Journal of Biomolecular NMR, 8 (1996) 477–486
ESCOM

J-Bio NMR 390

AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR*

Roman A. Laskowski[a,**], J. Antoon C. Rullmann[b,**], Malcolm W. MacArthur[a], Robert Kaptein[b] and Janet M. Thornton[a,***]

[a]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, U.K.*
[b]*Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands*

Received 10 July 1996
Accepted 9 August 1996

Keywords: Computer software; Structure validation; Restraint analysis; Protein geometry

Summary

The AQUA and PROCHECK-NMR programs provide a means of validating the geometry and restraint violations of an ensemble of protein structures solved by solution NMR. The outputs include a detailed breakdown of the restraint violations, a number of plots in PostScript format and summary statistics. These various analyses indicate both the degree of agreement of the model structures with the experimental data, and the quality of their geometrical properties. They are intended to be of use both to support ongoing NMR structure determination and in the validation of the final results.

Introduction

The past few years have seen remarkable progress in the methodology of solving protein structures by solution NMR. A wide spectrum of experimental techniques, refinement protocols and computer programs for structure determination is now available. Simultaneously the number of NMR structures deposited in the Brookhaven Protein Data Bank, PDB (Bernstein et al., 1977), has steadily increased to the current level of more than 500 entries, i.e., almost 15% of the total number of coordinate entries in the PDB.

With so many NMR structures now in the PDB, and with their rate of deposition rising rapidly, it is important to have some assessment of each structure's 'quality' – that is, how reliable it is as a true and accurate representation of the molecule(s) in question. For example, for X-ray crystal structures the resolution and R-factor give a rough measure of the accuracy that can be expected of the corresponding protein model and of the reliance that can be placed on it, while, locally, the atomic B-factors and occupancies can give an indication of which are the structure's more ordered and disordered regions.

For NMR structures, however, comparatively little attention has been given to methods for validating and assessing their 'quality'. Most of the effort has been directed at quantifying the 'precision' of the structure determination rather than its accuracy. The overall precision of an NMR structure is usually expressed either as an average pairwise root-mean-square deviation (rmsd) of the coordinates across the members of the final ensemble of structures, or an rmsd of the structures relative to the mean coordinates. The rmsd value can be calculated over the whole molecule, or only over those segments that are considered to have a 'well-defined' conformation. Widely varying estimates have been given for the maximum obtainable precision (Havel, 1991; Clore et al., 1993; Zhao and Jardetzky, 1994). These rmsd values have often been

*The AQUA and PROCHECK-NMR programs and operating instructions are available free of charge. The programs are supplied with script files for running on UNIX operating systems and can be obtained by anonymous ftp from 128.40.46.11 in directory pub/procheck. Further information can be obtained on http://www.biochem.ucl.ac.uk/~roman/procheck_nmr/procheck_nmr.html and http://www-nmr.chem.ruu.nl/users/rull/aqua.html, or by sending e-mail to roman@bsm.bioc.ucl.ac.uk or rull@nmr.chem.ruu.nl.

**R.A.L. and J.A.C.R. are joint first authors.

***To whom correspondence should be addressed.

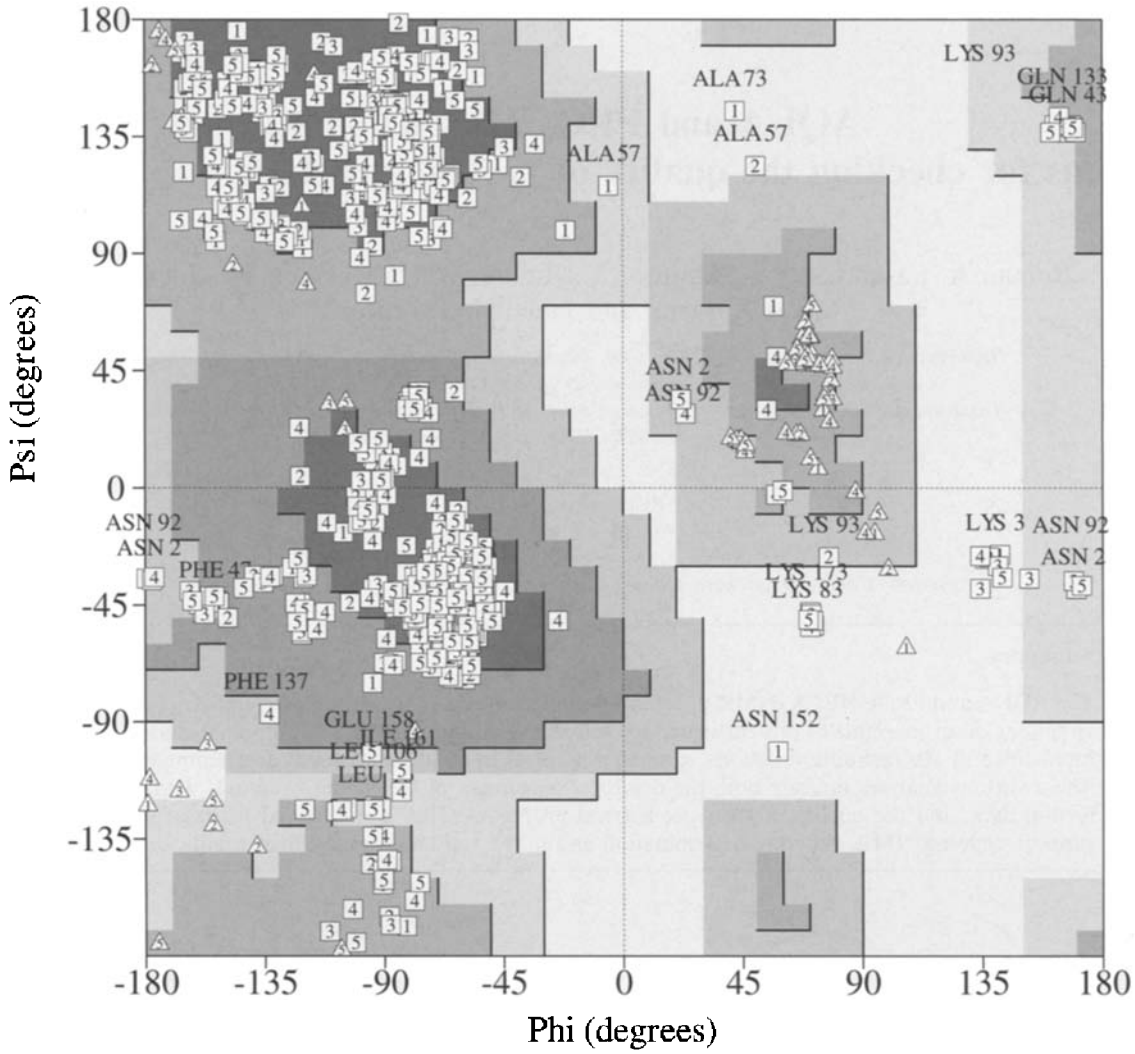

Fig. 1. The Ramachandran plot shows the distribution of ϕ–ψ values for all the residues in the structure. Here, only models 1 to 5 have been selected from the entire ensemble of 25 models. Each data point is labelled with its model number, while the names of any residues in disallowed regions of the Ramachandran plot are printed above their respective points. The shading indicates the favourable and unfavourable regions of the plot, the darker the shading the more favourable the region. A separate plot can be generated for each model in the ensemble, and even for each residue (see Fig. 2).

taken as the single most important quality criterion for NMR structures.

However, the problem with such statistics lies in the question of how representative the ensemble from which they are drawn is; how well has the conformational space been sampled (Hoch, 1991). The final ensemble of structures may correspond to a precise result, rather than an accurate one (Clore et al., 1993). In particular, the rmsd can be made small by adjusting the various parameters used during the different stages of structure refinement, such as the shape of the target function and the values of the force constants used for the various energy terms in restrained energy minimization, simulated annealing and restrained molecular dynamics (Chazin, 1992). Furthermore, there are no generally accepted criteria for selecting structures to be included in the final ensemble, either in terms of how many to select and which ones are the most representative. The size of ensembles currently deposited in the PDB varies from 1 to 77 models.

In addition to the rmsd there are a number of other measures which can also give some indication of the overall quality of an NMR-determined structure. One of the most important is the agreement with the experimental data. This is usually expressed in terms of numbers and sizes of violated distance restraints. Often the total or average restraint violation is quoted, together with the size of the largest violation. In direct refinement studies an R-factor is given (Gonzalez et al., 1991; Thomas et al., 1991).

A measure of the quality of the experimental data itself can be provided by the number of restraints per residue. Various analyses have shown that this is a significant

quality determinant of NMR structures (Clore et al., 1993; MacArthur and Thornton, 1993; Rullmann et al., unpublished data).

The 'stereochemical quality' of the protein provides another measure of quality and one which is independent of the experimental data and, for some parameters, can also be independent of the refinement procedures employed by the authors. This involves checks on the geometrical properties of the protein (e.g. bond lengths, bond angles, dihedral angles, etc.). These are based on comparisons with what is known about standard protein structure and geometry from the wealth of high-resolution X-ray structures already in the PDB. They can assess how 'normal' or 'abnormal' a given model is, compared with the standard values. The best-known example is the Ramachandran plot (Ramachandran et al., 1963) which defines combinations of ϕ–ψ dihedral angles that are favourable, unfavourable, or disallowed. These and several other, so-called 'coordinate-validation' methods have been reviewed recently by MacArthur et al. (1994).

Most of the above quality indicators are now commonly reported when an NMR structure is published. Their use is strongly recommended in a forthcoming report of an IUPAC-IUBMB-IUPAB Interunion Task Group (to be published). While all these indicators are relevant, it is crucial to recognize that individual values, especially of global averages, convey little information. Real insight into the quality of a structure can only be obtained by considering a broad range of indicators, since they are not independent and one quantity may be optimized at the expense of another. Important information may be obtained by considering the correlation between various quantities, e.g. along the residue sequence. This can highlight differences between surface and core residues, or differences between residue types, which are important for understanding the dynamic nature of protein molecules in solution (MacArthur and Thornton, 1993). Correlations between quality indicators may also help to identify local errors in the structure.

In some cases it may be much more relevant to report outlier values than to quote overall averages. For example, where bond lengths, peptide and ring planarity, dihedral angles, etc. are restrained to ideal values, the overall averages provide no test of quality. However, a look at how the values are distributed (i.e. whether the distribution is Gaussian, and how small or large its standard deviation is) can indicate whether the restraints might be too lax or too tight, as well as highlighting significant outliers.

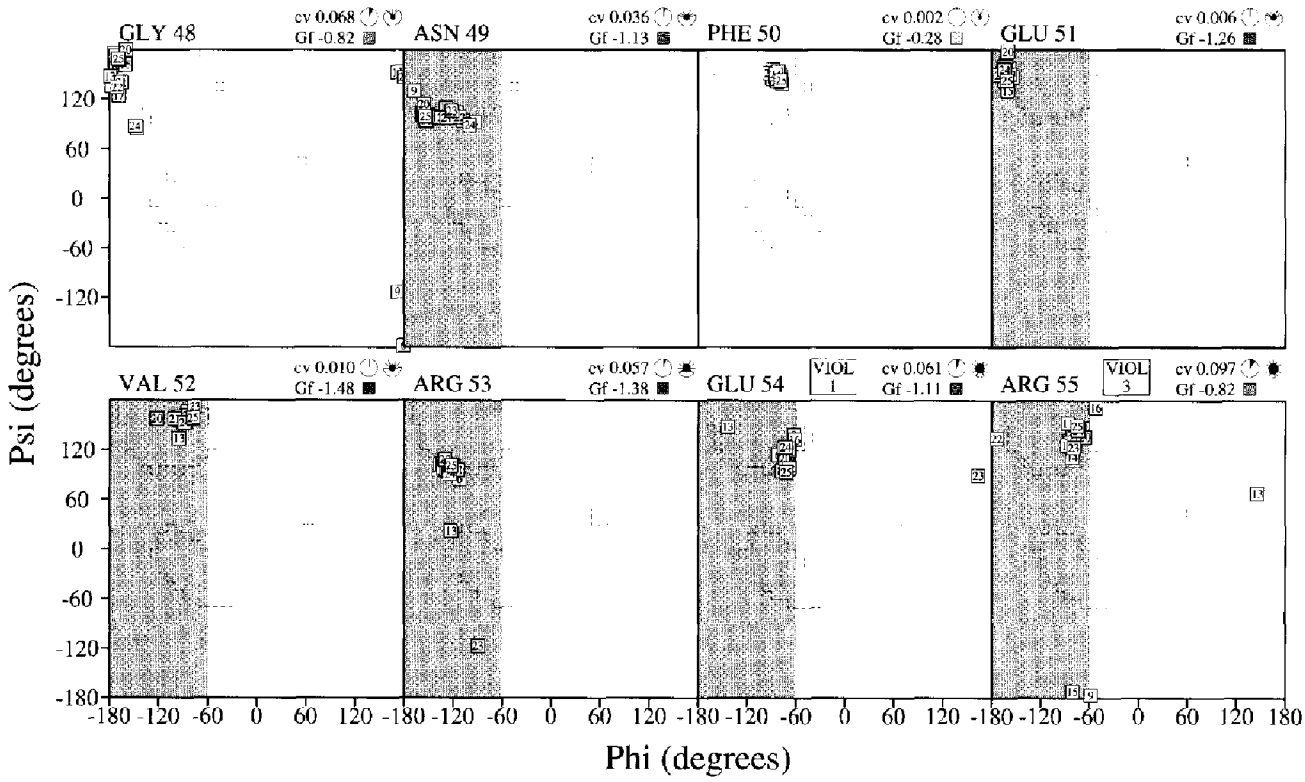

Fig. 2. Individual residue-by-residue Ramachandran plots show the distributions of ϕ ψ values for each residue across all the models of the ensemble. Here, only residues 48 to 55 are shown. Each data point is labelled with its model number. Where either the ϕ or ψ angles have been restrained during structure refinement, the region bounding the restraint limits is shown in a darker shade. The number of any points lying outside this region, corresponding to restraint violations, is shown in the box labelled 'VIOL' above the graph. Also shown above each graph are the circular variance, labelled 'cv' and the G-factor, labelled 'Gf'. Schematic diagrams to the right of these figures give a visual representation of these numbers. Also shown is a schematic depiction of the residue's solvent accessibility, the darker the icon the more solvent accessible it is.

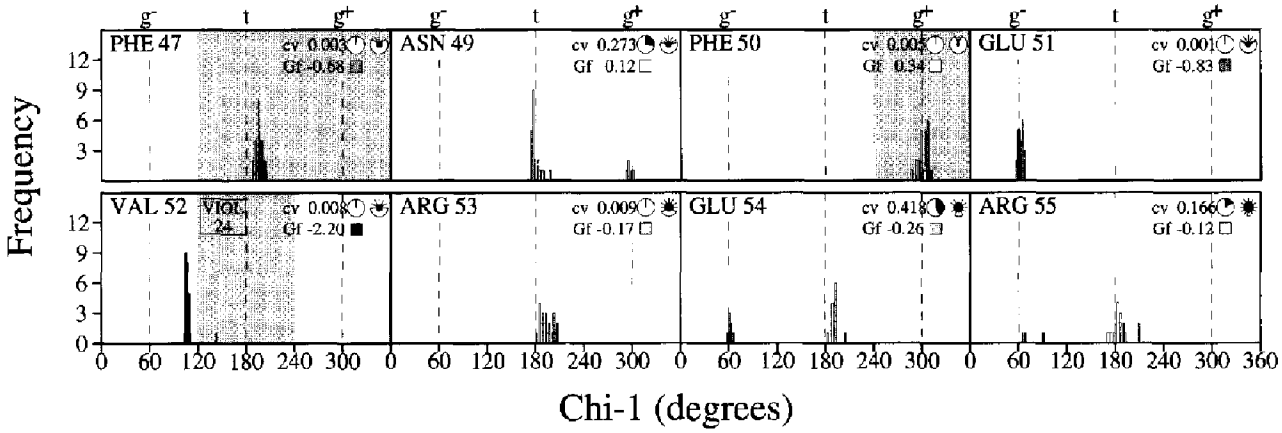

Fig. 3. Histograms of the χ_1 torsion angles for residues 47 to 55 across all 25 models of the ensemble. The dashed lines labelled g^-, t and g^+ correspond to the preferred *gauche minus*, *trans* and *gauche plus* conformations. As in Fig. 2, the dark-shaded regions correspond to the restraints applied during refinement, with the numbers of models in which the values fall outside this range shown in the box labelled 'VIOL'. The other numbers and schematics are as in Fig. 2.

In this paper we describe two related suites of programs, called AQUA and PROCHECK-NMR, which aim to perform a large number of validation checks on a given NMR ensemble of protein structures. The programs generate analyses in terms of tables of data, plots in PostScript format (Adobe Systems Inc., 1985), and summary statistics suitable for reporting in publications. The validation checks compare the final protein models both against the experimental data from which they were generated and against standard stereochemical properties derived from known, high-resolution X-ray crystal structures.

The analyses produced by the programs can complement those already generated by the different structure calculation packages used in the NMR field (Sutcliffe, 1993). Furthermore, as the AQUA/PROCHECK-NMR analyses are independent of the method of structure determination, exactly the same analyses can be performed on different structures regardless of how they were solved.

Another important characteristic of the programs is that they are not restricted to a specific type of analysis, but present a comprehensive overview of the quality of a set of structures: covalent geometry, torsion angles, chirality, planarity, precision (both rmsd and circular variance), accessibility, and distributions of restraints and restraint violations. The results are presented as plots suitable for publication. But the results can also be used to support ongoing structure determinations: the plots highlight the problematic regions using a number of different indicators which can be compared along the residue sequence.

Methods

The first set of programs, AQUA – Analysis of QUAlity (J.A.C. Rullman) – analyses the restraints obtained from the experimental data and computes the restraint violations between the models and data. The restraints will primarily consist of the interproton distance ranges representing the observed NOE (nuclear Overhauser effect) cross peaks. The NOE restraints may be complemented by amide hydrogen-exchange data, coupling constant measurements, known disulphide linkages and other restraints (such as those used to model bound metal ions).

The programs can interpret a number of commonly used restraint-file formats; for example, DISGEO (Havel and Wüthrich, 1984), X-PLOR (Brünger, 1992), BIOSYM (Biosym Technologies Inc., San Diego, CA, U.S.A.) and DIANA (Güntert et al., 1991). Where the restraints involve pseudo-atoms, AQUA generates their coordinates. Presently, this is done by calculating the geometric average of the corresponding proton positions.

The outputs provide detailed breakdowns of the restraints and their violations (largest violations, total violations, etc.) by model, by restraint and by residue. Various tools allow one to extract data on a keyword basis or produce more condensed summaries – e.g. average and rmsd violation values, numbers of (violated) restraints, sorted lists of first n largest violations, and so on.

The second set of programs, PROCHECK-NMR (R.A.L. and M.W.M.), links directly to the outputs generated by AQUA and produces a large number of colour, or black-and-white, plots in PostScript format. The plots are of two types: the first comprises analyses and comparisons of the geometry of the model structures making up the NMR ensemble, and the second consists of analyses of the restraints and restraint violations.

The first group of plots is an extension of the PROCHECK programs (Laskowski et al., 1993) used for assessing the stereochemical quality of X-ray structures. The extensions are primarily geared to coping with ensembles of model structures rather than with single struc-

tures. Like the original PROCHECK programs, the NMR versions make use of the stereochemical parameters found by Morris et al. (1992) to be good indicators of geometrical quality. Other parameters, specific to ensembles of structures, are also used, such as the circular variance of dihedral angles (Allen and Johnson, 1991; MacArthur and Thornton, 1993) which provides estimates of the precision of the local structure.

The second group of plots generated by PROCHECK-NMR analyses the restraints and restraint violations computed and tabulated by AQUA. These show individual restraints (and violations), analyses by residue and overall summaries covering the entire ensemble of model structures. The plots can give a picture of which regions of the protein are poorly defined in terms of numbers of restraints, and which regions have a high degree of restraint violation. These features can be visually compared with the regions of the protein where the geometry may be unfavourable or unusual, as shown in the first group of plots.

Results and Discussion

Figures 1 to 8 show examples of some of the plots produced by PROCHECK-NMR. The structure used is the solution structure of the histone-like HU protein from *Bacillus stearothermophilus* (Vis et al., 1995) and the ensemble comprises 25 models. It comes from an early stage of refinement and deliberately shows more restraint violations than in the final published coordinates. The plots shown are all in black-and-white. It should be noted that the colour versions usually give a much clearer picture of

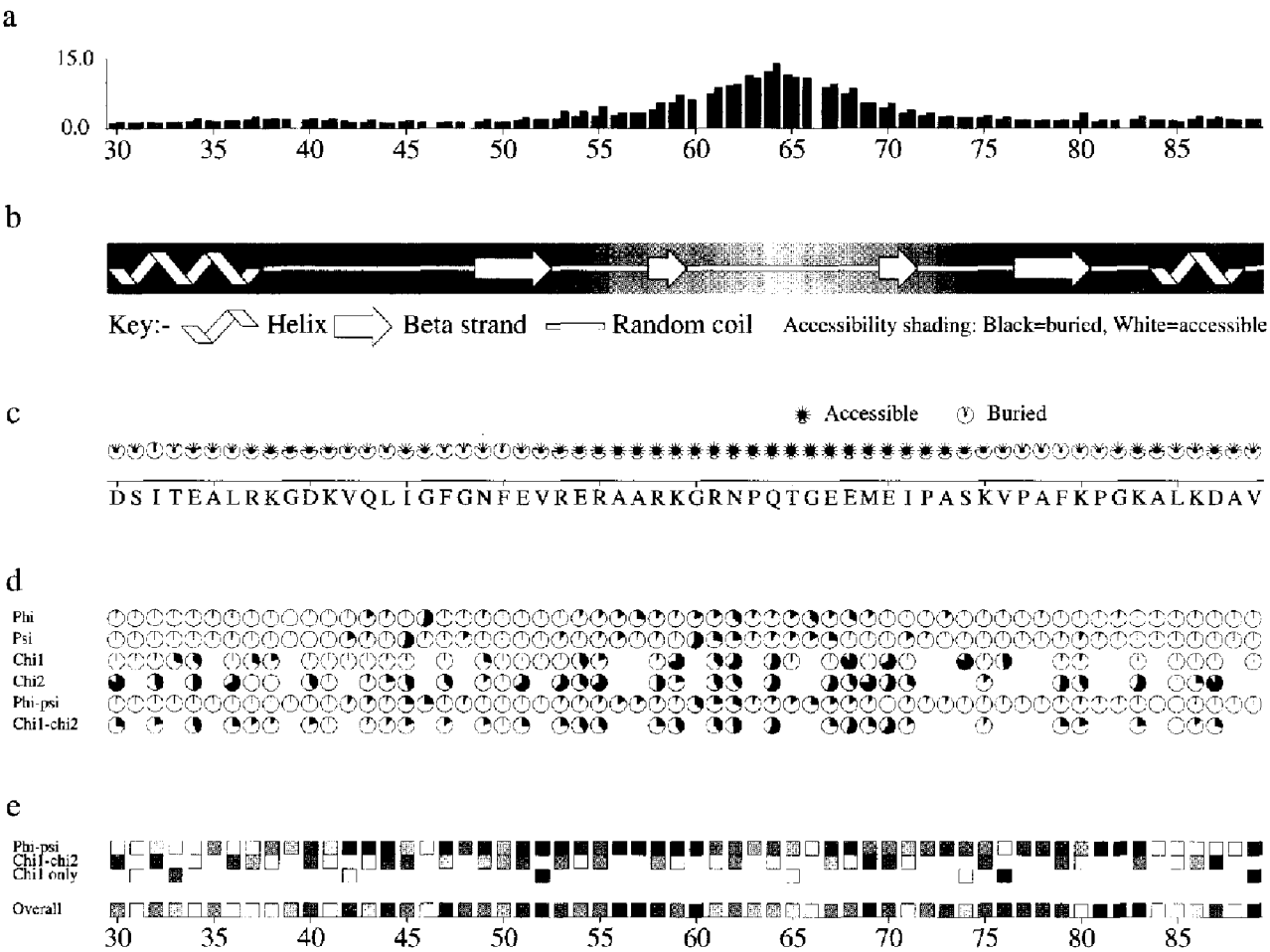

Fig. 4. A summary of the ensemble geometry for residues 30 to 89 of the test structure. The top graph (a) shows the rms deviations of main-chain (black) and side-chain (grey) heavy atoms from the mean coordinates of the structure, exhibiting a large peak corresponding to one of the structure's flexible loops. The schematic picture in (b) shows the protein's secondary structure, in particular the α-helices and β-strands, as defined using the Kabsch and Sander (1983) assignments, and averaged across all the members of the ensemble. In (c) the protein's sequence is shown together with schematic symbols showing each residue's solvent accessibility. The dark regions correspond to surface loops, while the lighter ones represent buried residues. The dials in (d) show the circular variances for the dihedral angle distributions: ϕ, ψ, χ_1 and χ_2; and combinations: ϕ–ψ and χ_1–χ_2. These indicate how tightly clustered or spread-out these values are across the ensemble; the larger the black area on the dial the greater the spread (see for example Figs. 2 and 3). The shaded boxes in (e) represent the G-factors for the ϕ–ψ and χ_1–χ_2 dihedral angle distributions (or the χ_1 only distribution for those residues without a χ_2). The darker the square the more 'unusual' are the corresponding dihedral angle values for the residue-type in question (as, for example, for the ϕ–ψ distributions of Ala[57], Pro[81] and Lys[83], cf. Fig. 1, and the χ_1–χ_2 distribution of Met[69]).

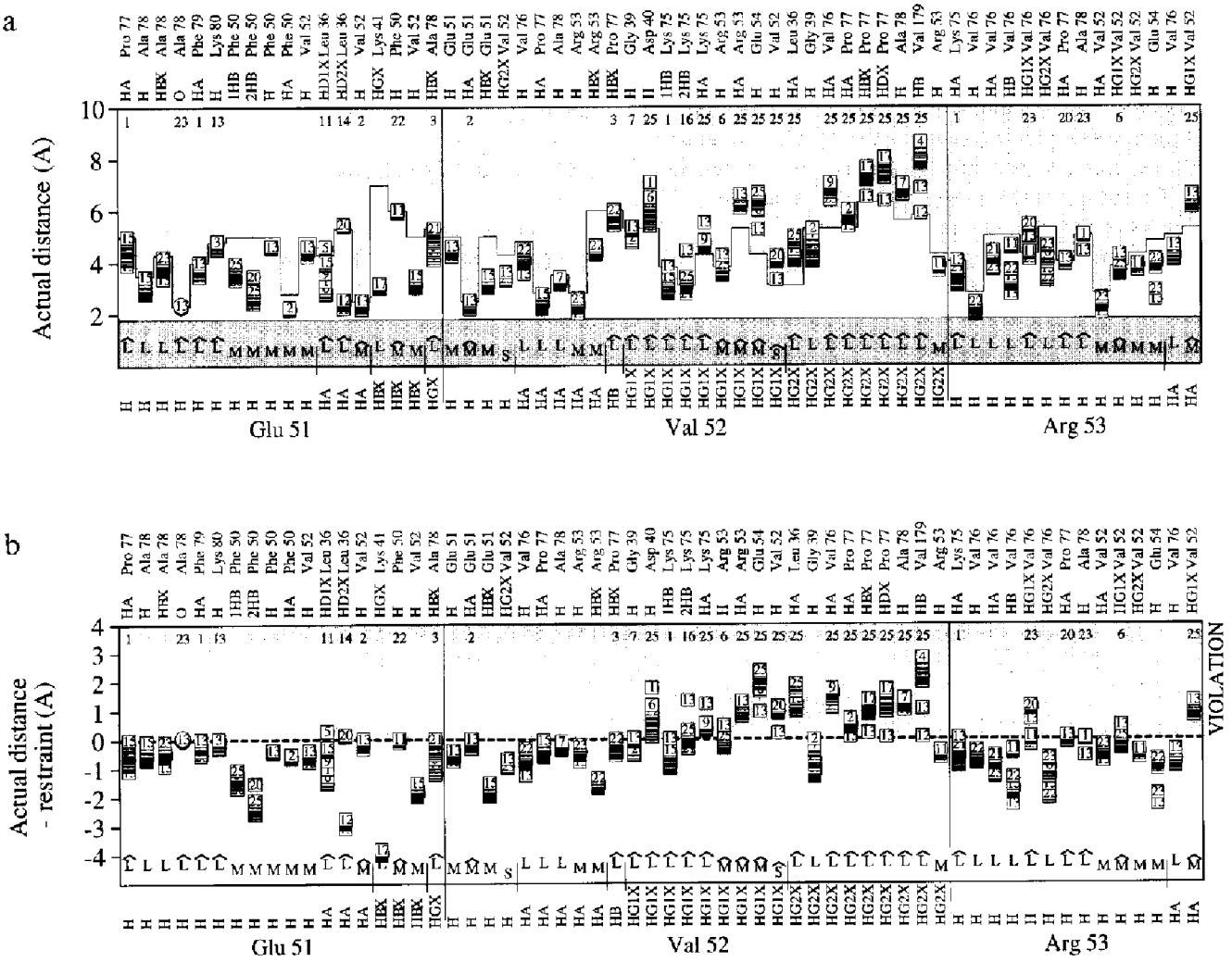

Fig. 5. Two plots showing the individual distance restraints used in solving the NMR structure, including any violations of those restraints. Here, only the restraints for Glu[51], Val[52] and some of those for Arg[53] are shown. The top graph (a) shows the actual distances in all the NMR models corresponding to each of the restraints. Each data point is labelled with its model number. The shaded regions at the top of each plot correspond to the upper distance bounds, while those at the bottom of each plot correspond to the lower-distance bounds. Any data points that appear in these shaded regions represent upper- or lower-bound violations, respectively. The atoms between which the restraints apply are named at the top and bottom of the graph; atom names ending in 'X' correspond to pseudo-atom positions. The atoms on the bottom are grouped by residue and atom type. So, for example, for Val[52], one can see that nearly all the restraints involving the HG1X and HG2X pseudo-atoms are violated, many being violated in all 25 models. The square point-markers correspond to NOE distance restraints while circles represent H-bond distance restraints. Each restraint is classified on the plot as 'short-, medium- or long-range' (S, M or L, respectively), according to the sequence separation of the residues involved: short-range are restraints within the same residue, medium-range are those between residues whose sequence separation along the polypeptide chain is four residues or less, and long-range are those where the sequence separation is greater than four residues. A caret above the S, M or L indicates that the restraint is violated in one or more of the models, the exact number of violations being shown at the top of the graph. The lower graph (b) shows the same restraints as in (a), but here the values plotted correspond to the differences between the actual and restraint distances. Positive values in the upper, darker-shaded part of the plot show the magnitudes of any restraint violations. This brings out the violations in the Val[52] residue more clearly. Also of interest are any large negative values which correspond to restraints that are much larger than the actual distances between the atoms in question as it may indicate the restraints are too slack (for example, the restraint between HBX of Glu[51] and the HGX of Lys[41]).

possible problem areas in the structures and of the regions that need to be looked at more closely.

Figure 1 shows a Ramachandran plot for the first five models only. As in PROCHECK, all the outliers in the unfavourable regions are labelled. The appearance of the plot can be changed by specifying smaller or filled-in data points, no model numbers, etc. It is possible to select which models and which residues are to be included in all the AQUA and PROCHECK-NMR analyses and plots. Additionally, in the case of the Ramachandran plot, separate plots can be produced for each model in the ensemble, for each of the 20 residue types, and even for each residue in the protein, as shown in Fig. 2.

Various dihedral angles can be compared on a residue-

by-residue basis, with the plots showing how tightly clustered or spread out the values are across the ensemble. Figure 3 shows an example for the χ_1 torsion angle. The degree of spread of each residue's χ_1 values across the 25 models of the ensemble is quantified by the 'circular variance' (Allen and Johnson, 1991; MacArthur and Thornton, 1993) which, for a given dihedral angle θ, is defined as:

$$\mathrm{Var}(\theta) = 1 - R_{av} \qquad (1)$$

where $R_{av} = R/n$, the parameter R being given by the expression:

$$R^2 = \left(\sum_{i=1}^{n} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n} \sin\theta_i\right)^2 \qquad (2)$$

where n is the number of members in the ensemble.

The value of the circular variance varies from 0 to 1, with the lower the value the tighter the clustering of the values about a single mean value (as, for example, for Glu[51] in Fig. 3 which has a circular variance of 0.001).

Also shown on each plot is the average 'G-factor' for each residue's χ_1 values. The G-factor provides a measure of how 'normal', or alternatively how 'unusual', a given stereochemical property is. The properties for which G-factors are computed in PROCHECK-NMR are: the residue's ϕ–ψ combination, its χ_1–χ_2 combination, and its χ_1 value.

The G-factor is essentially a log-odds score based on the observed distribution of the given property in high-resolution X-ray crystal structures. The dataset of structures comprised 163 nonhomologous protein chains chosen from structures solved by X-ray crystallography to a resolution of 2.0 Å or better and an R-factor no greater than 20%. No two of the 163 chains shared a sequence homology greater than 35%, and all atoms having zero occupancy were excluded from the analysis.

For example, to derive the G-factors for the ϕ–ψ combinations on the Ramachandran plot a separate plot was built up for each of the 20 different residue types. Each of these 20 plots was divided into 45×45 cells and the numbers of observations in each cell gave the probability of the given residue-type having that particular range of ϕ–ψ values. The probabilities were, in turn, converted to log-odds scores and then normalised across the different residue types to give comparable ranges of G-factor values.

When applied to a given residue, a low G-factor indicates that the property corresponds to a low-probability conformation. So, for example, in Fig. 3 the χ_1 values of Val[52] have low G-factors (average = −2.20) and hence are in unusual conformations, whereas the χ_1 values of Asn[49] appear to be in favourable conformations (average = 0.12). Thus, although the dihedral angles might be clustered very tightly (as indicated by a small circular variance), they may all be in a most unusual conformation (as is the case for Val[52]).

Additionally in Fig. 3, where a given residue's dihedral angles have had restraints applied, the ranges are shown as darker regions on the plot (as for Phe[47], Phe[50] and Val[52]). It is sometimes noticeable, as for Val[52] in Fig. 3, how the values cluster at the borders of, or just outside, these restraint ranges, as though trying to escape from them and suggesting the ranges may be incorrect. In the case of Val[52], most of the models violate the applied χ_1 restraint. We will see later that this is probably related to violations of this residue's NOE distance restraints. Plots similar to Fig. 3 can be produced for the ϕ, ψ and χ_2 dihedral angles.

Figure 4 shows a summary of various geometrical properties along the residue sequence giving a visual guide to which regions may be suspect. The properties plotted include the rms deviations of the main-chain and side-chain atoms from the mean coordinates of the whole ensemble, a schematic diagram of the protein's secondary structure, and the various dihedral angle circular variances and G-factors. It can be seen straight-away in Fig.

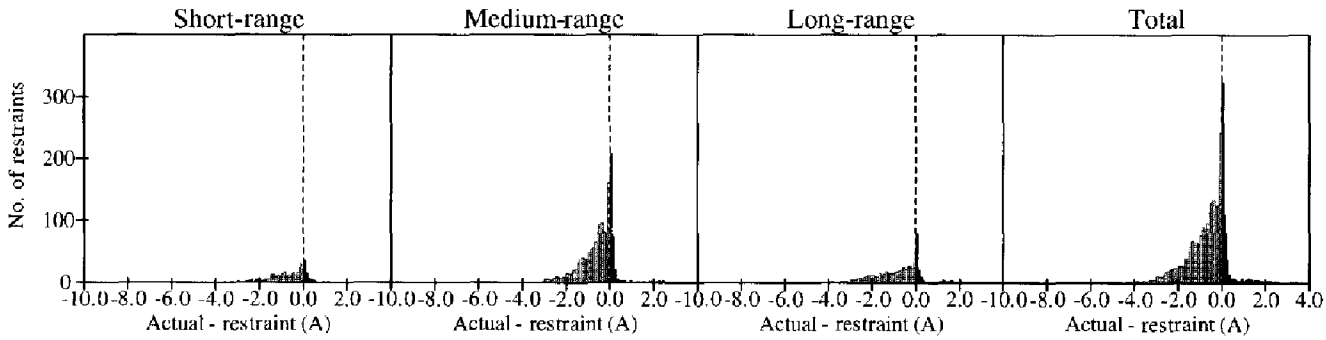

Fig. 6. Histograms of the differences between the actual interatomic distances and corresponding restraint distances. The positive x-axis values in each graph (black bars) correspond to the magnitudes of any upper-bound (and lower-bound) violations. The negative values (grey bars) give an indication of the excess of the restraint distances over the actual distances. Thus, many large negative values would suggest that the restraints may have been set too loosely. Essentially, the plots summarise the data shown in Fig. 5b. The three left-most graphs show the restraints classified into short-, medium- and long-range, as in Fig. 5, with the right-most graph giving the distribution over all the distance restraints.

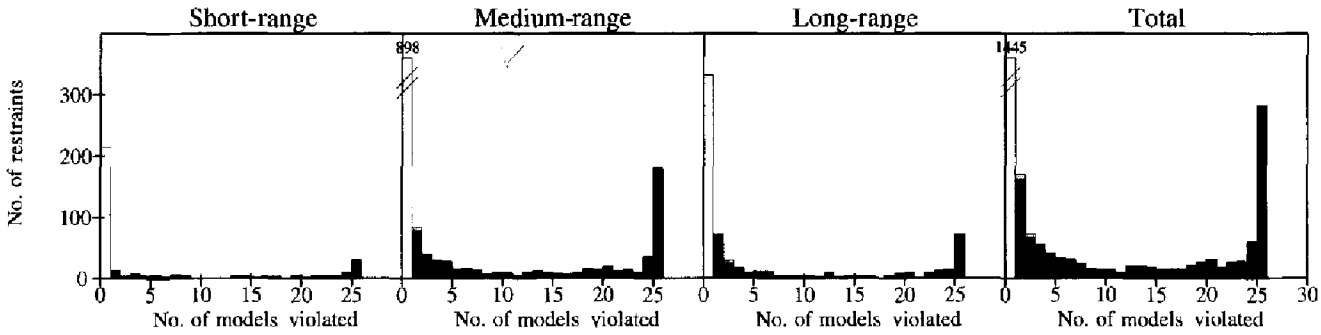

Fig. 7. Plots showing the violation frequencies of the distance restraints in the test structure. The black bars show the numbers of upper-bound distance restraints that are violated in 1, 2, ..., 25 of the models. The tall white bars represent the numbers of restraints that are not violated in any model. The small grey bars correspond to lower-bound distance violations. As in Fig. 6, the plots show the restraints classified into short-, medium- and long-range restraints, with the right-most graph giving the distribution over all the distance restraints. It can be seen that around 250 of the distance restraints are violated in all 25 models.

4 that there is a flexible loop between residues 60 and 69 exhibiting considerable conformational variation across the ensemble both in terms of atomic rms deviations from the mean coordinates and large variances in the torsion angles. The G-factors, on the other hand, do not find too much that is unusual in the conformation of this region. Conversely, the loop between residues 81 and 84 appears not to vary much across the ensemble, yet has unusual main-chain geometry.

In Figs. 5a and 5b, the individual distance restraints are shown, grouped by residue, and compared with the corresponding actual distances in all 25 models of the ensemble. In the top graph, Fig. 5a, the ranges defining the lower- and upper-bound of each restraint are shown as the darker shaded areas, while the actual distances in each of the models are plotted relative to this range. The numbers of violations (i.e. points outside the bounds of these ranges) are shown at the top of the plot. In the bottom graph, Fig. 5b, the differences between these actual distances and the upper-bound restraint are shown. Here, large positive values highlight large restraint violations while, conversely, large negative values may suggest that the restraints may be too slack or overgenerous. Figure 6 shows the distribution of positive and negative values from this bottom graph. This distribution can give an overview of how bad the worst violations are and/or how overgenerous some of the slacker restraints might be. In these graphs the restraints are classified into short-, medium- and long-range, according to the sequence separation of the residues involved, as explained in the figure legends. An alternative classification, based on distance ranges, can also be defined by the user.

Figure 7 shows a summary of the restraint violations across the ensemble of NMR structures. It shows how many distance restraints are violated in all 25 models, how many in 24 of the 25, and so on, down to how many restraints are not violated in any of the 25 models (white bars).

Finally, Fig. 8 shows a summary of the numbers of restraints and their restraint violations for each residue along the sequence. This shows the dihedral angle restraints as well as the distance restraints and can highlight

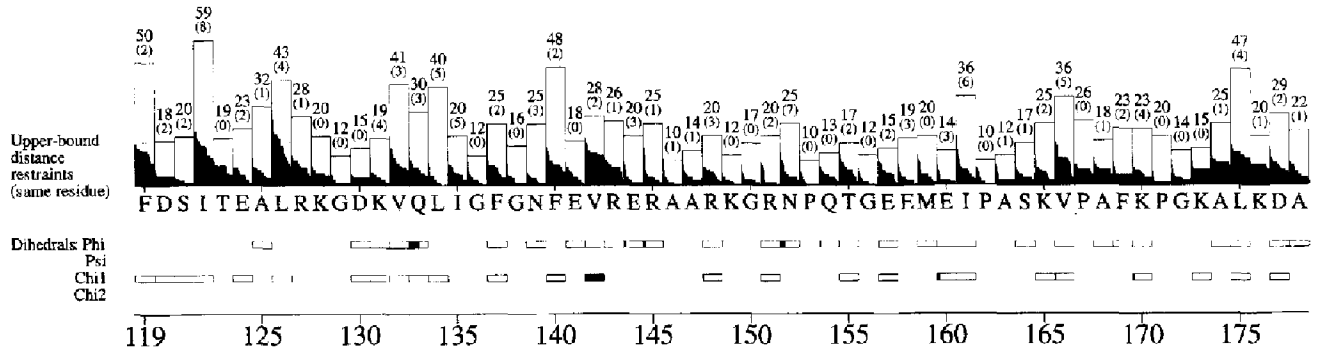

Fig. 8. Numbers of distance and dihedral angle restraints, and their violations, for residues 119 to 178. The number at the top of each stack gives the number of distance restraints for the residue, which is also represented by the height of the stack; the number in brackets just below gives the number of intraresidue restraints within the residue. The black regions illustrate the number of violated restraints. The regions are made up of horizontal bars stacked on top of one another, each bar corresponding to a violated restraint. The width of each bar is proportional to the number of models in which the restraint is violated. The restraints violated in the largest number of models are shown at the bottom. The higher bars correspond to those restraints violated in fewer and fewer numbers of models, and so stretch across only part of the width. Below the stacks are the single bars that indicate which residues have dihedral angle restraints and the degree of violation of these.

poorly defined or consistently violated regions of the protein structure. As mentioned above, the examples in Figs. 1 to 8 are only a subset of those produced by the programs.

In addition to these plots, PROCHECK-NMR can also generate PDB-format files that allow one to view the restraints and restraint violations in three dimensions using standard molecular modelling software. Figure 9 shows an example for a single residue, displayed in QUANTA™ (Molecular Simulations Inc., Burlington, MA, U.S.A.). It depicts a liquorice-bond representation of Val^{52} together with the 3D representations of its restraint violations. The restraints are represented by red-tipped white bars in which the combined length of each pair of red tips is equivalent to the size of the violation. The two ends of the side chain have very many significant violations. As mentioned above, the Val^{52} residue's χ_1 dihedral angle restraint is violated in most of its models (Fig. 3). In Fig. 9 it can be seen that the distance restraints on the residue's two ends appear to be pulling from opposite sides of the valine side chain. This suggests that the NOE distance restraints and the χ_1 dihedral angle restraint might either be in conflict, and are pulling the side chain in opposite directions, or have been affected by subsequent renaming of the valine's two C^{γ} atoms without the changes being taken into account in the restraint files.

As well as showing the violations for individual residues, the program can show all the model's violations. This can help identify red 'hot-spots' corresponding to atoms, or whole regions, of high violation. Any combination of short-, medium- or long-range restraints can be selected. Additionally, the satisfied, as well as the violated, restraints will be extracted (and placed in separate PDB-format files) for input into any molecular graphics package. Similar features can be found in the MOLMOL graphics program of Koradi et al. (1996), which also includes similar analyses to those presented here though the presentation of the data is somewhat different.

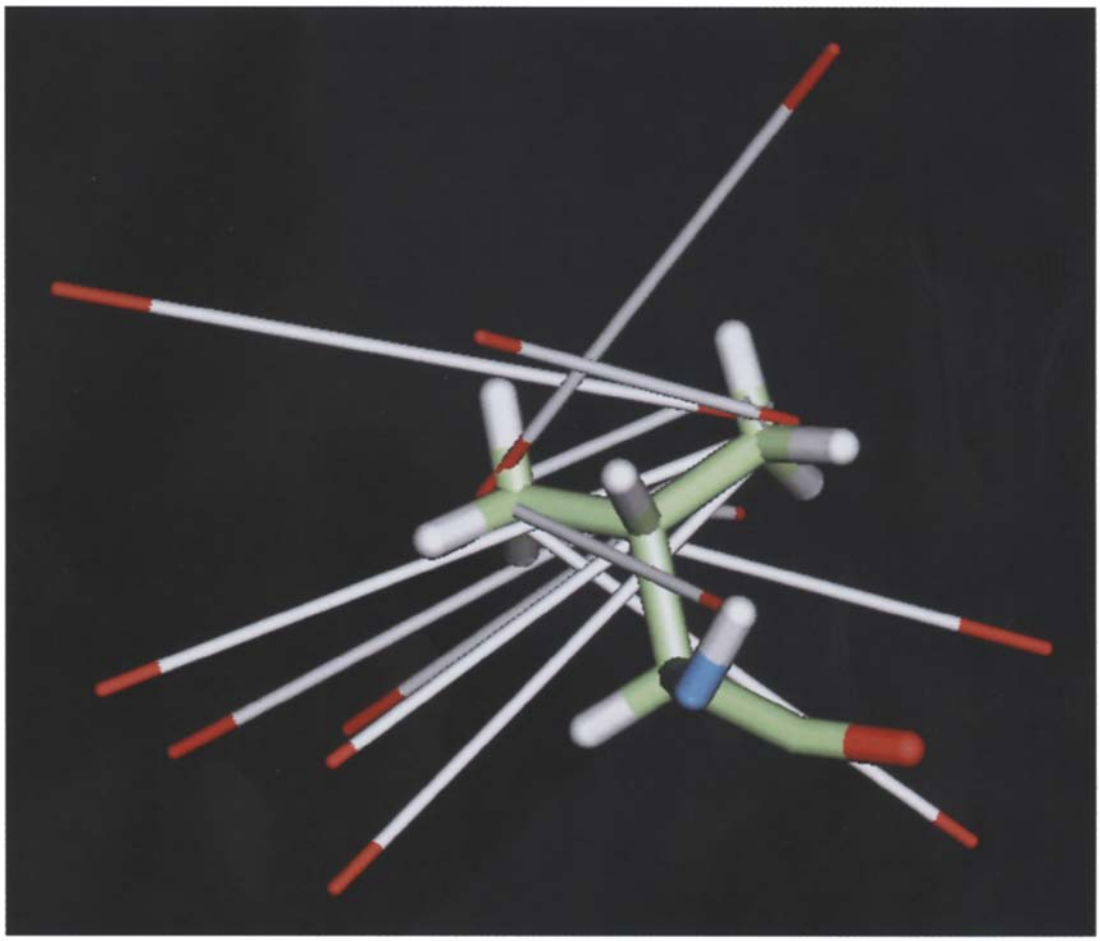

Fig. 9. A liquorice-bond representation of Val^{52} in the test structure plus all its restraint violations shown as red-tipped white bars. The red tips of each bar end on the atoms (or pseudo-atom coordinates) between which the restraint applies (the other residues are not shown here). The length of the white part of each bar corresponds to the size of the distance restraint applied; hence the combined length of the two red tips represents the size of the violation (i.e. the excess of the actual distance over the restraint distance). It can be seen that the restraints from the pseudo-atoms close to the valine's $C^{\gamma 1}$ and $C^{\gamma 2}$ atoms are quite large (around 1–2 Å) and are pulling in opposite directions. This suggests that they might be better satisfied by altering the residue's χ_1 dihedral angle. Figure 3 shows that the χ_1 is already subject to restraint; indeed, it appears to be pulling against the restraints currently applied, suggesting that something is wrong either with the assignments or the atom-labelling.

Conclusions

In conclusion, the analyses reported by the AQUA and PROCHECK-NMR programs should prove useful during the solution and refinement of NMR structures to help check for possible errors. For example, high levels of restraint violations can indicate assignment problems, deficiencies of the algorithms or protocol errors. The programs can also be used as an effective tool for demonstrating the quality of the resultant protein models when presenting the results for publication.

Acknowledgements

This work has been funded by the BIOTECH program of DGXII of the Commission of the European Union, under contract no. BIO2CT-920524. We would like to thank Jurgen Doreleijers, Ben Davis, Dave Love, Markus Bluemel, Werner Klaus, Alexandre Bonvin, Kris Boulez, Bryan Finn and Sunil Patel for helpful comments and suggestions, and Hans Vis for providing the coordinates and restraint files for the HU protein.

References

Adobe Systems Inc. (1985) *PostScript Language Reference Manual*, Addison-Wesley, Reading, MA, U.S.A.

Allen, F.H. and Johnson, O. (1991) *Acta Crystallogr.*, **B47**, 62–67.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Brünger, A.T. (1992) *X-PLOR v. 3.1. A system for X-ray crystallography and NMR*, Yale University Press, New Haven, CT, U.S.A.

Chazin, W.J. (1992) *Curr. Opin. Biotechnol.*, **3**, 326–332.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82–102.

Gonzalez, C., Rullmann, J.A.C., Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1991) *J. Magn. Reson.*, **91**, 659–664.

Güntert, P., Braun, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 517–530.

Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Biol.*, **46**, 673–698.

Havel, T.F. (1991) *Prog. Biophys. Mol. Biol.*, **56**, 43–78.

Hoch, J.C. (1991) In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Ed., Hoch, J.C.), Plenum Press, New York, NY, U.S.A., pp. 253–267.

Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.

MacArthur, M.W. and Thornton, J.M. (1993) *Proteins*, **17**, 232–251.

MacArthur, M.W., Laskowski, R.A. and Thornton, J.M. (1994) *Curr. Opin. Struct. Biol.*, **4**, 731–737.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins*, **12**, 345–364.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) *J. Mol. Biol.*, **7**, 95–99.

Sutcliffe, M.J. (1993) In *NMR of Macromolecules. A Practical Approach* (Ed., Roberts, G.C.K.), IRL Press, Oxford, U.K., pp. 359–392.

Thomas, P.D., Basus, V.J. and James, T.L. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 1237–1241.

Vis, H., Mariani, M., Vorgias, C.E., Wilson, K.S., Kaptein, R. and Boelens, R. (1995) *J. Mol. Biol.*, **254**, 692–703.

Zhao, D. and Jardetzky, O. (1994) *J. Mol. Biol.*, **239**, 601–607.