RECOORD: A Recalculated Coordinate Database of 500+ Proteins from the PDB Using Restraints from the BioMagResBank

Aart J. Nederveen,[1] Jurgen F. Doreleijers,[2] Wim Vranken,[3] Zachary Miller,[4] Chris A.E.M. Spronk,[5] Sander B. Nabuurs,[5] Peter Güntert,[6] Miron Livny,[4] John L. Markley,[2] Michael Nilges,[7] Eldon L. Ulrich,[2] Robert Kaptein,[1] and Alexandre M.J.J. Bonvin[1]*

[1]*Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
[2]*Center for Eukaryotic Structural Genomics, University of Wisconsin-Madison, Madison, Wisconsin, USA*
[3]*Macromolecular Structure Database, European Bioinformatics Institute, Wellcome Trust, Genome Campus, Hinxton, Cambridge, United Kingdom*
[4]*Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA*
[5]*Center for Molecular and Biomolecular Informatics, Radboud University, Nijmegen, The Netherlands*
[6]*Tatsuo Miyazawa Memorial Program, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Japan*
[7]*Unité de Bioinformatique Structurale, Institut Pasteur, Paris, France*

ABSTRACT **State-of-the-art methods based on CNS and CYANA were used to recalculate the nuclear magnetic resonance (NMR) solution structures of 500+ proteins for which coordinates and NMR restraints are available from the Protein Data Bank. Curated restraints were obtained from the BioMagResBank FRED database. Although the original NMR structures were determined by various methods, they all were recalculated by CNS and CYANA and refined subsequently by restrained molecular dynamics (CNS) in a hydrated environment. We present an extensive analysis of the results, in terms of various quality indicators generated by PROCHECK and WHAT_CHECK. On average, the quality indicators for packing and Ramachandran appearance moved one standard deviation closer to the mean of the reference database. The structural quality of the recalculated structures is discussed in relation to various parameters, including number of restraints per residue, NOE completeness and positional root mean square deviation (RMSD). Correlations between pairs of these quality indicators were generally low; for example, there is a weak correlation between the number of restraints per residue and the Ramachandran appearance according to WHAT_CHECK ($r = 0.31$). The set of recalculated coordinates constitutes a unified database of protein structures in which potential user- and software-dependent biases have been kept as small as possible. The database can be used by the structural biology community for further development of calculation protocols, validation tools, structure-based statistical approaches and modeling. The RECOORD database of recalculated structures is publicly available from http://www.ebi.ac.uk/msd/recoord. Proteins 2005;59:662–672.** © 2005 Wiley-Liss, Inc.

Key words: NMR; structure calculation; structure refinement; experimental restraints; structure validation; water-refinement

INTRODUCTION

Over 3800 macromolecular structures determined with nuclear magnetic resonance (NMR) spectroscopy are available in the Protein Databank (PDB).[1] Although the NMR structures account for only a small subset of the available structures, they represent 25% of the structural folds in the PDB.[2] The cornerstone of NMR structure determination consists mainly of three classes of structural restraints that are based on experimental data: distance restraints, dihedral angle restraints and orientational restraints.[3,4] These restraints have been combined with diverse software, protocols and force fields since the early days of NMR structure determination.[4] Initially, structures were obtained using metric matrix distance geometry calculations.[5] A method based on a variable target function in torsion angle space was introduced in 1985.[6] Restrained molecular dynamics (MD), an approach that has become very powerful in NMR structure determination, was introduced in the same year by Kaptein et al.[7] Later these methods were expanded using several other concepts, such as simulated annealing and the simplification of force fields using an approximation of the nonbonded interactions by a repulsive potential. The NMR community has used two groups of programs most extensively for structure determinations: XPLOR/CNS[8,9] and DYANA/CYANA.[10] Both are based on simulated annealing driven by MD calculations, as performed in torsion angle space within CYANA and both torsion angle and Cartesian space within XPLOR/CNS.

A.J. Nederveen and J.F. Doreleijers contributed equally to this work.

*Correspondence to: Alexandre M.J.J. Bonvin, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@chem.uu.nl.

Received 6 September 2004; 21 October 2004; Accepted 1 November 2004

Published online 8 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20408

The quality of NMR structural models is typically assessed by various quality indicators; these indicators can be related to the NMR data themselves (restraint violations, NOE completeness,[11] information content[12]), to the precision of the ensembles or to several structural features of the coordinate data, often expressed as normality scores relative to an X-ray database.[13,14] Several studies have pointed out that the quality of NMR structures is very heterogeneous.[15,16] Consequently, NMR models are often not the first choice for use in database approaches and drug design.[2] However, the protocols and force fields that have been used for calculating the NMR structures have improved greatly during the last decade. Nowadays, after the simulated annealing stage, a refinement in explicit water is commonly performed.[17] In this stage a full nonbonded potential is used together with explicit solvent molecules, thus mimicking the natural environment of a protein in solution. Recently, Nabuurs et al.[18] launched the DRESS database with 100 re-refined NMR structures and proved the benefit of refinement in explicit solvent for those entries.

So far no attempt has been presented to standardize NMR entries with respect to the calculation protocols used. The improvement in protocols and force fields, however, gives a rationale for recalculating structures from PDB entries for which NMR restraints are available. Up to now, a major hindrance to such an attempt has been the inconsistency between coordinate and restraint data and the absence of reliable conversion software between different data formats. This problem was recently solved for 545 proteins, for which curated restraints are now available from the BioMagResBank (BMRB).[19] This database is described in detail in another article.[20] As described here, we used state-of-the-art protocols in CNS and CYANA to recalculate the structures of these 545 proteins. This effort provides the structural biology community with a unified database of NMR protein structures in which potential user- and software-dependent biases are kept as small as possible. Average values and standard deviations (SDs) of bond lengths and bond angles for heavy atoms in the RECOORD database are consistent with the reference values derived from model structures in the Cambridge Structural Database (CSD).[21]

This article assesses the outcome of the recalculation effort. We also address the relationships in the recalculated coordinate sets among NMR data density, NOE completeness, restraint violations, protein type, normality relative to a X-ray database and precision of the NMR ensembles.

MATERIALS AND METHODS

NMR Data

Most of the restraints now available from the PDB cannot be loaded directly into structure calculation programs. We used restraints in NMR–STAR 3 format from the BioMagResBank[19] Filtered Restraints Database (FRED), corresponding to a set of 545 entries with distance restraints originally deposited in XPLOR/CNS, Discover[22] or DYANA/CYANA format. The construction of the FRED database is described in detail elsewhere.[20] However, because we analyzed restraint violations in the deposited structures and compared them to the recalculated sets, we will reiterate how the restraints in the FRED database were made consistent with the coordinates. Several categories, including residue numbering and IUPAC atom nomenclature,[23] were corrected. Stereospecific assignments present in the original data were conserved. Redundant restraints and other surplus were removed from the restraint list using a new module in the program Wattos.[20] The NMR–STAR files were converted to CNS and CYANA formats using the FormatConverter software that was developed in the framework of the Collaborative Computing Project for the NMR Community (CCPN).[24]

To assess the correctness of the restraint conversion process and to curate the converted restraints, a violation analysis of the deposited entries was performed during the construction of FRED. Since it was difficult to determine the kind of averaging and/or pseudo-atom usage that was originally used for each entry, sum averaging was used for all entries. By using this method as a conservative approach, no calculated distances between atom pairs larger than those originally calculated by the authors were expected. Sum-averaging defines the effective distance between two atom sets as[25]

$$d_{eff} = \left(\sum_{ij} d_{ij}^{-6} \right)^{-1/6}$$

where d_{ij} is the distance between atom i and atom j and the sum runs over all pairwise distances between the two selected sets of atoms. Sum averaging ensures that the effective distance d_{eff} is always shorter than the minimum distance d_{ij} entering the sum.

All hydrogen positions were regenerated in CNS before the violation analysis was carried out to ensure consistency with the CNS topology. Diastereotopic methyl groups and methylene protons were swapped in the restraint list if the NOE energy was lower after swapping in more than 75% of the deposited models. Consecutively, restraints involving stereospecific assignments to prochiral groups were deassigned and made ambiguous if they gave rise to violations larger than 2 Å in one of the models or violations larger than 1 Å in more than 50% of the models.

Figure 1 presents histograms for the number of residues per entry, the NMR data density (number of restraints per residue, both dihedral and distance restraints) together with the secondary structure content and the year of deposition of all entries in the database.

Recalculation Setup

Structure recalculations were carried out in both CNS version 1.1[9] and CYANA version 1.0.6.[10,26] CNS topologies were generated with a modified version of the CNS script 'generate_easy.inp.' Disulfide bridges were automatically detected from the coordinates of the deposited models and incorporated into the topology. Patches for *cis* peptide bonds were added to the topology if the absolute value of the omega angle in the first model of each entry was

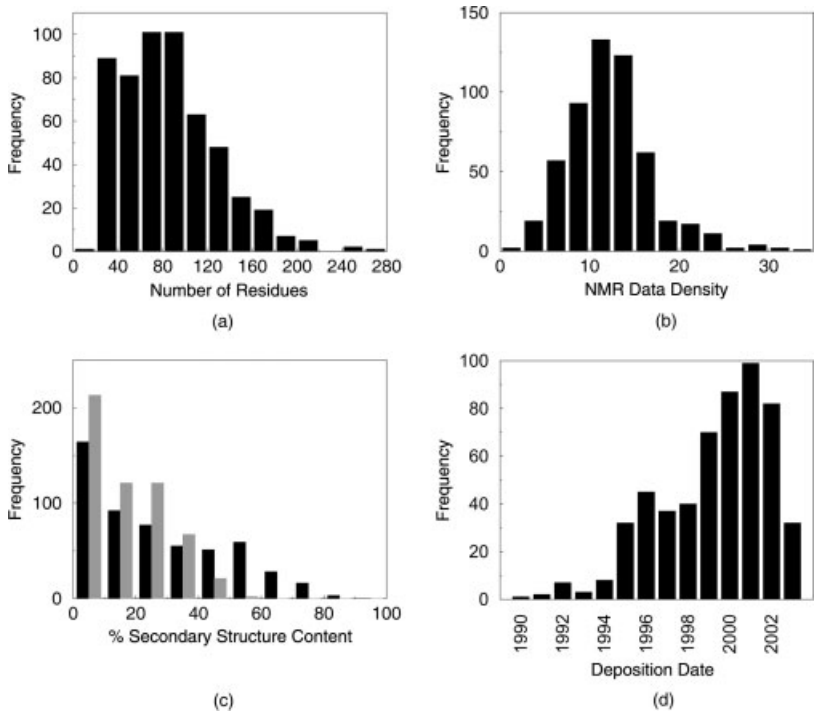

Fig. 1. Histograms for (a) the number of residues per entry (entries 1EZP and 1EZO, each with 370 residues, are not shown); (b) NMR data density (number of distance and dihedral restraints per residue, redundant restraints not counted); (c) percentage secondary structure content: alpha helical content (black bars) and sheet content (grey bars) and (d) year of deposition.

smaller than 25°. The protonation states of the histidines in the sequence were conserved from the original PDB coordinates. The CNS topologies were automatically converted to CYANA sequence files containing information for disulfide bridges, *cis* peptides and histidine protonation states. Both in CNS and CYANA, sum averaging was used for implementing distance restraints. We used both distance and dihedral angle restraints, when available. Residual dipolar coupling (RDC) data, present for 24 entries, were not used in this study.

For CNS, we implemented a stand-alone version of the ARIA 1.2[27] protocols consisting of four MD simulated annealing stages, both in torsion angle and Cartesian space: (i) a high-temperature torsion angle dynamics phase at 10,000K (2,000 steps of 24 fs integration time steps); (ii) torsion angle dynamics cooling phase from 10,000K to 50K (2,000 steps of 24 fs); (iii) Cartesian dynamics cooling phase from 2,000K to 1,000K (8,000 steps of 3 fs); (iv) Cartesian dynamics cooling phase from 1,000K to 50K (8,000 steps of 3 fs). We calculated 200 structures for each entry and sorted the structures with respect to the total energy. Before doing this, the protocol was fine-tuned automatically by trying to calculate one model without distance restraint violations larger than 0.5 Å within 10 trials. If this could not be achieved, the subsequent calculation of the 200 models was performed with the total number of time steps doubled, which was the case for 133 entries. For CYANA, the standard simulated annealing protocol was used.[10] This protocol comprises a high-temperature phase at 9,600K of 2,000 steps and a cooling phase to 0K in 8,000 steps. In CYANA, the time step has no predefined value but depends on the energy change per time step. Again, 200 conformers per entry were calculated and sorted by the value of their target function. For entries with more than 120 residues, the total number of time steps was increased to 15,000 in order to allow for the larger sequences to be folded by the CYANA algorithm.

The two sets of 50 lowest-energy structures obtained with both programs were finally water-refined using a scheme similar to that used in ARIA,[17] since no water refinement is available with CYANA. The explicit solvent refinement consisted of the following steps: (i) immersion in a 7.0 Å shell of water molecules and energy minimization; (ii) slow heating from 100 to 500K in 100K temperature steps with 200 MD steps per temperature step (time step 3 fs), with harmonic position restraints on the protein heavy atoms that were slowly phased out during the heating stage; (iii) refinement at 500K with 2,000 MD steps (time step 4 fs); (iv) slow cooling from 500K to 25K in 25K temperature steps with 200 MD steps per temperature step (time step 4 fs); (v) final energy minimization (200 steps). The scaling of the force constants for bonds, angles, impropers and omega angles during the cooling stage was modified slightly with respect to the original protocol[17] to allow for the naturally occurring variation in these parameters as described by Engh and Huber.[21] The non-bonded interactions were calculated using a 8.5 Å cutoff with full Van der Waals (Lennard–Jones) and electrostatic potentials that incorporated the OPLS non-bonded parameters implemented in the PARALLHDG 5.3

force field.[17] Parameters were added to the OPLS nonbonded parameters to avoid proton–proton overlap.

The calculations were carried out in April 2004 on the Condor computer cluster[28,29] of the Department of Computer Sciences at the University of Madison-Wisconsin. The calculations (and test runs) were distributed over more than 800 processors. The cluster was used through a scheduler that allowed jobs to be started, check-pointed and restarted after interruption according to a user's base priority and a correction for used resources. This significantly decreased the maintenance of large batches of jobs such as this recalculation effort.

Quality Assessment

Violations of distance and dihedral restraints for all models were calculated with CNS.[9] NOE completeness[11] of the datasets was calculated with a new module of the program Wattos[20] with a cutoff distance of 4 Å. The program DSSP[30] was used for assessing the secondary structure content in all sets. Three categories were studied: helical content (α-helices), sheet content (extended strands) and overall secondary structure content (also including 3–10 helices, π-helices, hydrogen-bonded turns and bends). The programs PROCHECK[13] and WHAT_CHECK[14] were used to calculate additional quality scores. Most of the checks carried out in WHAT_CHECK make use of reference values from an X-ray database. Statistical Z-scores are used to express how many SDs a certain property is away from the database average. For a detailed explanation of the different WHAT_CHECK quality indicators we referred to http://www.cmbi.ru.nl/gv/pdbreport/checkhelp/. Backbone circular variances[13] were calculated as a running average over a window of three residues. For the mean circular variance per ensemble, only well-defined residues with window-averaged circular variance <0.2 were used. The QUEEN method was used to calculate the experimental structural uncertainty, as derived from the restraints.[31]

We used a paired t-test to determine whether differences for the various quality indicators were statistically significant assuming normal distributions. Differences were considered to be significant if $p < 0.05$. Correlations between various indicators were analyzed with Pearson's correlation coefficient.

Positional RMSD values in the original set were calculated for 484 entries. For each of the remaining entries, only one model was originally deposited by the authors. RMSDs relative to the coordinate-averaged model were calculated for backbone atoms (N, C^{α}, C) of both the entire sequence and of the well-ordered residues after superposition onto the backbone of the well-ordered residues only. The well-ordered residues for each entry were automatically derived from the original ensemble with the script 'wellordered.inp' available within ARIA and described in detail previously.[32] For the recalculated sets, RMSDs were calculated both relative to the average coordinates of the recalculated set and to the average coordinates of the originally deposited set.

RESULTS AND DISCUSSION

For each protein in our database, 200 structures were recalculated in both CNS and CYANA and the 50 best (lowest total energy) models were subjected to a final refinement in explicit water (see Materials and Methods). All deposited models and the 25 best models before and after water-refinement were selected for analysis and labeled as follows:

ORG, originally deposited models as present in the PDB;
CNS, models recalculated in CNS;
CYA, models recalculated in CYANA;
CNW, models recalculated in CNS and water-refined in CNS;
CYW, models recalculated in CYANA and water-refined in CNS.

All coordinate data, along with the analysis results and the scripts used in this recalculation project are available at http://www.ebi.ac.uk/msd/recoord.

This massive structure calculation effort was achieved through the implementation of both CNS and CYANA within the Condor framework for grid computing. The integrated Condor environment allows large numbers of calculations to be launched simultaneously with ease and the progress of the calculations to be monitored over time. The scheme implemented here is available for future projects and can be extended or upgraded easily with new versions of CNS or CYANA as they become available. On the Condor cluster, the total central processing unit (CPU) time including test runs needed to calculate the different sets amounted to 31,169 h, equivalent to 3.5 years on a single workstation. As an illustration of the computational requirements, a 2.2 GHz P4 computer requires 340 s to calculate one model for entry 2EZM (101 amino acids and 15.5 restraints per residue) with CNS and 31 seconds with CYANA, whereas the subsequent water-refinement in CNS takes 670 s.

Agreement with Experimental Distance Restraints

The RMS value of the distance restraint violations was found to decrease significantly after recalculation and refinement (Table I), which is also evident from the histogram depicted in Figure 2(a). In the ORG set, 44 entries have RMS distance restraint violations larger than 0.2 Å (see inset). This number decreases to eight for the CNW and CYW sets, seven of which are also in the set of 44 entries, indicating that only one of these entries was originally (in the ORG set) below the 0.2 Å cutoff. The RMS violation of this entry (1EZP) increases from 0.02 (ORG) to 0.20 (CNW) Å. Entry 1EZP is a 370 amino acid protein that could not be properly refolded in the CNW set. For this entry, many RDC restraints were used in the original calculations that were not taken into account here. For 135 entries, the RMS distance restraint violations increase relative to the ORG set after recalculation and water-refinement (CYW and CNW). The entry with the largest increase is the aforementioned 1EZP that increases from 0.02 (ORG) to 0.2 (CNW) Å RMS violation.

We also investigated whether the number of distance restraint violations is biased by the programs that were

TABLE I. Quality Indicators of Database Before and After Recalculation and Refinement

	Original	Recalculation		Water-refinement	
	ORG	CNS	CYA	CNW	CYW
Violation analysis					
RMS distance restraint violations (Å)	0.08 ± 0.14[a]	0.04 ± 0.06	0.04 ± 0.05	0.04 ± 0.05	0.03 ± 0.04
# Consistent violations > 0.5 Å[b]	3.9 ± 13.7	0.3 ± 1.5	1.5 ± 5.2	0.1 ± 0.6	0.1 ± 0.6
RMS dih. restr. violations (degrees)	1.6 ± 4.6	0.5 ± 0.7	0.5 ± 0.7	0.5 ± 0.5	0.5 ± 0.6
# Bumps per 100 residues[c]	73 ± 63	11 ± 9	87 ± 37	10 ± 7	9 ± 7
WHAT_CHECK Z-scores					
2^{nd} Generation packing quality	−3.5 ± 1.9	−4.1 ± 1.9	−4.1 ± 1.8	−2.5 ± 2.0	−2.5 ± 2.0
Ramachandran plot appearance	−4.6 ± 1.6	−4.6 ± 1.2	−5.7 ± 1.1	−3.4 ± 1.4	−3.5 ± 1.4
χ_1/χ_2 Rotamer normality	−3.7 ± 1.5	−0.9 ± 1.3	−4.6 ± 0.4	−0.9 ± 1.0	−1.2 ± 1.0
Backbone conformation	−4.0 ± 3.0	−3.4 ± 2.6	−4.9 ± 2.9	−3.8 ± 2.7	−4.0 ± 2.7
DSSP secondary structure analysis					
Helical content	26.1 ± 22.4	22.3 ± 20.2	23.2 ± 20.8	25.6 ± 22.2	25.5 ± 22.2
Sheet content	15.8 ± 13.6	14.6 ± 13.1	13.3 ± 12.5	17.8 ± 15.0	18.1 ± 15.1
Secondary structure content[d]	71.3 ± 10.3	69.3 ± 10.9	68.0 ± 11.4	73.7 ± 9.0	73.4 ± 9.2
PROCHECK results					
Most favored regions	69.3 ± 14.2	69.0 ± 13.1	60.5 ± 14.1	76.1 ± 11.3	75.0 ± 11.5
Allowed regions	25.4 ± 10.6	26.0 ± 9.9	31.7 ± 9.9	19.6 ± 8.5	20.4 ± 8.5
Generously allowed regions	3.9 ± 3.8	3.7 ± 3.2	5.9 ± 4.6	2.5 ± 2.1	2.7 ± 2.2
Disallowed regions	1.3 ± 1.7	1.3 ± 1.4	1.9 ± 1.8	1.8 ± 1.8	1.9 ± 2.0
Precision NMR ensemble[e]					
Backbone RMSD (Å)	2.3 ± 2.8	2.9 ± 3.2	2.8 ± 3.7	2.9 ± 3.1	2.8 ± 3.6
Well-ordered RMSD (Å)	0.6 ± 1.0	1.0 ± 1.4	1.0 ± 1.4	1.1 ± 1.4	1.0 ± 1.4
Backbone RMSD ORG (Å)[f]	n/a	3.7 ± 3.9	4.0 ± 4.7	3.7 ± 3.8	3.9 ± 4.6
Well-ordered RMSD ORG (Å)	n/a	1.4 ± 1.7	1.5 ± 1.7	1.5 ± 1.6	1.5 ± 1.7
Circular variance	0.04 ± 0.03	0.05 ± 0.05	0.05 ± 0.03	0.05 ± 0.03	0.05 ± 0.03

[a]SDs in this table sometimes suggest the possibility of non-existing negative values, due to the presence of outliers in the distribution of specific indicators.
[b]A violation is considered to be consistent if it occurs in more than 50% of the models.
[c]Bumps are calculated for heavy atoms only; in WHAT_CHECK two atoms are said to 'bump' if they are closer than the sum of their Van der Waals radii minus 0.4 Å.
[d]Also including 3–10 helices, π helices, hydrogen bonded turns and bends.
[e]RMSDs are calculated for the backbone heavy atoms of both the entire sequence and of the well- ordered residues after superposition onto the well-ordered residues only.
[f]RMSD relative to average coordinates of original ensemble from ORG set.

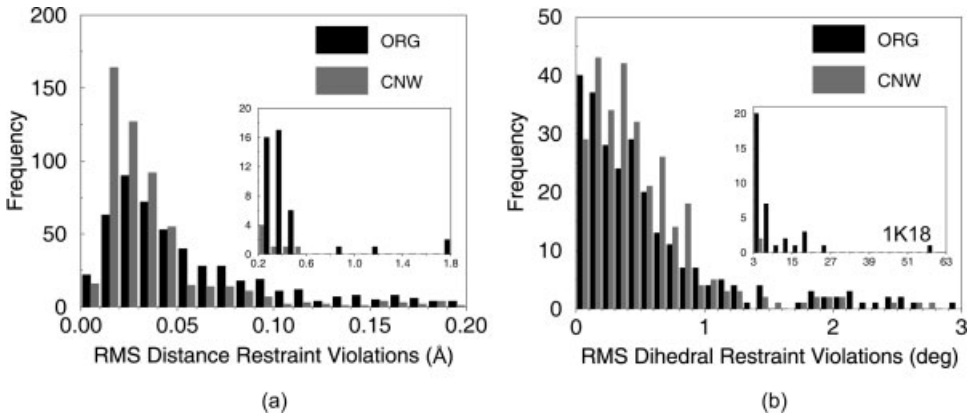

Fig. 2. Histograms for RMS values of (a) distance and (b) dihedral angle restraint violations for both the original entries (ORG) and the recalculated and refined entries (CNW). The bin sizes for the insets are 0.1 Å and 3°, respectively.

initially used by the authors for refining their structures. In Table II, we compare the RMS distance restraint violations before and after recalculation and water-refinement (sets ORG and CNW). The possible differences observed should be taken with circumspection, since the number of entries that were initially calculated with each program differs substantially (see Table II). Entries corresponding to different programs exhibit relatively large differences in distance restraint violations: for example, the Discover entries have a RMS distance restraint viola-

TABLE II. Quality Comparison of Entries Originally Calculated with CNS, CYANA or Discover for Sets ORG and CNW

	CNS (414)[a]	CYANA (97)	Discover (34)
RMS distance restraint violations (Å) (ORG)	0.06 ± 0.07	0.10 ± 0.15	0.27 ± 0.41
RMS distance restraint violations (Å) (CNW)	0.04 ± 0.03	0.03 ± 0.03	0.09 ± 0.13
2^{nd} Generation packing quality (ORG)	−3.6 ± 1.8	−3.5 ± 2.1	−2.7 ± 2.2
2^{nd} Generation packing quality (CNW)	−2.5 ± 1.9	−2.4 ± 2.1	−2.0 ± 2.4
Ramachandran plot appearance (ORG)	−4.6 ± 1.7	−4.7 ± 1.5	−4.3 ± 1.3
Ramachandran plot appearance (CNW)	−3.4 ± 1.4	−3.4 ± 1.4	−3.3 ± 1.8

[a]The numbers in parentheses indicate the number of entries with restraints in a specific format.
For some entries, the format of the deposited restraints might not match the final software used by the authors of the original structures.

tion of 0.27 Å, while the average value for the entire database is 0.08 Å (see Table I). Some of the violations in this set may be explained by the presence of pseudo-atom corrections that were not included in the restraint list as deposited at the PDB. In addition, some of the violations for the Discover entries may have been introduced during the regeneration of the proton coordinates in ideal geometry that was performed for the ORG set.

The average RMS value for the violations in the set of original models considered here (0.08 ± 0.14 Å) is larger than that reported previously for a smaller database of 97 NMR entries (0.06 ± 0.04 Å).[15] This increase of both the average and the SD mainly originates from the Discover entries. In Table II, the mean RMS distance restraint violation for the XPLOR/CNS entries (0.06 Å) is equal to that for the earlier database. Furthermore, the RMS value for the violations decreases significantly after recalculation, which indicates that many violated restraints could easily be satisfied.

Agreement with Experimental Dihedral Restraints

The RMS values for the dihedral angle restraint violations also decrease after recalculation. RMS values larger than 3° were found for 36 entries [see inset Fig. 2(b)] in the ORG set. The entry with the largest RMS violation (58.7°) is 1K18, with 24 ϕ/ψ restraints, which are almost all (20) severely violated. These large dihedral angle restraint violations disappeared in all entries after recalculation (in the CYW and CNW sets) except for three (1K18, 1IY6 and 1BCT). All 36 entries with large RMS values yielded smaller RMS values following recalculation. In spite of this, 129 of 289 entries exhibited increased RMS values in the CNW set for dihedral angle restraint violations after recalculation and refinement. The largest increase was for 1EZO, from 0.3 to 2.8°.

For the CYA set, we initially found a RMS value for the dihedral violations of 2.7 ± 8.5°. This increase is caused by the fact that in the BMRB FRED database dihedral restraints with definitions containing prochiral hydrogen atoms were not converted into the CYANA format.[20] This affected 92 of the recalculated entries. These restraints were thus not enforced during structure calculation in CYANA but were included in the violation analysis. If those entries are excluded from the analysis, a RMS value of 0.5 ± 0.7° is found. Since no significant differences were noticed for all other quality indicators, no entries were left out of the CYA set for analysis.

Quality of Recalculated Structures

Table I further analyzes several quality indicators for assessing the performance of the recalculations. The Z-score of the packing quality does not yet improve in the CNS and CYA sets. The Z-scores for Ramachandran plot appearance are similar for the CNS set and the ORG set, whereas the CYA set moves away by more than one SD from the ORG set. This difference between the CYANA and CNS sets disappears after water-refinement; after this step, both sets (CYW and CNW) are better than the original structures (ORG) by one SD, both for packing and Ramachandran. These improvements are very similar to those reported for the DRESS database.[18]

The improvement of the Ramachandran score is also visible in the PROCHECK results in Table I. Here again the CYA set does not improve in Ramachandran performance. Interestingly, while the percentages of residues in the most favored regions increase by approximately 8%, the number of residues in the disallowed regions for the refined sets (CNW and CYW) increases by small but significant amounts. This effect was also noticed in the refinement of IL-4 that was studied previously with almost the same refinement protocol.[17]

Figure 3(a) shows the correlation between the improvement of the Z-scores for packing quality and for Ramachandran plot appearance for the set CNW. The plot is divided into quadrants. In our database, 420 entries (77%) fall in the upper right quadrant, indicating that both the packing and the Ramachandran Z-score improved. There are 27 entries falling in the lower left quadrant, indicating that those entries decrease in quality for both scores. For the set CYW, this number is 28. The union between these two subsets from CNW and CYW is 32 and the intersection 23 entries. These numbers indicate that for about 6% of the entries in our database (32/545) no improvement could be reached after recalculation and refinement. In this set of 32, there were 15 entries that only deteriorated slightly (less than 1 SD for both Ramachandran and packing in both CYW and CNW). The remaining 17 entries were checked manually in order to discover the reason for their poor performance in our protocol. We concluded from the original papers that for 14 entries the authors used more

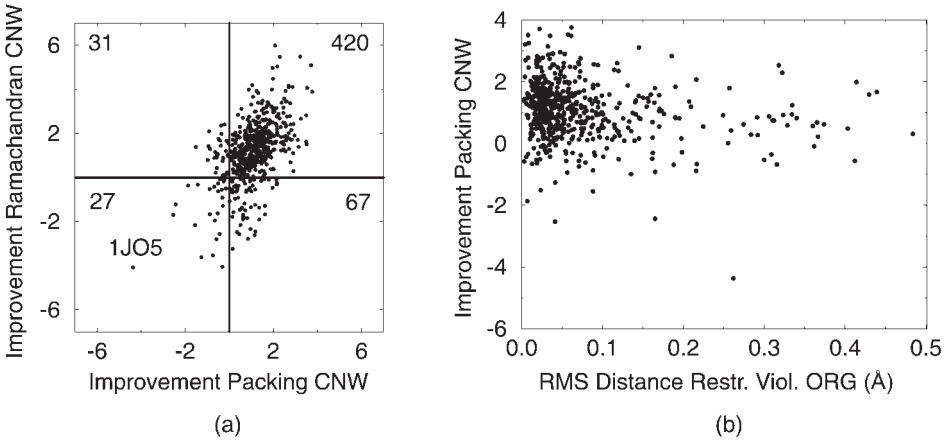

Fig. 3. (a) Improvement in Z-score ($Z_{CNW}-Z_{ORG}$) for the packing quality versus the Ramachandran plot appearance and (b) RMS distance restraint violations in the ORG set versus the packing quality Z-score improvement. RMS values above 0.5 Å are not shown in the plot (4 entries off scale).

restraints (either distance or dihedral restraints) in the original calculations than actually available from the PDB (entry codes: 1JO5, 1ALG, 1LS4, 1O8T, 1FYJ, 1SUT, 1K18, 1EZY, 1MM5, 1MM4, 1QK6, 1EQ3, 1EFE and 1P1T). For two entries (1KOY and 1GHH) dipolar couplings were used, and three entries (1O8T, 1MM4 and 1MM5) had a NMR data density of less than three restraints per residue. No straightforward reason could be found for the poor performance of entry 1BM4. Two entries in this set of 17 entries performed differently in CYANA and CNS (i.e. difference in improvement by more than 1 SD): 1LS4 (CYANA better) and 1O8T (CNS better).

Figure 3(b) shows the correlation between the RMS distance restraint violations in the ORG set and the improvement of the packing quality after recalculation and refinement (CNW). No clear correlation is present ($r = -0.16$). To find out if there were individual entries for which the number of violations in the ORG set was linked to a poor performance in the recalculation, we compared some of the subsets discussed above. The intersection of the subset of 44 entries having RMS distance restraint violations larger than 0.2 Å in the ORG set and the subset of 125 entries in the left quadrants and the lower right quadrants of Figure 3(a) contains 16 entries, indicating again that no general conclusions could be drawn from violations in the original set.

We also investigated whether the outcome of our recalculation effort was biased by the programs the initial authors used to determine their structures. Some of the important quality indicators both before and after recalculation and refinement (ORG and CNW sets) are compared in Table II. The Ramachandran plot appearance and the packing quality Z-scores do not differ notably for the different calculation programs. The Discover entries exhibited better packing before and after recalculation. These findings indicate that all categories benefit equally from the recalculation effort.

The improvement in rotamer normality is evident in Table I and is mainly an effect of the force field used in the recalculation for sets CNS, CYW and CNW (PARALLHDG 5.3). The force field used in CNS includes a dihedral angle potential describing rotameric states.[17] The number of bumps decreases dramatically in our database, from 73 heavy atom close contacts per 100 residues (ORG) to 10 (CNS, CYW and CNW). This closely approaches the number of bumps per 100 residues (7.3 ± 5.8) that was reported for the internal WHAT_CHECK X-ray database.[33] The Z-score for the backbone conformation improves only slightly for the CNS and CNW sets. The backbone normality score is a measure of how normal a given backbone conformation is compared to high-resolution X-ray structures and is based on the C^{α} positions of five sequential residues. Apparently, water-refinement is unable to improve this score, possibly due to the presence of unstructured loops. This difficulty in improving the backbone conformation was also observed in the DRESS database.[18]

The results for the secondary structure analysis in Table I again emphasize the effectiveness of water-refinement for increasing the secondary structure content. The helical content, however, decreased slightly (by 0.5%) after recalculation and water-refinement for the CNW set ($p = 0.04$) compared to the ORG set, whereas the sheet content increased significantly by 2.0%.

We identified 26 cases for which X-ray coordinates are available for the same (or highly homologous) monomeric proteins. Calculation of the positional backbone RMSDs for both the ORG and the CNW ensembles relative to the corresponding X-ray structures revealed a small but not yet significant improvement after recalculation (data not shown). This indicates that our structures are as accurate as the original ones. Together with the fact that for several quality indicators (Ramachandran, packing, rotamer normality) our RECOORD database is much closer to the reference database of X-ray structures, this makes us confident that the overall quality of the recalculated structures has improved.

Precision of Recalculated Ensembles

Table I presents RMSD values, which describe the precision of the recalculated conformer ensembles. Our

discussion is limited to the results for the CNW set as representative for the results after recalculation and refinement, since no important differences exist between this set and the CYW set. The RMSD values of the ensembles significantly increase after recalculation and refinement. The well-ordered RMSDs increase by 0.4 Å on average (i.e. from 0.6 to 1.0 Å). These results are in agreement with a previous paper from Spronk et al.[33] showing that the RMSD of many reported NMR ensembles can be increased without compromising the experimental restraints. The RMSDs for 44 entries increase by more than 1.0 Å and for 20 of those by more than 2.0 Å. For six entries of those 20 (1JO5, 1LS4, 1O8T, 1MM4, 1MM5 and 1EFE), the larger positional variability is explained by the lack of experimental restraints, as discussed in the previous subsection.

Similar observations can be made for the circular variances of the various sets. On average, this indicator slightly increases after recalculation, indicating that the variability of the backbone torsion angles is smaller in the ORG set.

Furthermore, the recalculated sets move slightly away from the ORG set, as shown by the RMSD values relative to the mean model of the ORG set (listed as 'well-ordered RMSD ORG' in Table I). The value of 1.4/1.5 Å for the well-ordered RMSD ORG is significantly higher than the 1.0/1.1 Å for the well-ordered RMSD relative to the mean coordinates of the recalculated sets. Here 102 entries are more than 1.0 Å away from the mean ORG model (as indicated by the difference [well-ordered RMSD ORG] − [well-ordered RMSD] $>$ 1 Å), and 28 entries are more than 2.0 Å away. Of those 28 entries, 27 also have a positional variability in the CNW set at least 1 Å larger (increase in well-ordered RMSD by more than 1 Å) than in the ORG set. The entry 1L7B is an interesting exception to this trend; it has well-ordered RMSDs of 0.4 Å and 0.9 Å for the ORG and CNW sets, respectively, and a well-ordered RMSD relative to the original mean coordinates of 3.1 Å. The RMS value for the distance restraint violations for this entry is 1.1 Å in the ORG set and 0.02 Å in the CNW set, suggesting a discrepancy between the restraint lists used in the two sets.

Comparison Between Programs

The programs CNS and CYANA differ in many respects, including the standard geometry of amino acid residues, the simplified force fields, the implementations of torsion angle dynamics and the calculation strategies. The different approaches manifest themselves in some parameters related to details of the structures, such as the number of bumps and the Z-scores for Ramachandran appearance, rotamer normality and backbone conformation reported by WHAT_CHECK (see Table I). However, the fact that the structures from both approaches become virtually indistinguishable after the unified refinement in explicit solvent shows that the small local differences in the structures originating from the two programs are not of a fundamental but rather of a technical nature. For example, the larger number of short non-bonded contacts in the CYA set (listed as bumps in Table I) is due to the use of slightly smaller repulsive core radii in CYANA compared to CNS.

The major difference between the two calculation programs used manifests itself in the computation time needed for the calculation: the generation of a model in CYANA is roughly 10 times faster than in CNS. The higher computational speed of CYANA is due in part to an implementation of torsion angle dynamics that requires only one force evaluation per integration step[10] instead of four in the corresponding CNS algorithm[34] and in part to the use of a pure torsion angle dynamics protocol, whereas with CNS a hybrid torsion angle/Cartesian space approach is used.

Relation Between NMR Data Quality and Structural Quality

The present recalculation effort provides a unified database of NMR structures that can be used to extract potential relationships between data quality and structural quality. Correlations between various parameters were investigated and summarized in Table III assuming linear correlation: both before (ORG, below diagonal) and after (CNW, above diagonal) recalculation and water-refinement. In general, it appears that, after recalculation and refinement, the correlations increase. This fulfills our initial goal of standardizing the calculation protocols and diminishing software and user biases in the deposited NMR structures. Still, the correlations remain low and only a few of them have an absolute value above 0.5. Some of the most interesting correlations are discussed below.

The strongest correlation is found between the uncertainty present in the restraints and the backbone RMSD ($r = 0.69$) [Fig. 4(a)]. In this plot we marked entries 1G9L, 1KKD and 1BCT as outliers. Entry 1G9L consists of a well-ordered core and a relatively large unstructured part. Entry 1KKD is an unstructured protein with only a few well-ordered residues. Entry 1BCT has two well-defined helices connected by a flexible linker, making a RMSD calculation less meaningful. In Figure 4(a) it becomes clear that the structural uncertainty present in the experimental restraints imposes a lower limit on the RMSD that can be achieved.

In Figure 4(b) the Ramachandran Z-score from WHAT_CHECK is plotted against the circular variance ($r = -0.67$). The outlier in this plot is 1SUT, a 22 amino acid peptide with an unusual helical structure. The two entries in the upper part with circular variance of 0.2 are 1MM4 and 1MM5, which could not be folded properly with the few restraints available (less than one restraint per residue). Here again an upper limit is visible: a protein that has a high Ramachandran normality (less than 2 SDs away from the database average) will have a circular variance smaller than 0.05.

It appears from Table III that the well-ordered RMSD is only weakly correlated to the circular variance (ORG: $r = 0.32$ and CNW: $r = 0.42$). The different nature of these quality indicators accounts for the weak correlation. The circular variance is a local measure only calculated for the well-defined residues (circular variance $<$0.2; see Materi-

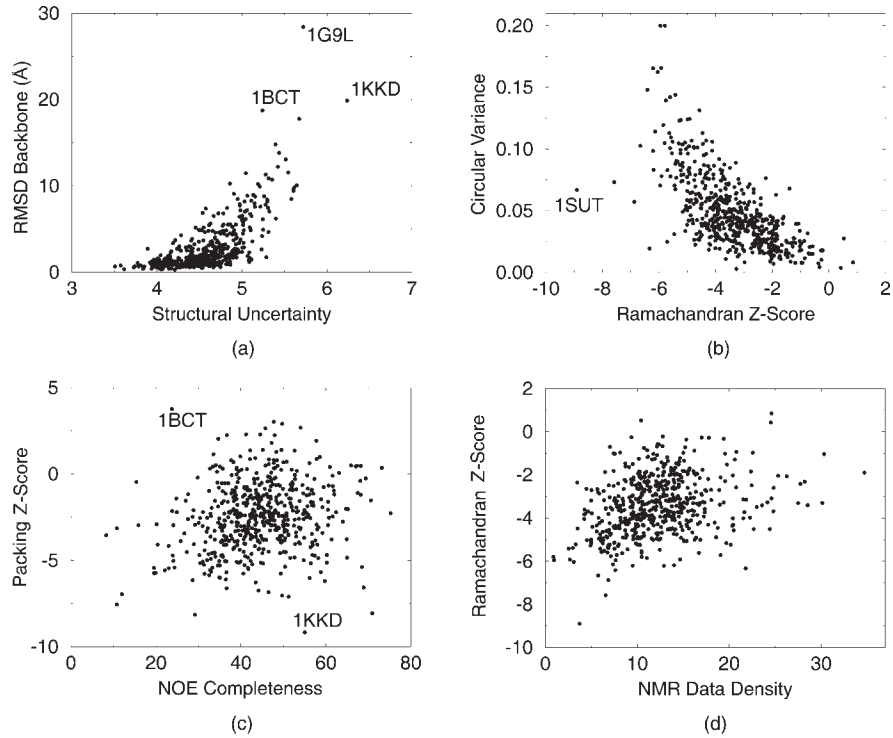

Fig. 4. Correlation plots for various quality indicators in the CNW set.

TABLE III. Correlation Coefficients of Different Parameter

ORG[a] \ CNW[b]	Data Dens.	Compl.	Uncertainty	RMS Viol.	B. RMSD	W. RMSD	C.V.	Packing	Ramach.	Bumps
Data density		**0.56**[c]	−0.14	−0.23	−0.31	−0.31	−0.45	0.34	0.31	−0.03
NOE completeness	**0.58**		−0.29	−0.36	−0.21	−0.17	−0.18	0.19	0.26	−0.05
Structural uncertainty[d]	−0.14	−0.32		0.04	**0.69**	**0.54**	0.23	−0.22	−0.11	−0.27
RMS violations	−0.11	−0.19	−0.10		0.08	0.21	0.22	−0.25	−0.37	**0.58**
Backbone RMSD	−0.24	−0.12	**0.60**	0.02		**0.54**	0.30	−0.39	−0.21	−0.08
Well-ordered RMSD	−0.18	−0.03	0.39	0.01	0.45		0.42	−0.31	−0.27	0.03
Circular variance	−0.32	−0.04	0.09	0.00	0.21	0.32		**−0.60**	**−0.67**	0.25
Packing quality	0.31	0.21	−0.23	−0.06	−0.33	−0.18	−0.49		**0.69**	−0.39
Ramachandran	0.16	0.22	−0.01	−0.11	−0.13	−0.16	−0.48	0.48		**−0.51**
# Bumps per 100 res.	0.04	−0.26	−0.06	0.04	−0.10	0.00	0.07	−0.21	−0.47	

[a]Correlation coefficients for the set ORG are listed below the diagonal.
[b]Correlation coefficients for the set CNW are listed above the diagonal.
[c]Correlations larger than 0.5 or smaller than −0.5 are printed in bold.
[d]The uncertainty present in the restraints does not depend on the coordinates and is thus the same in sets ORG and CNW.

als and Methods), whereas the well-ordered RMSD is a more global measure and might still include residues that are highly mobile if the ensemble as a whole is not well-defined.

A correlation between the RMS distance restraint violations and the number of bumps is present in the CNW set (r = 0.58) and absent in the ORG set (r = 0.05). This correlation in the CNW indicates that strain caused by NOE restraints can easily cause close contacts in the structure.

The correlation between packing and Ramachandran Z-scores for the CNW set (r = 0.69) is expected, but still informative, since both highlight distinct structural features. Only weak correlations are observed between indicators based on the NMR restraints, such as data density, NOE completeness and uncertainty and the WHAT_CHECK quality Z-scores for packing and Ramachandran. This shows that the NMR data based indicators are generally ineffective at indicating the normality of a structure, which can be determined from high quality X-ray structures. Two examples are presented in Figure 4(c and d). In Figure 4(c), the correlation between the NOE completeness and the packing quality Z-score is shown (r = 0.20). Entry 1BCT is an outlier, since it is almost entirely helical (helical content of 84%). Entry 1KKD is another outlier, since it is highly unstructured, as mentioned earlier. This is not reflected in the NOE completeness, which is as high as 55%, as this measure is based on the coordinates of the model itself. In Figure 4(d), the NMR data density is plotted against the Ramachandran Z-score (r = 0.31); this again illustrates the rather weak correlation between data and structural quality.

Surprisingly, the correlations between the data density and the various quality indicators decrease when they are based only on long-range NOEs (data not shown), which were shown to be the most important ones to define the 3D structure.[18] This indicates that, in general, the force field and protocols used perform well in terms of Ramachandran and packing normality, irrespective of the number of restraints provided.

We further analyzed the normality of beta and α structures in our dataset. For α-proteins (defined as α-helical content >40% and sheet content <5% in the ORG set) the packing and Ramachandran Z-scores are −1.5 and −2.9 respectively, in the CNW set. These numbers are much closer to the mean of the X-ray database than the numbers for the entire database (−2.5 and −3.4 for CNW, see Table I). For β-proteins (defined as α-helical content <10% and sheet content >30% in the ORG set) this effect is less apparent: the packing and Ramachandran Z-scores are −2.1 and −3.7, respectively, in the CNW set. These numbers suggest that α-helical structures solved by NMR will in general be closer to the database average regarding packing and Ramachandran appearance. This is somewhat surprising considering that β-proteins show a better dispersion of signals in NMR spectra. In addition, the collection of long-range NOE connectivities involving backbone protons is typically easier than for α-proteins in which sidechain–sidechain contacts are needed to pack the helices against each other. On the other hand, these differences might reflect the wider range of angles in the definition of β-sheet conformations compared to α-helical ones.

The entries in our database were deposited from 1990 to 2003 [see Fig. 1(d)]. There exists no correlation with an absolute coefficient larger than 0.15 between the year of deposition and any of the indicators listed in Table III both for the ORG and CNW sets. For example, there is a very weak trend indicating an increase in the number of restraints per residue with the year of deposition (r = 0.08). This trend might have become higher if RDCs were included.

CONCLUSION

The RECOORD database introduced here contains recalculated and refined coordinates for 545 proteins. The quality indicators for packing and Ramachandran appearance moved on average one SD closer to the mean of the reference database. Only 6% of the entries did not improve in our protocol, mainly due to incomplete deposited experimental data compared to the ones reported in the corresponding papers. The agreement with the experimental restraint data is also generally better than in the originally deposited models.

Although the basic limitation on further improving the structural quality seem to lay on the restraint data available, new calculation algorithms may improve the normality scores of the entries that could be reached in this paper. The use of additional data not considered here such as RDCs and chemical shifts should also lead to further improvement.

The RECOORD database provides a benchmark for future efforts to improve the tools used for NMR structure determination and validation. RECOORD also can serve as a unified high quality NMR structure database for various modeling and statistical applications.

ACKNOWLEDGMENTS

This work was supported by the European Community program NMRQUAL (contract number QLG2-CT-2000-0313) and by grant P50-GM64598 from the Protein Structure Initiative of the National Institute of General Medical Sciences, National Institute of Health. We thank Dr Rolf Boelens for stimulating discussions.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Laskowski RA. Structural quality assurance. Methods Biochem Anal 2003;44:273–303.
3. Wüthrich K. NMR of Proteins and Nucleic Acids. New York: John Wiley & Sons; 1986.
4. Güntert P. Structure calculation of biological macromolecules from NMR data. Q Rev Biophys 1998;31:145–237.
5. Havel TF. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Prog Biophys Mol Biol 1991;56:43–78.
6. Braun W, Go N. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. J Mol Biol 1985;186:611–626.
7. Kaptein R, Zuiderweg ER, Scheek RM, Boelens R, van Gunsteren WF. A protein structure from nuclear magnetic resonance data. *lac* repressor headpiece. J Mol Biol 1985;182:179–182.
8. Brünger AT. X-PLOR Manual (Version 4.0). Department of Molecular Biophysics and Biochemistry: Yale University; 1996.
9. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
10. Güntert P, Mumenthaler C, Wüthrich K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
11. Doreleijers JF, Raves ML, Rullmann T, Kaptein R. Completeness of NOEs in protein structure: a statistical analysis of NMR. J Biomol NMR 1999;14:123–132.
12. Nabuurs SB, Spronk CAEM, Vriend G, Vuister GW. Concepts and tools for NMR restraint analysis and validation. Concepts in Magnetic Resonance 2004;22A:90–105.
13. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.
14. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature 1996;381:272.
15. Doreleijers JF, Rullmann JA, Kaptein R. Quality assessment of NMR structures: a statistical survey. J Mol Biol 1998;281:149–164.
16. Spronk CA, Linge JP, Hilbers CW, Vuister GW. Improving the quality of protein structures derived by NMR spectroscopy. J Biomol NMR 2002;22:281–289.
17. Linge JP, Williams MA, Spronk CA, Bonvin AM, Nilges M. Refinement of protein structures in explicit solvent. Proteins 2003;50:496–506.
18. Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AMJJ, Vuister GW, Vriend G, Spronk CAEM. DRESS: a database of REfined solution NMR structures. Proteins 2004;55:483–486.
19. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL. 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 2003;26:139–146.

20. Doreleijers JF, Nederveen AJ, Vranken W, Lin J, Bonvin AMJJ, Kaptein R, Markley JL, Ulrich EL. BioMagResBank databases DOCR and FRED with converted and filtered sets of experimental NMR constraints and coordinates from over 500 protein PDB structures. J Biomol NMR 2004, in press.
21. Engh RA, Huber R. Accurate bond and angle parameters for X-ray proteinstructure refinement. Acta Crystallogr A 1991;47:392–400.
22. http://www.accelrys.com/insight/discover.html.
23. Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wüthrich K. Recommendations for the presentation of NMR structures of proteins and nucleic acids–IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. Eur J Biochem 1998;256:1–15.
24. Fogh R, Ionides J, Ulrich E, BoucherW, VrankenW, Linge JP, Habeck M, Rieping W, Bhat TN, Westbrook J, Henrick K, Gilliland G, Berman H, Thornton J, Nilges M, Markley J, Laue E. The CCPN project: an interim report on a data model for the NMR community. Nat Struct Biol 2002;9:416–418.
25. Nilges M. A calculation strategy for the structure determination of symmetric dimers by ^{1}H-NMR. Proteins 1993;17:297–309.
26. Güntert, P. Automated NMR protein structure calculation. Prog NMR Spectrosc 2003;43:105–125.
27. Linge JP, Habeck M, Rieping W, Nilges M. ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 2003; 19:315-316.
28. Litzkow M, Livny M, Mutka M. Condor - a hunter of idle workstations. In: Proceedings of the 8th international conference of distributed computing systems, 1988.
29. Thain D, Tannenbaum T, Livny M. Condor and the grid. In: Berman F, Fox G, Hey T, editors. Grid computing: making the global infrastructure a reality. New York: John Wiley & Sons Inc; 2002; p 299–335.
30. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
31. Nabuurs SB, Spronk CA, Krieger E, Maassen H, Vriend G, Vuister GW. Quantitative evaluation of experimental NMR-restraints. J AmChem Soc 2003;125:12026–12034.
32. Nilges M, Clore M, Gronenborn AM. A simple method for delineating well defined and variable regions in protein structures determined from interproton distance data. FEBS Lett 1987;219:11–16.
33. Spronk CA, Nabuurs SB, Bonvin AM, Krieger E, Vuister GW, Vriend G. The precision of NMR structure ensembles revisited. J Biomol NMR 2003;25:225–234.
34. Rice LM, Brunger AT. Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. Proteins 1994;19:277–290.