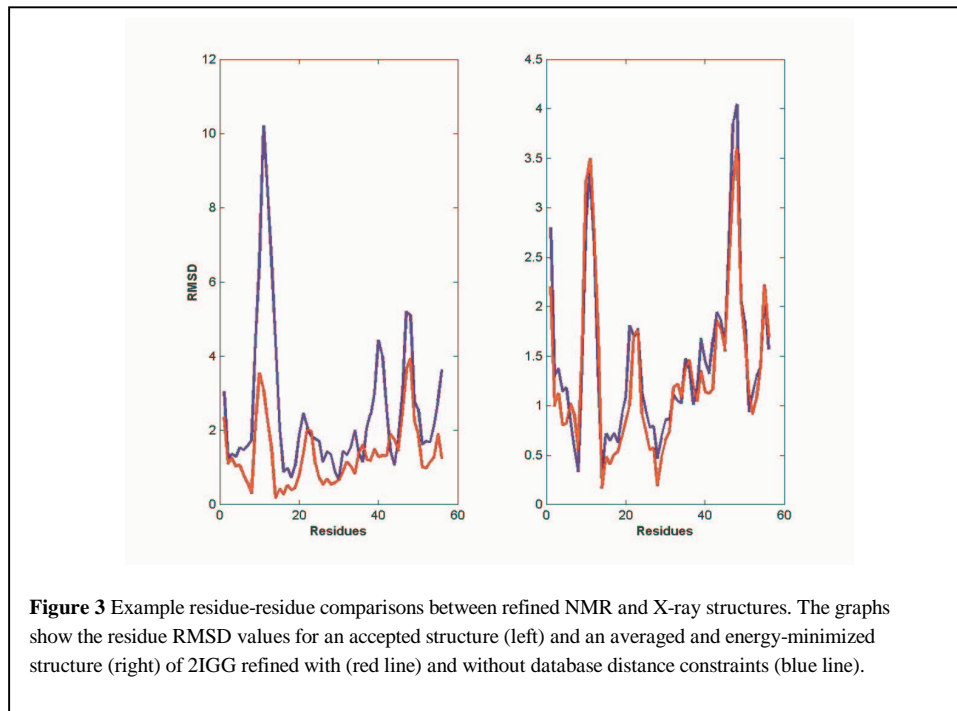
The refined NMR structures for five proteins (1CEY, 1CRP, 1E8L, 1ITL, and 1PFL) are compared with their corresponding X-ray structures for the RMSD values of the pairs of NMR and X-ray structures. Since each protein has an ensemble of NMR structures, the mean and standard deviation of the RMSD values of the member structures are calculated and used as an assessment for the whole ensemble of structures. As shown in Table 4,

Table 4 Refined NMR structures compared to X-ray structures

NMR ID	X-Ray ID	#R	Means \pm Standard Deviations [*]	
			NMR [†]	NMR + DB [‡]
1CEY	3CHY	128	1.85 \pm 0.19	1.80 \pm 0.17
1CRP	1IAQ_A	166	1.77 \pm 0.29	1.60 \pm 0.26
1E8L	193L	129	2.05 \pm 0.22	2.02 \pm 0.19
1ITL	1RCB	129	2.88 \pm 0.76	2.79 \pm 0.21
1PFL	1FIK	139	1.66 \pm 0.07	1.65 \pm 0.07

^{*}The means and standard deviations of the RMSD values for the ensembles of NMR structures compared with their X-ray structures; [†]Refined with only NMR distance constraints; [‡]Refined with NMR and database distance constraints.

both means and standard deviations of the RMSD values for the ensembles of structures refined with additional database distance constraints are clearly smaller than those refined without them, indicating strongly that the structures agree more closely with the X-ray



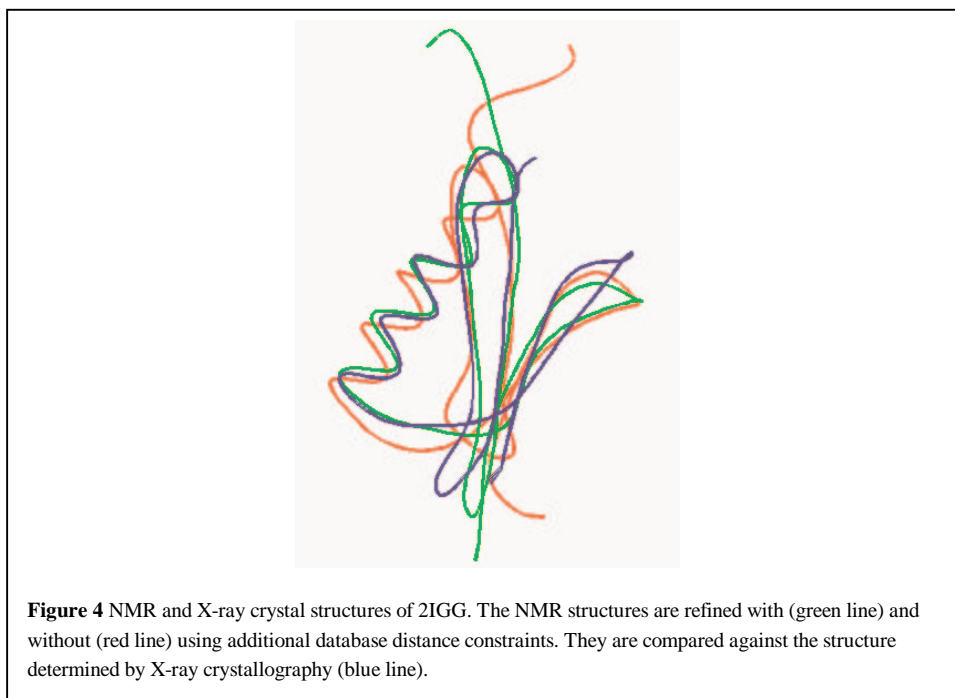
structures after being refined with the additional database distance constraints.

The RMSD values are average measures on overall structural differences. Therefore, the small RMSD differences between structures refined with or without database distance constraints as shown in Table 4 may imply large local structural differences. A detailed residue-residue comparison for a particular protein 2IGG²⁰ is illustrated in Figure 3, where the RMSD values for all corresponding pairs of residues of NMR and X-ray structures are plotted. The two curves in each graph show the residue RMSD values for two NMR structures of 2IGG, one refined with original NMR distance constraints and another with additional database distance constraints, when they are compared with the corresponding X-ray structure. The graph on the left is for two accepted structures randomly chosen from their corresponding ensembles of structures. The graph on the right is for two averaged and energy-minimized structures. Both graphs demonstrate the differences between NMR structures refined with and without additional database distance constraints, although the differences in the latter graph are not as large as the former. Figure 4 further displays in 3D graphics the differences among three structures determined for 2IGG, one refined with NMR distance constraints only, one with additional database distance constraints, both without being averaged and energy-minimized, and one determined by X-ray crystallography. The picture shows clearly that the NMR structure determined by using additional database distance constraints agree

with the X-ray structure better in many regions than the one without using additional database distance constraints, especially in loops and tails, where the structure is not well defined by the NMR experimental data.²⁰

5. Discussions

The analysis of NMR-determined protein structures by comparing selected cross-residue inter-atomic distances with the distributions of the distances in known protein structures can always provide a statistical estimate of the accuracy of the NMR structures. While some of the deviations of the inter-atomic distances in NMR structures may be attributable to the additional flexibilities of the NMR structures in solution beyond the crystalline state, many of them must originate in modeling errors, as justified indirectly by the higher acceptance rates and smaller RMSD values of the ensembles of structures when selected distances are confined in high probability regions of their distributions. However, how to distinguish the variations of the distances due to the flexibilities of the NMR structures from those caused by modeling errors is not so clear and remains a question to pursue in future studies. Several approaches may be taken to determine the fluctuations of NMR structures, such as based on the order parameters or temperature factors that can be obtained from NMR or X-ray diffraction data, respectively, or using the Gaussian Network Model²¹ or the Normal Mode Analysis.²² If the fluctuations of NMR structures can be determined, the structural variations inconsistent with the



fluctuations may be better targeted for refinement.

While a distance constraint can be derived for every selected pair of cross residue atoms based on the distribution of the distance in known protein structures, not all the

constraints are necessary for the refinement of a given NMR structure since some distances may not necessarily be incorrect even if they deviate significantly from their average distributions. In this work, the distance constraints for all pairs of atoms N, C_α, C, O, C_β in nearby residues along the protein backbone are derived based on their distribution functions. However, only four such constraints (C_β-C_β, C-C_β, N-C_α, O-C_β) are selected for pairs of neighboring residues and one (C_β-R-C_β) for every two separated residues, where R represents the separating residue. In general, the constraints may be most effective for distances or interactions in regions that are underdetermined by NMR experimental data.

On the other hand, the atom types used can certainly be extended to include more side-chain atoms and longer-range interactions. In general, the backbone and other non-hydrogen atoms are perhaps most likely to have distances among them disagreeing with their distributions in known protein structures, since the non-hydrogen atoms usually do not have as much distance data available as hydrogen atoms and therefore cannot be determined as directly and accurately. Indeed, some initial test results show that for many NMR structures, the RMSD values of the ensembles of structures compared with the corresponding X-ray structures in terms of all non-hydrogen atoms are much larger than the RMSD values of the ensembles in terms of hydrogen atoms, while the RMSD values of the ensembles in terms of only backbone atoms are in between the two cases (data not shown).

Although it may be beyond the scope of this investigation, it would be ideal if there were further experimental supports for the correctness of the refined structures using the database derived distance constraints, other than just the computational criteria. In the absence of sufficient experimental evidences, it should be cautioned that the refined structures are partially knowledge-based, and may not reflect the true structures of the proteins. The results in this paper try to demonstrate the reliability of the knowledge-based structures in terms of several standard measures. Additional direct and indirect justifications have also been obtained and reported elsewhere. For example, when the (short-range) database derived distance constraints are introduced, the experimental torsion angle constraints derived from J-coupling can actually be reduced without affecting the quality of the refined structures, showing that the database constraints can be used to enhance or even replace some of the experimental constraints.²³ The database derived distance constraints have also been applied specifically to improving the structures of several critical loops of the human prion variant E200K, which are not well defined in their NMR models because of the lack of enough experimental data. The improved loops have been validated through careful comparisons with both NMR and X-ray experimental structures of several other wild type prions and mutants.²⁴

Acknowledgments

The authors would like to thank Peter Vedell and Di Wu for their helpful discussions on the paper. The work is partially supported by the research funds from the Department of Mathematics, the Graduate Program on Bioinformatics and Computational Biology, and the Lawrence Baker Center for Bioinformatics and Biological Statistics at Iowa State University.

References

1. Creighton, T.E. *Proteins: Structures and Molecular Properties*, 2nd Edition. Freeman and Company (1993).

2. Doreleijers, J. F., Rulmann, J. A. C., and Katein, R. Quality assessment of NMR structures: A statistical survey. *J. Mol. Biol.* **281**. 149-164 (1998).
3. Spronk, C. A. E. M., Natuurs, S. B., Bonvin, A. M. J. J., Krieger, E., Vuister, G. W., and Vriend, G. The precision of NMR structure ensembles revisited. *Journal of Biomolecular NMR* **25**. 225-234 (2003).
4. Tjandra, N. and Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* **278**. 1111-1114 (1997).
5. Clore, G. M. and Gronenborn, A. M. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. USA* **95**. 5891-5898 (1998).
6. Prestegard, J. H. New techniques in structural NMR – an-isotropic interactions. *Nat. Struct. Biol.* **5 Suppl.** 517-522 (1998).
7. Wang, L. and Donald, B. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Journal of Biomolecular NMR* **29**. 223-242 (2004).
8. Kuszewski, J., Gronenborn, A. M., and Clore, G. M. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science* **5**. 1067-1080 (1996).
9. Grishaev, A. and Bax, A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *J. Am. Chem. Soc.* **126**. 7281-7292 (2004).
10. Bourne, P. E. and Weissig, H. *Structural Bioinformatics*. John Wiley & Sons, Inc. (2003).
11. Miyazawa, S. and Jernigan, R. L. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**. 534-552 (1985).
12. Miyazawa, S. and Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* **256**. 623-644 (1996).
13. Sippl, M. J. Calculation of conformational ensembles from potentials of mean force. *J. Mol. Biol.* **213**. 859-883 (1990).
14. Sippl, M. J. and Weitckus, S. Detection of native-like models for amino acid sequence of unknown three-dimensional structure in a database of known protein conformations. *Proteins: Structure, Function, and Genetics* **13**. 258-271 (1992).
15. Rojnuckarin, A. and Subramaniam, S. Knowledge-based potentials for protein structure. *Proteins: Structure, Function, and Genetics* **36**. 54-67 (1999).
16. Wall, M. E., Subramaniam, S., and Phillips, Jr. G. N. Protein Structure Determination Using a Database of Inter-Atomic Distance Probabilities. *Protein Science* **8**. 2720-2727 (1999).
17. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, Shindyalov, L. N., and Bourne, P. E., The Protein Data Bank. *Nuc. Acid. Res.* **28**. 235-242 (2000).
18. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. Crystallography and NMR System: A new software suite for macromolecular structure determination. *Acta Cryst.* **D54**. 905-921 (1998).
19. Doreleijers, J. F., Mading S., Maziuk D., Sojourner K., Yin, L., Zhu, J., Makley, J. L., and Ulrich, E. L. BioMagResBank database with sets of experimental NMR

- constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *Journal of Biomolecular NMR* **26**. 139-146 (2003).
20. Lian, L. Y., Derrick, J. P., Sutcliffe, M. J., Yang, J. C., and Roberts, G. C. K. Determination of the solution structures of domain II and III of protein G from *Streptococcus* by ^1H nuclear magnetic resonance. *J. Mol. Biol.* **228**. 1219-1234 (1992).
 21. Bahar, I., Atilgan, A. R., Demirel, M. C., and Erman, B. Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability. *Phys. Rev. Lett.* **80**. 2733-2736 (1998).
 22. Levitt, M., Sander, C., and Stern, P. S. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease, and lyozyme. *J. Mol. Biol.* **181**. 423-447 (1985).
 23. Cui, F., Jernigan, R., and Wu, Z. Enhancement of Torsion Angle Constraints in NMR Structure Refinement via Database Derived Distance Constraints, submitted, 2005.
 24. Cui, F., Mukhopadhyay, K., Young, W., Jernigan, R., and Wu, Z. Improvement of Under-Determined Loop Regions of Human Prion Protein by Database Derived Distance Constraints, submitted, 2005.



Bioinformatics.

Feng Cui received a Bachelor degree in Medical Sciences from Hunan Medical University of China in 1995 and a Master's degree in Biology from Trument State University of United States in 2000. He was a research associate in the National Laboratory of Medical Genetics in China from 1995 to 1999. He is currently a Ph.D. candidate in the Program on Bioinformatics and Computational Biology of Iowa State University. He is the recipient of the 2004-2005 ISU-Pioneer Fellowship in



molecular biology of Iowa State University and also the Director of the ISU Baker Center for Bioinformatics and Biological Statistics.

Robert Jernigan received his B.S. degree in Chemistry from California Institute of Technology in 1963 and his Ph.D. degree in Physical Chemistry from Stanford University in 1968. He did postdoctoral research at University of California at San Diego from 1968 to 1970. He joined NIH from 1970 to 2002 and was the Deputy Chief and Chief of the Laboratory of Experimental and Computational Biology and Section of Molecular Structures of NCI. He is currently professor of biochemistry, biophysics, and



Zhijun Wu received his Bachelor degree in Computer Science from Huazhong University of Science and Technology, China, and his Ph.D. degree in Computational Mathematics from Rice University, United States, in 1982 and 1991, respectively. He did postdoctoral research at Cornell University and Argonne National Lab from 1991 to 1997. He is currently Associate Professor in the Department of Mathematics of Iowa State University and also a joint faculty member of the ISU Program on Bioinformatics and Computational Biology.