KNOWLEDGE-BASED VERSUS EXPERIMENTALLY-ACQUIRED DISTANCE AND ANGLE CONSTRAINTS FOR NMR STRUCTURE REFINEMENT

By

Feng Cui
Robert Jernigan
and
Zhijun Wu

IMA Preprint Series # 2207

(May 2008)

INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS
UNIVERSITY OF MINNESOTA
400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455–0436
Phone: 612/624-6066 Fax: 612/626-7370
URL: http://www.ima.umn.edu

Knowledge-Based versus Experimentally-Acquired Distance and Angle Constraints for NMR Structure Refinement

Feng Cui[1], Robert Jernigan[2,3], and Zhijun Wu[2,4*]

[1]Laboratory of Cell Biology
National Cancer Institute, National Institutes of Health
Bethesda, Maryland 20892

[2]Program on Bioinformatics and Computational Biology
[3]Department of Biochemistry, Biophysics, and Molecular Biology
[4]Department of Mathematics
Iowa State University
Ames, Iowa 50011

Running title: NMR Constraints and Structure Quality

***Corresponding Author:** Zhijun Wu, Program on Bioinformatics and Computational Biology and Department of Mathematics, Iowa State University, Ames, Iowa, 50011, U.S.A., **Phone:** 515-294-8165, **Fax:** 515-294-5454, **Email:** zhijun@iastate.edu

Abstract NOE distance constraints and torsion angle constraints are major conformational constraints for NMR structure refinement. In particular, the number of NOE constraints has been considered as an important determinant for the quality of the NMR structures. Of course, the availability of torsion angle constraints is critical for the formation of correct local conformations as well. In our recent work, we have shown how a set of knowledge-based short-range distance constraints can also be utilized for NMR structure refinement, as a complementary set of conformational constraints to the NOE and torsion angle constraints. In this paper, we show the results from a series of structure refinement experiments by using different types of conformational constraints, NOE, torsion angle, or knowledge-based constraints, or their combinations, and make a quantitative assessment on how the experimentally-acquired constraints contribute to the quality of the structural models and whether or not they can be combined with or substituted by the knowledge-based constraints. We have carried out the experiments on a small set of NMR structures. Our preliminary calculations have revealed that the torsion angle constraints contribute substantially to the quality of the structures, but require to be combined with the NOE constraints to be fully effective. The knowledge-based constraints can be functionally as crucial as the torsion angle constraints, although they are statistical constraints after all and are not meant to be able to replace the latter.

Keywords: NOE distance constraints, torsion angle constraints, knowledge-based distance constraints; precision and accuracy of NMR structures; NMR structure refinement

Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is one of the major experimental techniques for protein structure determination (Wüthrich, 1986). The latest Protein Data Bank (PDB) release (February 2007) showed that approximately 17% of total entries (6063 out of 35095) are determined by NMR spectroscopy. One of the advantages of NMR over other structure determination approaches such as X-ray crystallography is that by using NMR, proteins can be studied in solution, i.e., an environment similar to that in living cells, and both structures and motions of a protein can be determined in NMR experiments. However, similar to other techniques, due to limited experimental data, the structures determined by NMR are not necessarily always as accurate as desired and further refinement of the structures is often required (Wüthrich, 1986).

The quality of NMR structures depends certainly on the Nuclear Overhauser Effects (NOE) data, which constitute the main source of geometric information obtained in NMR experiments. The NOE data, however, cannot be obtained in a complete and accurate manner. The NOE intensity between a pair of magnetically interacting protons is inversely proportional to the sixth power of the distance between the two atoms, which can be detected only if the two atoms are in a short distance (<5.0 Å). Thus, the distances may be estimated through NOE only when the pairs of protons are spatially very close. In addition, the estimation of the distances is semi-quantitative: the distances are assigned to different ranges (e.g., <2.5, <3.5, or <5.0 Å), depending on the strength of the NOE signals (Creighton, 1993). With these distance constraints derived from NOE (along with torsion angle constraints from J-couplings and other constraints), an ensemble of structures instead of a single structure within the ranges of the constraints is generated. Additional experimental (*e.g.*, residue dipolar couplings) (Clore *et al.*, 1998; Tjandra *et al.*, 1997) and theoretical constraints (*e.g.* knowledge-based constraints and potential functions) (Cui *et al.*, 2005; Grishaev and Bax, 2004; Kuszewski *et al.*, 1996; Wall *et al.*, 1999; Wu *et al.*, 2007) may be applied for further refinement of the structures.

An NMR structure ensemble can be evaluated in terms of the precision and accuracy of the ensemble (Spronk *et al.*, 2003). In a given ensemble, the structures are fluctuated around their 'mean' structure. The degree of fluctuation varies in different local regions, depending on how well the local structures are determined. The precision estimates the variation of each atomic position around its 'mean' position, and thus is usually expressed as the average coordinate root mean square deviation (RMSD) of the structures in the ensemble against their average structure. On the other hand, the accuracy of the ensemble measures the closeness of the structures in the ensemble to the 'true' structure, for which the X-ray structure of the same molecule is often used as a reference. It is usually expressed as the average RMSD of the structures in the ensemble against their 'true' structure.

The number of NOE constraints per residue is shown to be the most important factor shaping the precision and accuracy of NMR structure ensembles (Clore *et al.*, 1993; Liu *et al.*, 1992). As the number of NOE constraints per residue is increased, the precision and accuracy of an ensemble are improved significantly until a plateau is reached at about 15 NOE constraints per residue. However, 'real' NMR data are unlikely to have such a high number of NOE constraints per residue. For example, an analysis on 97 NMR structures deposited in PDB (Doreleijers *et al.*, 1999; Doreleijers *et al.*, 1998) revealed that 81 out of the 97 structures (84%) have less than 15 NOE constraints per residue. For the whole set of 97 structures, the average number of NOE constraints per residue is 10.5±0.4, far below the 'ideal' number of 15 NOE constraints per residue. The shortage of NOE data implicates room in NMR structures for improvement in precision and accuracy, especially in the structures with low numbers of NOE constraints per residue, by introducing other types of experimental or theoretical constraints. While alternative constraints have been applied to refining NMR-derived structures successfully (Clore *et al.*, 1998; Cui *et al.*, 2005; Grishaev and Bax, 2004; Kuszewski *et al.*, 1996; Tjandra *et al.*, 1997), it is not so clear how much they can contribute to the quality of the resulting structures. A rigorous assessment in this respect would help to identify effective types of constraints complementary to NOE.

In this paper, we take a small step towards this goal by assessing the contribution of NOE, torsion angle, and a special class of knowledge-based distance constraints derived from structural databases in Cui *et al.*, 2005 to the precision and accuracy of NMR structure ensembles. Of the three types of constraints, NOE and torsion angle constraints are most commonly used experimental data. The knowledge-based distance constraints are derived from the distributions of the distances (between two atoms across two residues in sequence) in databases of known protein structures. Cui *et al.*, 2005 showed that the use of these distance constraints may improve the quality of NMR structures in terms of both precision and accuracy, compared with the structures refined only by experimental constraints, although the improvement has to be examined carefully because the constraints are knowledge-based statistical constraints, not experimental constraints that can be trusted relatively completely.

In this work, a set of NMR structures is selected, with varying numbers of NOE constraints per residue (from 6.7 to 19.7, see Table 1). Because the selected structures have a wide range of NOE constraints per residue, we consider them as good representatives of NMR structures in terms of their numbers of NOE constraints per residue (Doreleijers *et al.*, 1999; Doreleijers *et al.*, 1998). The NOE and torsion angle constraints of selected structures are downloaded from PDB (Berman *et al.*, 2000) or BioMagResBank (Doreleijers *et al.*, 2003). A set of database-derived distance constraints are generated with the lower and upper bounds on the distances equal to the 'mean' distances in their database distributions minus and plus two standard deviations (see Methods). The selected structures are then refined separately by NOE constraints only, NOE and torsion angle constraints, or NOE and database-derived constraints, using the standard structure refinement protocols of Crystallography and NMR System (CNS) (Brunger *et al.*, 1998). The precision and accuracy of the resulting ensembles are calculated and compared. Our results show that the contribution of the database-derived distance constraints to both precision and accuracy is comparable to that of the torsion angle constraints, which supports positively the use of the database-derived distance constraints. However, it does not mean that the database-derived distance constraints can be applied equally as the experimental constraints, because they are statistical constraints after all, and they should be applied with caution and only when there are not enough experimental constraints.

Methods

Collection of structures and distances

A total of 2090 X-ray structures with resolution of 2.0 Å or higher and sequence similarity of 70% or less were downloaded from PDB (with the Advanced Search scheme), for the computation of the distributions of the inter-atomic distances in known proteins. No additional filters were used. Using 70% sequence similarity cutoff was a bit arbitrary. There was a 'historical' reason that we used this cutoff value: The structures were collected in 2003 – 2004. At that time, only two cutoff values were available, 70% and 90%. To be conservative, we chose 70%. The coordinates of the atoms in all the segments were extracted except for those with alternative locations (symbolized as ALT in PDB), or in unknown residue types (symbolized as UNK), or identified as heterogen atoms (symbolized as HETATM). The distances between selected heavy atoms in two residues separated by one or zero residue were calculated. The distance data obtained was used for NMR structure refinement.

Calculation of distance distributions

Let D be the distance between two given atoms, A_1 and A_2 the types of the two atoms, R_1 and R_2 the types of the two residues the two atoms are associated with, respectively, and S the number of residues between R_1 and R_2 in sequence. Then, the distribution of the distance D between atoms A_1 in R_1 and A_2 in R_2 with R_1 and R_2 being separated by S residues in sequence can be represented by using a distribution function $P[A_1,A_2,R_1,R_2,S](D)$. For each set of A_1, A_2, R_1, R_2, and S, the corresponding distances in the downloaded structures were collected and grouped into a set of uniformly divided distance intervals $[D_i, D_{i+1}]$, where $D_i = 0.1 * i$ Å, i = 0, 1, …, n. The function value, $P[A_1,A_2,R_1,R_2,S](D)$, for any D in $[D_i, D_{i+1}]$, was then defined to be the number of distances in $[D_i, D_{i+1}]$ divided by the number of distances in all the intervals. For each distribution function P, the mean μ and standard deviation σ were also calculated and stored. Note that in this study, only five different types of atoms were considered. They were the amide N, the carbon C_α, and the carbonyl C and O along the backbone and the carbon C_β in the side chain. The residue types included all twenty different amino acid types. The separation S was either one or zero. So there are total 5 * 5 * 20 * 20 * 2 = 20,000 possible distance types.

Two example distance distribution functions are plotted in Figure 1. The first figure is for the distances between atom C in arginine at position *i* and atom O in isoleucine at position *i*+1 along a polypeptide chain (Figure 1a). The second one is for the distances between atom C_β in alanine at position *i* and atom N in leucine at position *i*+2 along a polypeptide chain (Figure 1b). Both distributions appear to be Gaussian-like.

Distance constraints and refinement protocols

For each distribution function P, the mean (μ) plus and minus 2 standard deviations (σ) were used as the upper and lower bounds for the corresponding distance D. For a protein to be refined, a selected set of distance bounds was generated and stored in the same format as the NOE distance constraints. A standard torsion angle dynamic simulated annealing protocol implemented in CNS was used for structure refinement.

Results

Seven different NMR structures with different numbers of NOE constraints per residue were selected as test cases (Table 1). They included the structures for 1E8L, 1EPH, 1GB1, 2IGG, 1CEY, 1CRP, and 1PFL. The last four structures were selected because their corresponding X-ray structures were available for comparison. All the structures were refined using the standard torsion angle dynamic simulated annealing protocols of CNS with default settings. The resulting structures, obtained separately with NOE constraints only, NOE and torsion angle constraints, NOE and database-derived distance constraints, were compared and assessed in terms of several standard measures used in NMR modeling, including the RMSD values of the ensembles of structures (precision), and the RMSD values of the structures compared with their X-ray reference structures (accuracy). Finally, the structure ensemble of 1CRP (Kraulis *et al.*, 1994) with the highest number of NOE constraints per residue (19.7 NOE constraints per residue) among selected structures, was assessed as a case study in terms of residue-residue RMSD values and the Ramachandran plots (Laskowski, 1993; Morris *et al.*, 1992).

Note that 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 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) was calculated by minimizing the violations of the experimental and database-derived constraints and the CNS built-in energy potentials.

NOE and heavy atom fluctuations

The NOE data measures the distances for pairs of spatially closed hydrogen atoms, *e.g.*, HN and HA (<5 Å) that are linked to heavy atoms of the protein backbone through covalent bonds. Thus, the distance constraints derived from NOE are imposed directly to the hydrogen atoms but not to the backbone heavy atoms. It is therefore expected that the hydrogen atoms (HN and HA) are less fluctuated compared to the backbone atoms and other non-hydrogen side-chain atoms. To see if the number of NOE constraints per residue influences the degree of fluctuation, we refined the structures for 1E8L, 1GB1, 1EPH and 2IGG using experimental constraints (NOE and torsion angle constraints) and computed the ensemble RMSD values for HN and HA, backbone atoms, and all non-hydrogen atoms (Table 2). The first two structures have high numbers of NOE constraints per residue, while the last two have low numbers of NOE constraints per residue (Table 1).

Table 2 show that the backbone and non-hydrogen atoms are indeed more fluctuated compared to HN and HA atoms. In addition, the degree of fluctuation is correlated to the number of NOE constraints per residue. For example, 1EPH, 1GB1, and 2IGG are 'small' proteins of similar size (~60 aa) but with different numbers of NOE constraints per residue (6.7, 16.5, and 7.4 NOE constraints per residue, respectively). The ensemble RMSD for the backbone atoms of 1GB1 (0.45 Å), for example, is significantly lower than those of 1EPH (2.04 Å) and 2IGG (2.62 Å), indicating that the backbone atoms of 1GB1 are much better determined due to the high number of NOE constraints per residue. On the other hand, the 'big' protein 1E8L (129 aa) with 13.1 NOE constraints per residue exhibits a higher ensemble RMSD value (2.30 Å) compared to that of 1EPH (2.0 Å) with 6.7 NOE constraints per residue. This may be due to the size of the protein (1E8L is twice as big as 1EPH) and the (possibly non-uniform) distribution of the NOE constraints along the polypeptide chain. The increase in the degree of fluctuation in backbone, especially for proteins with low numbers of NOE constraints per residue, may show the flexibilities of the structures, but may also suggest some modeling errors caused by the lack of experimental constraints.

Note that the RMSD values we calculated for the structural ensembles are very different from their originally published RMSD values. This is because (1) the original values were calculated only for ordered regions while we considered the whole structures; (2) the original values were tuned for each individual structure while we refined all the structures using the same software and the same default setting; (3) we only applied NOE and torsion angle constraints for the refinement because we were interested in assessing only these two types of constraints in this work while most of the structures were originally refined with additional constraints such as the residue dipolar coupling constraints.

The reason we included all the regions of the structures, ordered and disordered, in our RMSD calculations is that we are interested in the fluctuations of the structures in ordered as well as disordered regions and their correlations with the availabilities of the NOE constraints. Our belief is that the availabilities of the NOE constraints affect the fluctuations of the whole structures, which is justified by the results in Table 2. The reason we applied the same software with the same default setting to the refinement of all the structures is that we wanted to see the influence on the fluctuation of the structures from the availability of the NOE constraints under the same software condition. The reason we only considered the NOE and torsion angle constraints is that we focused on analyzing these constraints and combining them with knowledge-based constraints. We are certainly interested in including other types of constraints as well, which we will definitely consider in our future investigation.

Distance deviations of NMR structures

To see the deviations of the distances in NMR structures from their average distributions in all known structures, we derived the reference distributions of the distances (mean (μ) ± 2 × standard deviations (σ)) for certain pairs of atoms from a database of high resolution X-ray structures (see Methods). We then compared the distances in NMR-determined structures with their reference distributions to see if they were beyond the ranges of μ ± 2σ. A survey on 462 averaged and energy-minimized NMR structures deposited in PDB showed that in each of the 462 structures, on average, 22% of all pairs of residues that are separated by zero or one residue have inter-atomic distances deviated by more than 2 standard deviations from their 'mean' distances (Cui *et al.*, 2005).

The large deviations of inter-atomic distances in NMR structures from their average distributions in known protein structures are clear indications of the differences of NMR structures from average protein structures. Again, the differences may come from the structural fluctuations revealed by NMR or the modeling errors caused by the lack of constraints for some pairs of atoms. The latter is indeed possible because we also observed a clear decrease in the deviations with increasing the availability of the constraints, either experimentally-acquired or database-derived. For example, we refined the structures for the seven selected structures using NOE constraints alone (NOE), NOE and torsion angle constraints (NOE+TOR), and NOE and database-derived distance constraints (NOE+DB_DIST), and then counted the deviated inter-atomic distances and affected residue pairs in the averaged and energy-minimized structures. We were able to see, as shown in Table 3, that the structures refined by NOE constraints alone (NOE) generated a lot of deviated distances between heavy atoms, but the number of such distances could be significantly reduced by adding torsion angle constraints (NOE+TOR) or knowledge-based distance constraints (NOE+DB_DIST). Also, this reduction in the number of deviated distances appeared to be negatively correlated with the number of NOE constraints per residue: the higher the number of NOE constraints per residue, the lower the number of deviated distances that could be reduced by additional constraints. For example, the averaged and energy-minimized structure of the 1EPH ensemble (6.7 NOE constraints per residue) had 244 largely deviated distances, which were dramatically reduced to 58 and 55 after adding torsion angle and database-derived constraints, respectively. However, for 1E8L (with a high number of NOE constraints per residue, 13.1 NOE constraints per residue), the reduction in deviated distances was only about 30% (Table 3). This indicates that the structures with a higher number of NOE constraints per residue may have less room for improvement in terms of reducing largely deviated distances. On the other hand, the contributions of the torsion angle and database-derived distance constraints to the reduction of deviated distances are clearly comparable. For example, 1E8L had 93 and 86 deviated distances after being refined by NOE plus torsion angle constraints and by NOE plus knowledge-based distance constraints, respectively. Together, our results showed that the inclusion of torsion angle and knowledge-based distance constraints helped to reduce the number of deviated distances in the refined structures. Also, interestingly, the 1GB1 ensemble had no structures accepted by CNS if refined using NOE constraints only, but could be refined significantly after including the torsion angle or knowledge-based distance constraints, showing that the additional constraints were critical in refining the structures even if sufficient NOE constraints were provided.

The precision of NMR structural ensembles

As we have mentioned in the introduction section, the precision of an NMR structural ensemble is expressed as the average coordinate RMSD of the structures in the ensemble. There are two ways to calculate the ensemble RMSD: 1) Compute the RMSD values for all the pairs of structures in the ensemble first, and then, average them. 2) Compute the RMSD values for all the structures in the ensemble against their average structure first, and then, average them. We used the second approach, which is one of the default approaches in CNS, in our calculations. To see how much the ensemble can be improved in terms of precision by introducing torsion angle or database-derived distance constraints, we calculated the RMSD of the resulting ensembles refined separately by NOE constraints only (NOE), NOE and torsion angle constraints (NOE+TOR), and NOE and database-derived distance constraints (NOE+DB_DIST) (Table 4).

Table 4 shows that in general the precision of the structural ensembles can be improved significantly after including torsion angle or knowledge-based distance constraints. In addition, the amount of improvement is correlated with the number of NOE constraints per residue. For example, in case of 1EPH (6.7 NOE constraints per residue), the inclusion of torsion angle constraints or database-derived distance constraints decreased the ensemble RMSD for the backbone atoms by up to 23% (2.65 Å for NOE and 2.04 Å for NOE+TOR) and 21% (2.08 Å for NOE+DB_DIST), respectively. The use of the two types of additional constraints also greatly improved the precision for all the non-hydrogen atoms (14% for NOE+TOR and 17% for NOE+DB_DIST). Apparently here, the torsion angle or database-derived distance constraints provided proper constraints for the conformations of 1EPH NMR structures, which were not well defined by insufficient NOE constraints (6.7 NOE constraints per residue) in the first place. On the other hand, for proteins with high numbers of NOE constraints per residue, the use of the torsion angle or database-derived distance constraints also significantly increased the precision of the resulting ensembles. For example, the improvement on the precision of the 1E8L ensemble (13.1 NOE constraints per residue) reached 3.8% by using torsion angle constraints (2.39 Å for NOE and 2.30 Å for NOE+TOR) and 7.1% by using knowledge-based distance constraints (2.22 Å for NOE+DB_DIST). Overall, the contribution of the database-derived distance constraints to the increase of the ensemble precision was comparable to that of the torsion angle constraints.

Note that the average structure of the NMR ensemble that includes, in our case, 50 accepted structures by the default setting, is given by CNS as a standard output. The detailed description of the calculation can be found in the manual of CNS. We did not implement our own method to measure the dispersion of a given ensemble. Instead, we used this standard setting of CNS for the following reasons. First, the method is well-defined and accepted by the whole community. Second, the method is an integral part of the CNS software. It will be convenient for others to reproduce our results or compare their results with ours.

The accuracy of NMR structural ensembles

The accuracy of an NMR structure ensemble can be expressed as the average RMSD of the structures in the ensemble against the reference X-ray structure of the same molecule. We selected four NMR structures (1CEY, 1CRP, 1PFL and 2IGG) to evaluate the accuracy of the ensembles because their X-ray structures are available. The four structures were chosen deliberately: two of them (1CEY and 2IGG) have low numbers of NOE constraints per residue (~7 NOE constraints per residue), while the other two (1CRP and 1PFL) have high numbers of NOE constraints per residue (19.7 and 12.9 NOE constraints per residue, respectively), comparable to the set used above (1E8L, 1EPH, 1GB1 and 2IGG) (see Table 1). The structures were refined separately by using NOE constraints only (NOE), NOE and torsion angle constraints (NOE+TOR), or NOE and database-derived distance constraints (NOE+DB_DIST). The accuracy of the resulting ensembles was calculated (Table 5).

Table 5 shows that the use of the torsion angle constraints and database-derived distance constraints can substantially improve the accuracy of the ensembles, especially for the ones with low numbers of NOE constraints per residue. For example, for 1CEY (7.2 NOE per residue), the torsion angle constraints and database-derived distance constraints helped to increase the accuracy of the structural ensemble by 7% (1.99 Å for NOE and 1.85 Å for NOE+TOR) and 8% (1.84 Å for NOE+DB_DIST), respectively. For 1CRP (19.7 NOE per residue), however, the accuracy increased by only 3% and 5% after including the torsion angle and database-derived constraints, respectively. The contribution of the two different types of constraints to the accuracy of the ensembles was comparable (Table 5). Clearly, the improvement on the accuracy of the structural ensembles was not as significant as that for the precision. Nevertheless, the accuracy, evaluated by RMSD, is a measure on overall structural differences. The improvement of the structures at the local level may still be substantial, which can be analyzed case by case in practice (see below).

A detailed analysis of 1CRP: a case study

The structure 1CRP with an extremely high number of NOE constraints per residue (19.7 NOE constraints per residue) was chosen for a detailed analysis, which showed a negligible improvement on accuracy by using additional constraints (Table 5). We first analyzed the Ramachandran plots of the averaged and minimized structures of the 1CRP ensembles refined by using NOE constraints alone ('NOE' structure), NOE and torsion angle constraints ('NOE+TOR' structure), or NOE and database-derived distance constraints ('NOE+DB_DIST' structure).

Table 6 shows that both 'NOE+TOR' and 'NOE+DB_DIST' structures have better stereo-chemistry properties at the local level, compared to the 'NOE' structure: the number of residues in the disallowed regions was greatly reduced, while the number of residues in the most favoured regions and additionally allowed regions was increased. On the other hand, the plots for the 'NOE+TOR' structure and the 'NOE+DB_DIST' were not completely comparable in this particular case: the 'NOE+DB_DIST' structure had no residues in the disallowed regions but more residues in the most favoured regions, compared to the 'NOE+TOR' structure. This difference may be worth a detailed comparison as shown below.

The residue-residue RMSD values for the 'NOE' (magenta), the 'NOE+TOR' (green), and the 'NOE+DB_DIST' (blue) structures of 1CRP against the reference X-ray structure (1IAQ Chain A) are plotted (Figure 2). The RMSD values for the fragments from residues 30 to 37 and from residues 61 to 69 are zero because of the absence of the coordinate data in the X-ray structure (Spoerner *et al.*, 2001). The RMSD values for residues 59 and 60 are set to zero because the NMR and X-ray structures are significantly deviated. Among the three structures, the 'NOE+DB_DIST' structure (blue) has the smallest RMSD in several fragments between residues 7 and 10, between residues 47 and 50, and between residues 142 and 153, which are mainly located in the turn regions. This indicates that the database-derived distance constraints seem able to help precluding certain non-native conformations in these regions. Interestingly, most of the regions with differential RMSD values among the three structures are around one or more glycine residues (located at residues 10, 12, 13, 15, 48, 60, 75, 77, 115, 138, and 151), illustrated by vertical bars in cyan (Figure 2). Because glycine has no side chain, its conformation is extremely flexible and can be easily affected by the additional torsion angle or database-derived distance constraints. This flexibility of glycine conformation may in turn influence the conformations of the neighbouring residues, together causing the differences in the RMSD values.

Discussions

In this paper, we have assessed quantitatively the contribution of three different types of NMR constraints, NOE, torsion angle, and knowledge-based distance constraints, to the precision and accuracy of NMR structure ensembles. Based on the fact that the number of NOE constraints per residue contributes the most to the precision and accuracy, we primarily focused on how much improvement on the structures can be made, in the presence of NOE constraints, by other types of constraints, in our case, the torsion angle constraints and knowledge-based distance constraints. To investigate the importance of the number of available NOE constraints, we carefully selected a set of structures with high (~15) and low (~7) numbers of NOE constraints per residue and refined the structures separately by using NOE constraints only, NOE and torsion angle constraints, or NOE and knowledge-based distance constraints. We concluded, based on our test cases, that (i) NOE constraints alone can account for a major improvement on the precision and accuracy of the resulting ensembles; (ii) the ensemble structures can be improved at the local level after introducing torsional angle or knowledge-based distance constraints, in terms of stereo-chemistry, even though the overall improvement for accuracy is less pronounced; (iii) the number of NOE per residue affects the extent of the improvement: the increase in precision and accuracy of NMR structures with low numbers of NOE constraints per residue appears to be more siginificant compared to the ones with high numbers of NOE constraints per residue; (iv) the contribution of the knowledge-based distance constraints to the ensemble precision and accuracy is (at least) comparable to that of the torsion angle constraints.

Note that this paper is written for a different purpose with a different set of test results from our previously published two papers on related topics: Cui *et al.*, 2005 and Wu *et al.*, 2007. Cui *et al.* 2005 was on combining NOE and torsion angle constraints with database-derived distance bounds TOGETHER for NMR structure refinement. Wu *et al.* 2007 was on combining NOE and torsion angle constraints with database-derived distance potentials for NMR structure refinement. The database-derived distance bounds and potentials are different and result in different computational approaches to structure refinement, although both are based on database distributions of the inter-atomic distances. Therefore, the two papers are very different. The test cases are also very different. The current paper is more related to Cui *et al.*, 2005. However, the current paper is to evaluate different types of conformational constraints as they are applied to NMR structure refinement. The new results contained in this paper are the comparisons on the qualities of the structures refined by using NOE constraints only, by using NOE plus torsion angle constraints, and by using NOE plus database-derived distance constraints. These tests on different combinations of the three major conformational constraints have never been conducted, evaluated, and reported in either Cui *et al.* 2005 or Wu *et al.* 2007.

Assessment on the precision and accuracy of NMR structures is one of the most active areas of research in the bio-molecular NMR field, in part, due to the fact that there are no widely-accepted criteria (Snyder *et al.*, 2005). The 'current' approach to estimating the precision of an ensemble of structures is to measure the variability of the structures across the ensemble. The precision is usually expressed as the coordinate RMSD of all the backbone atoms from the 'mean' structure. Recently, internally well-defined "core atom set(s)" instead of all the backbone atoms were proposed as a measure of the precision of NMR structures (Snyder and Montelione, 2005). An advantage of using the "core atom set(s)" is that it helps to make a more meaningful alignment between the NMR structures and their reference structures – only taking into account those well-defined regions. Thus, RMSD will not be underestimated due to the regions with noisy coordinates and therefore reflects the precision of the well-defined regions of the structures. On the other hand, many other measures are developed to assess the quality of NMR structures, such as MOLPROBITY (Word *et al.*, 2000; Word *et al.*, 1999), PROCHECK (Laskowski *et al.*, 1996) and WHAT-CHECK (Hooft *et al.*, 1996). While calculating RMSD between NMR-derived structures and their reference X-ray structures is a straightforward measure of accuracy, several 'goodness-of-fit' measures such as RFAC (Gronwald *et al.*, 2000) and RPF (Huang *et al.*, 2005) are also considered as useful tools for the accuracy assessment.

A long held viewpoint on the precision and accuracy of NMR structures is that the number of NOE constraints per residue constitutes the most important determinant (Clore *et al.*, 1993), although the completeness of NOE constraints per residue has recently been considered to be more informative (Doreleijers *et al.* 1999). The importance of NOE as well as other types of experimental constraints, however, has never quantitatively assessed in the context of 'real' NMR data, for which the number of NOE constraints per residue varies substantially from about 6 to 19 (see Introduction). By comparing with torsion angle constraints and knowledge-based distance constraints, we found that the use of the torsion angle and knowledge-based distance constraints can also significantly increase the precision of the ensembles (by 4-23%), especially for the structures with low numbers of NOE constraints per residue. However, the accuracy *per se* appears not changed accordingly (by 3-8%). Our data indicated that an increase in precision does not necessarily correlate with an increase in accuracy, which is also consistent with some previous findings (Clore *et al.*, 1993).

In addition to NOE and torsion angle constraints, chemical shifts data and more importantly, residual dipolar coupling (RDC) data, are experimental constraints used to refine NMR structures (Clore *et al.*, 1998; Tjandra *et al.*, 1997). The RDC data are particularly useful since in a weak alignment media they can provide information about angles between the inter-nuclear axis, *e.g.,* the N-H bond of a residue, and the orientation of the applied magnetic fields. These long-range constraints provide additional useful structural information and can be used for NMR structure refinement. Such experimental data can be assessed in a similar fashion as described in this paper. On the other hand, several knowledge-based potential functions derived from dihedral angle and distance constraints from known protein structures have been used to refine several different sets of NMR and X-ray structures using different software and refinement protocols (Grishaev and Bax, 2004; Kuszewski *et al*., 1996; Wall *et al.*, 1999, Sun *et al*., 2007). A careful assessment on their contributions to the precision and accuracy of the refined structures using the data sets in the same refinement system is necessary as well for the identification of the types of constraints that are key to structural refinement in presence of low or high numbers of NOE constraints per residue.

Overall, our present study represents a small step towards the understanding, in a quantitative manner, the precision and accuracy of the NMR structures as influenced by the applications of different types of structural constraints and in particular, how NOE, torsion angle, and knowledge-based distance constraints contribute to the precision and accuracy of the refined protein structures. A comprehensive assessment on other types of experimental and theoretical constraints can be interesting as well and will be investigated in future.

Acknowledgements

The authors would like to thank the anonymous referees for their valuable remarks and suggestions on the work described in the paper. They would like to thank Ajith Gunaratne, Peter Vedell, and Di Wu for their helpful discussions and Department of Mathematics and Program on Bioinformatics and Computational Biology at Iowa State University for their research support. This work has also been partially supported by the NIH/NIGMS grant R01GM081680.

References

Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H. and Westbrook, J. (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*, **7 Suppl**, 957-959.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, **54**, 905-921.

Clore, G.M., Gronenborn, A.M. and Tjandra, N. (1998) Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J Magn Reson*, **131**, 159-162.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J Mol Biol*, **231**, 82-102.

Creighton, T.E. (1993) *Proteins : structures and molecular properties*. W.H. Freeman, New York.

Cui, F., Jernigan, R. and Wu, Z. (2005) Refinement of NMR-determined protein structures with database derived distance constraints. *J Bioinform Comput Biol*, **3**, 1315-1329.

Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L. and Ulrich, E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J Biomol NMR*, **26**, 139-146.

Doreleijers, J.F., Raves, M.L., Rullmann, T. and Kaptein, R. (1999) Completeness of NOEs in protein structure: a statistical analysis of NMR. *J Biomol NMR*, **14**, 123-132.

Doreleijers, J.F., Rullmann, J.A. and Kaptein, R. (1998) Quality assessment of NMR structures: a statistical survey. *J Mol Biol*, **281**, 149-164.

Grishaev, A. and Bax, A. (2004) 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.

Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. and Clore, G.M. (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science*, **253**, 657-661.

Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K.P. and Kalbitzer, H.R. (2000) RFAC, a program for automated NMR R-factor estimation. *J Biomol NMR*, **17**, 137-151.

Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.

Huang, Y.J., Powers, R. and Montelione, G.T. (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *J Am Chem Soc*, **127**, 1665-1674.

Kohda, D. and Inagaki, F. (1992) Three-dimensional nuclear magnetic resonance structures of mouse epidermal growth factor in acidic and physiological pH solutions. *Biochemistry*, **31**, 11928-11939.

Kraulis, P.J., Domaille, P.J., Campbell-Burk, S.L., Van Aken, T. and Laue, E.D. (1994) Solution structure and dynamics of ras p21.GDP determined by heteronuclear three- and four-dimensional NMR spectroscopy. *Biochemistry*, **33**, 3515-3531.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci*, **5**, 1067-1080.

Laskowski, R.A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. . (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283-291.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**, 477-486.

Lian, L.Y., Derrick, J.P., Sutcliffe, M.J., Yang, J.C. and Roberts, G.C. (1992) Determination of the solution structures of domains II and III of protein G from Streptococcus by 1H nuclear magnetic resonance. *J Mol Biol*, **228**, 1219-1234.

Liu, Y., Zhao, D., Altman, R. and Jardetzky, O. (1992) A systematic comparison of three structure determination methods from NMR data: dependence upon quality and quantity of data. *J Biomol NMR*, **2**, 373-388.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) Stereochemical quality of protein structure coordinates. *Proteins*, **12**, 345-364.

Schwalbe, H., Grimshaw, S.B., Spencer, A., Buck, M., Boyd, J., Dobson, C.M., Redfield, C. and Smith, L.J. (2001) A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Sci*, **10**, 677-688.

Snyder, D.A., Bhattacharya, A., Huang, Y.J. and Montelione, G.T. (2005) Assessing precision and accuracy of protein structures derived from NMR data. *Proteins*, **59**, 655-661.

Snyder, D.A. and Montelione, G.T. (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins*, **59**, 673-686.

Spoerner, M., Herrmann, C., Vetter, I.R., Kalbitzer, H.R. and Wittinghofer, A. (2001) Dynamic properties of the Ras switch I region and its importance for binding to effectors. *Proc Natl Acad Sci U S A*, **98**, 4944-4949.

Spronk, C.A., Nabuurs, S.B., Bonvin, A.M., Krieger, E., Vuister, G.W. and Vriend, G. (2003) The precision of NMR structure ensembles revisited. *J Biomol NMR*, **25**, 225-234.

Sun, X., Wu, D., Jernigan, R., and Wu, Z. (2007) PRTAD: A protein residue torsion angle distribution database, to appear in IEEE Transaction on Data Mining and Bioinformatics.

Tjandra, N., Omichinski, J.G., Gronenborn, A.M., Clore, G.M. and Bax, A. (1997) Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Biol*, **4**, 732-738.

Wall, M.E., Subramaniam, S. and Phillips, G.N., Jr. (1999) Protein structure determination using a database of interatomic distance probabilities. *Protein Sci*, **8**, 2720-2727.

Word, J.M., Bateman, R.C., Jr., Presley, B.K., Lovell, S.C. and Richardson, D.C. (2000) Exploring steric constraints on protein mutations using MAGE/PROBE. *Protein Sci*, **9**, 2251-2259.

Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S. and Richardson, D.C. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol*, **285**, 1711-1733.

Wu, D., Jernigan, R., Wu, Z. (2007) Refinement of NMR-determined protein structures with database derived mean-force potentials. *Proteins: Structure, Function, and Bioinformatics*, **in press**.

Wüthrich, K. (1986) *NMR of proteins and nucleic acids*. Wiley, New York.

***Table 1*. Data Sets**

Protein ID	Residues	NOE[a]	DA[b]
1CEY	128	7.2	112
1CRP	166	19.7	69
1E8L	129	13.1	110
1EPH	53	6.7	24
1GB1	56	16.5	93
1PFL	139	12.9	200
2IGG	64	7.4	39

NOE[a] – NOE distance constraints per residue.
DA[b] – dihedral angle constraints.

***Table 2*. RMSD of the Structural Ensembles Refined with Experimental Constraints**

PDB	HN and HA Atoms (Å)	Backbone Atoms (Å)	Non-hydrogen Atoms (Å)	Original Published (Å)
1E8L	1.48 ± 0.40	2.30 ± 0.50	2.97 ± 0.60	0.50+0.13[a]
1EPH	1.76 ± 0.38	2.04 ± 0.61	2.94 ± 0.70	0.87+0.29[b]
1GB1	0.33 ± 0.05	0.45 ± 0.12	1.04 ± 0.18	0.27+0.03[c]
2IGG	2.35 ± 0.97	2.62 ± 0.85	3.29 ± 0.83	0.90+0.20[d]

[a] (Schwalbe *et al.*, 2001)
[b] (Kohda and Inagaki, 1992)
[c] (Gronenborn *et al.*, 1991)
[d] (Lian *et al.*, 1992)

***Table 3.* Largely Deviated Inter-Atomic Distances and Related Residue Pairs in the Averaged and Energy-minimized Structures of Ensembles[†] (Distance/Residue Pair)**

Protein ID	NOE	NOE+TOR	NOE+DB_DIST
1CEY	265/76	126/38	113/36
1CRP	156/54	110/45	108/46
1E8L	134/42	93/34	86/29
1EPH	244/48	58/25	55/26
1GB1	N/A	28/15	16/12
1PFL	140/48	105/41	99/38
2IGG	78/31	75/31	27/18

[†] Each ensemble contains 50 structures accepted by the default criteria of CNS.

***Table 4.* RMSD of NMR Structure Ensembles[†] (Mean ± Standard Deviation)**

Protein ID	Data	Backbone Atoms (Å)	Non-H Atoms (Å)
1CEY	NOE	2.61 ± 0.71	3.54 ± 0.80
	NOE+TOR	2.45 ± 0.60	3.20 ± 0.69
	NOE+DB_DIST	2.41 ± 0.55	3.03 ± 0.73
1CRP	NOE	2.67 ± 0.64	3.23 ± 0.76
	NOE+TOR	2.51 ± 0.58	3.02 ± 0.67
	NOE+DB_DIST	2.54 ± 0.54	2.92 ± 0.61
1E8L	NOE	2.39 ± 0.49	3.31 ± 0.60
	NOE+TOR	2.30 ± 0.50	2.97 ± 0.60
	NOE+DB_DIST	2.22 ± 0.52	2.67 ± 0.57
1EPH	NOE	2.65 ± 0.64	3.42 ± 0.70
	NOE+TOR	2.04 ± 0.61	2.94 ± 0.70
	NOE+DB_DIST	2.08 ± 0.51	2.85 ± 0.61
1GB1	NOE	N/A	N/A
	NOE+TOR	0.45 ± 0.12	1.04 ± 0.18
	NOE+DB_DIST	0.54 ± 0.19	1.07 ± 0.24
1PFL	NOE	2.44 ± 0.51	3.56 ± 0.69
	NOE+TOR	2.40 ± 0.48	3.15 ± 0.58
	NOE+DB_DIST	2.35 ± 0.50	2.98 ± 0.54
2IGG	NOE	2.54 ± 0.77	3.25 ± 0.84
	NOE+TOR	2.62 ± 0.85	3.29 ± 0.83
	NOE+DB_DIST	2.32 ± 0.77	3.06 ± 0.78

[†] Each ensemble contains 50 structures accepted by the default criteria of CNS.

***Table 5.* RMSD of NMR Structures with Their Corresponding X-ray Structures[†] (Mean ± Standard Deviation)**

NMR ID	X-Ray ID	NOE (Å)	NOE+TOR (Å)	NOE+DB_DIST (Å)
1CEY	3CHY	1.99 ± 0.29	1.85 ± 0.19	1.84 ± 0.14
1CRP	1IAQ_A	1.82 ± 0.40	1.77 ± 0.29	1.72 ± 0.27
1PFL	1FIK	1.67 ± 0.08	1.66 ± 0.07	1.64 ± 0.09
2IGG	1FCC_C	1.93 ± 0.67	1.97 ± 0.79	1.83 ± 0.51

[†] Each ensemble contains 50 structures accepted by the default criteria of CNS.

***Table 6.* Statistics of the Ramachandran Plots of 1CRP[†]**

	NOE	NOE+TOR	NOE+DB_DIST
Most Favored Regions	121 (80.7%)	111 (74.0%)	118 (78.7%)
Additionally Allowed Regions	23 (15.3%)	36 (24.0%)	29 (19.3%)
Generously Allowed Regions	2 (1.3%)	1 (0.7%)	3 (2.0%)
Disallowed Regions	4 (2.7%)	2 (1.3%)	0 (0.0%)
Total	150 (100%)	150 (100%)	150 (100%)

[†] The Ramachandran plots are generated from the averaged and energy-minimized structures of the 1CRP structure ensembles. Each ensemble contains 50 accepted structures.

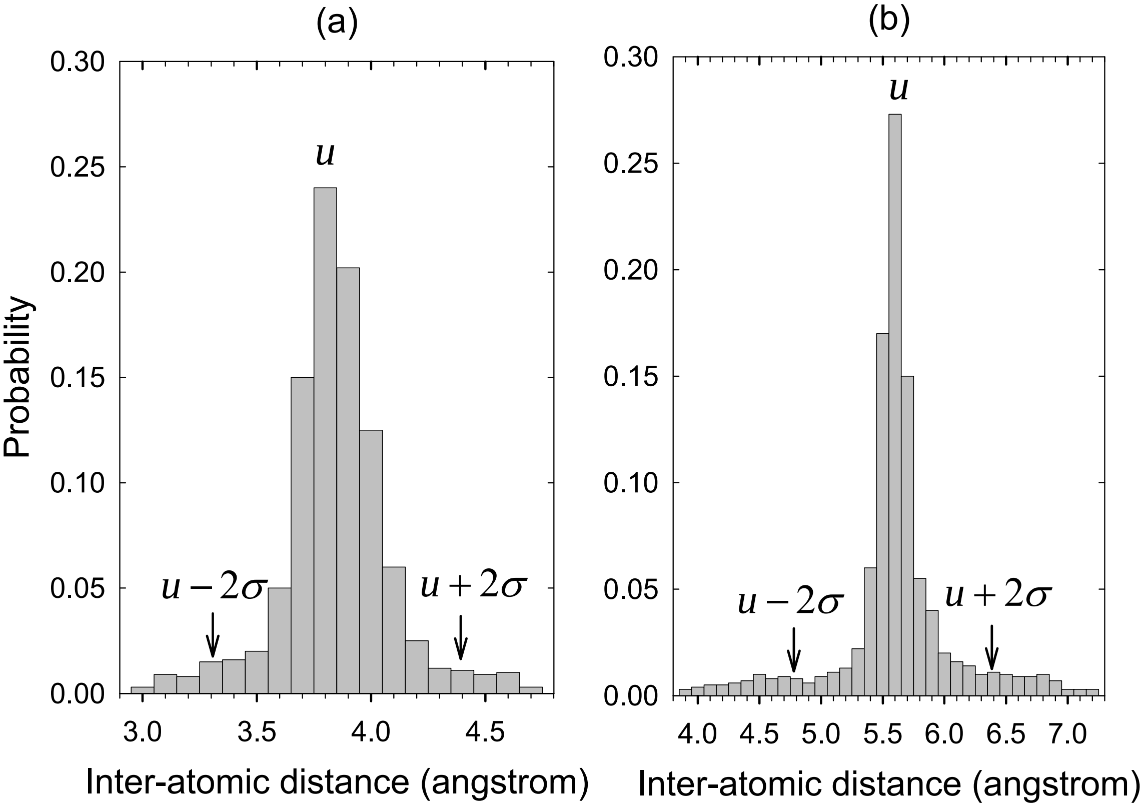

Figure. 1. Distributions of inter-atomic distances. (a) Distribution of the distances between atom C in residue ARG and atom O in residue ILE separated by zero residue (b) Distribution of the distances between atom C_β in ALA and atom N in LEU separated by one residue. The ranges of mean ± two standard deviations (μ ± 2σ) of the distributions are marked by arrows.

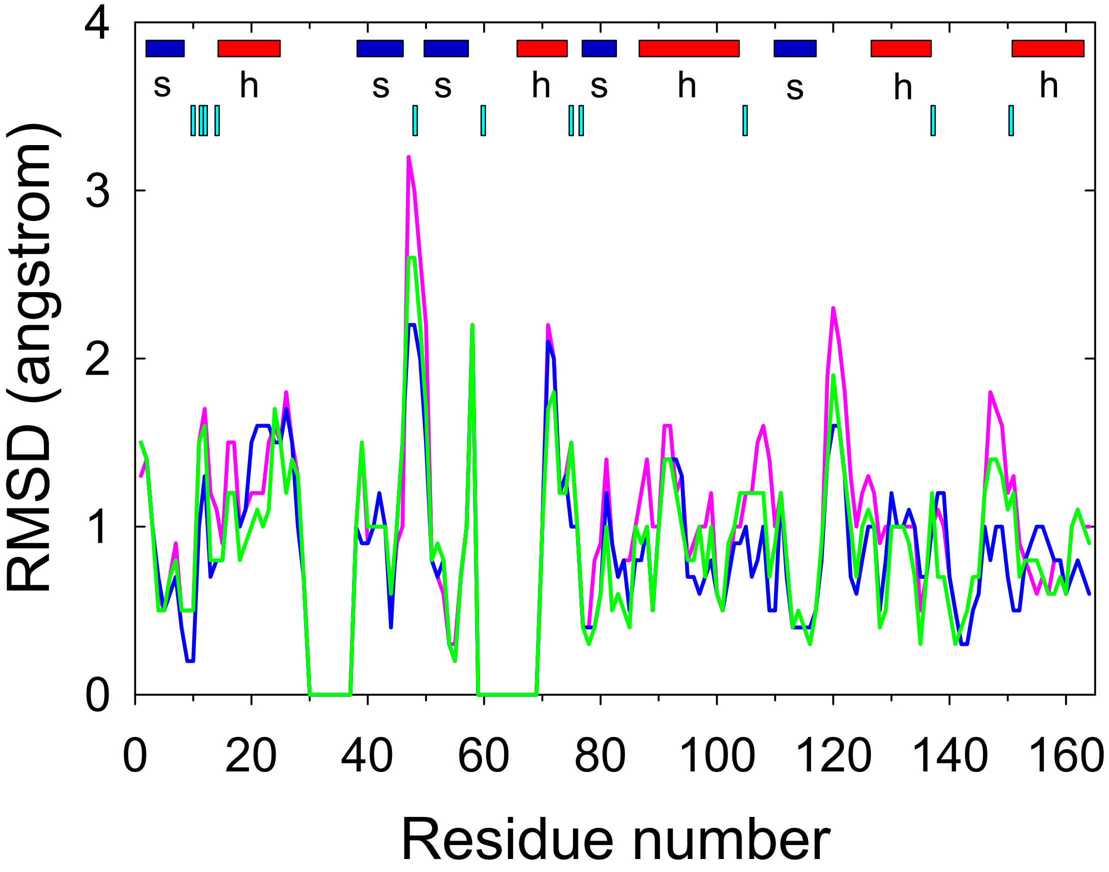

Figure. 2. Residue-residue RMSD between 1CRP NMR structures and the X-ray structure 1IAD chain A. The NMR structures of 1CRP are refined separately by NOE constraints only (NOE, magenta), NOE and torsion angle constraints (NOE+TOR, green), and NOE and database-derived distance constraints (NOE+DB_DIST, blue). The resulting ensembles of 50 structures are averaged and minimized. The backbone atoms of the averaged and energy-minimized structures for the ensembles are used for the calculation of RMSD against the X-ray structure. The secondary structures of the protein are marked on the top of the profiles with 'h' standing for helices, 's' for β-sheets, and 't' for turns. Glycine residues are marked by vertical bars in cyan.