

## Toward an NMR *R* Factor

C. GONZALEZ,\* J. A. C. RULLMANN, A. M. J. J. BONVIN,  
R. BOELEN, AND R. KAPTEIN†

*Department of Chemistry, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands*

Received October 2, 1990

Several indicators have been used in order to judge the quality of NMR structures. In most cases the NMR data are translated into distance constraints. The number and size of the constraint violations can then be used as a measure for the agreement between model structure and NMR data.

A more direct comparison between experimental and calculated data leads to the definition of a residual error function (1-7) analogous to the *R* factor used in X-ray structure determination. The conventional *R* factor is defined as the normalized mean deviation between structure factors derived from the model,  $F_{\text{cal}}$ , and from experimental data,  $F_{\text{exp}}$ ,

$$R(F) = \frac{\sum ||F_{\text{cal}}| - |F_{\text{exp}}||}{\sum |F_{\text{exp}}|}. \quad [1]$$

Also very common is the Booth *R* factor, defined as the normalized standard deviation

$$R(F) = \left( \frac{\sum (|F_{\text{cal}}| - |F_{\text{exp}}|)^2}{\sum |F_{\text{exp}}|^2} \right)^{1/2}. \quad [2]$$

Definition [2] is more closely related to the quantity  $(|F_{\text{cal}}| - |F_{\text{exp}}|)^2$  that is usually minimized in least-square crystallographic refinements.

To derive the three-dimensional structure of a macromolecule in solution from NMR data, the most useful experimental information is obtained from NOE signal intensities. Therefore, an NMR *R* factor may be defined as a measure of the agreement between intensities of cross peaks observed in NOESY spectra of the macromolecule.

NOE calculations can be based on relaxation matrix theory (7-9). In this approach, the Bloch equations are solved for the complete spin system obtaining a theoretical NOE intensity for all proton pairs at different mixing times. The normalized intensities can be written as

$$A(\tau_m) = \exp(-\mathbf{R}\tau_m), \quad [3]$$

where  $A_{ij}(\tau_m)$  is the NOE intensity of the spin pair (*ij*),  $\tau_m$  is the mixing time, and  $\mathbf{R}$  is the relaxation matrix.

\* Permanent address: Instituto de Estructura de la Materia, C. S. I. C., Madrid.

† To whom correspondence should be addressed.

Two general definitions of an NMR  $R$  factor, analogous to the X-ray definitions [1] and [2], are

$$R = \frac{\sum_{i,j,m} W_{ij}(\tau_m) |A_{ij}^{\text{calc}}(\tau_m) - A_{ij}^{\text{exp}}(\tau_m)|}{\sum_{i,j,m} W_{ij}(\tau_m) A_{ij}^{\text{exp}}(\tau_m)} \quad [4]$$

and

$$R = \left( \frac{\sum_{i,j,m} W_{ij}(\tau_m) |A_{ij}^{\text{calc}}(\tau_m) - A_{ij}^{\text{exp}}(\tau_m)|^2}{\sum_{i,j,m} W_{ij}(\tau_m) (A_{ij}^{\text{exp}}(\tau_m))^2} \right)^{1/2}. \quad [5]$$

Weight factors  $W_{ij}(\tau_m)$  should be included to take account of measurement errors, in particular noise levels, as well as inaccuracies of the theoretical procedures.

Estimation of the errors in the NOE signal intensities is a tricky problem. Many effects can contribute to them but, in general, it is clear that the errors will be higher at short mixing times, when the signal-to-noise ratio is lower. One possibility, therefore, is to set the weight factor equal to the mixing time. This and other specifications will be discussed below.

We have tested different NMR  $R$ -factor definitions with the structure of crambin. Crambin is a small protein that has been studied extensively by NMR (4, 10–12). Its structure in solution has been determined by a new iterative procedure that uses the full relaxation matrix in order to calculate precise distance restrictions (8, 9). In this procedure, called IRMA, theoretical NOE values are calculated with Eq. [3]. These values are replaced by the experimental ones when the latter are available. The new NOE matrix is transformed back to a corrected relaxation matrix, from which new distances are calculated. Structure calculations are performed using these new distance restrictions. The whole process is repeated until convergence is obtained. In all the calculations isotropic tumbling with correlation time  $\tau_c$  was assumed. A complete description of this method can be found elsewhere (8, 9, 13).

In the calculation of the solution structure of crambin (4), three cycles of IRMA have been performed, starting from a fully extended chain. In the first cycle eight distance-geometry structures were calculated. The best one was simulated for 20 ps with restrained molecular dynamics (RMD) followed by restrained energy minimization (REM), and then used as starting structure for the second cycle of IRMA. Using the new bounds, 25 ps of RMD was performed at 600 K. Five structures from the trajectory, one after each 5 ps, were then simulated for 35 ps at 300 K, averaged over the last 20 ps, and subjected to REM. The resulting structures are very similar, both in energy and in atomic positions (4). The one with the lowest restraint energy was used as starting point for the third cycle.

Now stereospecific assignments were included for a number of prochiral groups, both in IRMA calculations and in the RMD simulations. Assignments were based on experimental data, i.e.,  $J$  couplings and relative NOE strengths, as well as on a direct comparison between measured and calculated NOE values (12). The third cycle involved 17 ps of RMD employing a stepwise increase of the distance restraint force constant, followed by 20 ps which was used for averaging. The average structure was again subjected to REM. A more detailed account of this work will be given elsewhere.

The results of these IRMA cycles have been used to test the concept of an NMR  $R$  factor. Table 1 shows  $R$  factors for the resulting structures, as well as for a linear chain (the initial starting structure) and for the X-ray structure of 1.5 Å resolution obtained from the Brookhaven Protein Data Bank (14). Since the original X-ray structure contains a sequence error at position 25 (10), Ile-25 was replaced by Leu. Subsequently a short unrestrained energy minimization was carried out in order to remove bad contacts (4).

Four different  $R$ -factor definitions are employed in Table 1. The first two correspond to the definitions [4] and [5], setting  $W_{ij}(\tau_m) = 1$ , i.e.,

$$R_1 = \frac{\sum_{i,j,m} |A_{ij}^{\text{calc}}(\tau_m) - A_{ij}^{\text{exp}}(\tau_m)|}{\sum_{i,j,m} A_{ij}^{\text{exp}}(\tau_m)}, \quad [6]$$

$$R_2 = \left( \frac{\sum_{i,j,m} |A_{ij}^{\text{calc}}(\tau_m) - A_{ij}^{\text{exp}}(\tau_m)|^2}{\sum_{i,j,m} (A_{ij}^{\text{exp}}(\tau_m))^2} \right)^{1/2}. \quad [7]$$

Both the  $R_1$  definition (3, 6) and the  $R_2$  definition (1, 7) have been used before. The quantity  $R_r$ , used by us in earlier work (4), is similar to  $R_1$  but has  $W_{ij}(\tau_m) = \tau_m$ :

$$R_r = \frac{\sum_m \tau_m \sum_{i,j} |A_{ij}^{\text{calc}}(\tau_m) - A_{ij}^{\text{exp}}(\tau_m)|}{\sum_m \tau_m \sum_{i,j} A_{ij}^{\text{exp}}(\tau_m)}. \quad [8]$$

Alternatively, one may use the distance-like quantity  $A_{ij}^{-1/6}$  to define an  $R$  factor that is more closely related to the usual distance restraint energy than  $R_1$  or  $R_2$ . We call this quantity  $R_r$  and define it as

$$R_r = \left( \frac{\sum_{i,j,m} |(A_{ij}^{\text{calc}}(\tau_m))^{-1/6} - (A_{ij}^{\text{exp}}(\tau_m))^{-1/6}|^2}{\sum_{i,j,m} (A_{ij}^{\text{exp}}(\tau_m))^{-1/3}} \right)^{1/2}. \quad [9]$$

The results in Table 1 show similar trends for all definitions; i.e., the  $R$  factor decreases as the structure refinement progresses. More specifically, there is no essential

TABLE 1  
Comparison of  $R$ -Factor Definitions for Several Crambin Structures<sup>a</sup>

$R$ :	Interresidue				Inter- and intraresidue		
	$R_1$	$R_2$	$R_r$	$R_r$	$R_1$	$R_r$	$R_r$
Linear chain	0.96	1.03	0.95	11.65	0.53	0.52	7.98
After first IRMA cycle <sup>b</sup>	0.60	0.73	0.56	0.24	0.40	0.38	0.21
After second IRMA cycle	0.58	0.63	0.53	0.24	0.44	0.41	0.21
After third IRMA cycle	0.57		0.52	0.22	0.41	0.38	0.20
X-ray structure	0.51	0.56	0.47	0.24	0.42	0.39	0.20
Number of NOEs			264			646	

<sup>a</sup>  $R$ -factor definitions according to Eqs. [6]–[9]. The sums over  $i, j$  are taken over interresidue contacts only, or over all (i.e., both inter- and intraresidue) contacts.

<sup>b</sup> A cycle is defined as the derivation of distance constraints according to the IRMA procedure (8, 9), followed by structure refinement. See text for details on the refinement procedures.

difference between the quadratic ( $R_2$ ) and the linear ( $R_1$ ,  $R_r$ ) forms. The inclusion of mixing times as weight factors does yield lower  $R$  factors, as expected, but does not change the overall trend.

A possible disadvantage of the definitions [6] to [8] is that strong contacts, i.e., short distances, are overemphasized. This might be remedied by defining  $R$  as the average of individual ratios  $\Delta A_{ij}/A_{ij}$  as done by some authors (2, 3), but then the result is very sensitive to experimental noise; only values obtained at long mixing times can be used.

The  $R_r$  quantity defined in [9] also gives similar weight to short and long distances. The  $R_r$  value is indeed much higher than all of the other definitions for the linear chain, where many long-range constraints are violated. For the other structures, the  $R_r$  factors are rather similar, corresponding to the similarity of distance restraint energies, although a slow decrease can be observed.

From Table 1 it is also clear that the sensitivity of the  $R$  factor increases when only interresidue contacts are included in the summations, giving values that are spread over a wider range. Independently of the model, intraresidue pairs of protons cannot be very far away from each other. Therefore, even for a completely extended chain, theoretical intensities cannot be very different from experimental ones. The high number of intraresidue signals observed hides the conformational effects in the global  $R$  factor.

Table 2 shows partial  $R$  factors, calculated for interresidue contacts with the  $R_r$  definition. Backbone-backbone contacts have an  $R$  factor that is consistently lower than contacts involving one or two side chain groups. This shows that the backbone of the protein is much better defined than the side chains, as is expected from the RMS deviations of the structures determined by NMR. RMS deviations are always higher when side chains are included.

It is also clear that the different types of contacts were not all optimized simultaneously, although the total interresidue  $R$  factor dropped monotonously. One important factor is the absence of stereospecific assignments, leading to the introduction of pseudoatoms and concomitant relaxation of distance restraints by at least 1 Å. This does not affect drastically the structure of the backbone, but it leads to an important loss of

TABLE 2  
Partial  $R$  Factors on Interresidue Contacts for Several Crambin Structures<sup>a</sup>

	All interresidue	Backbone- backbone	Side chain- side chain	Backbone- side chain
Linear chain	0.95	0.92	0.94	0.97
After first IRMA cycle	0.56	0.39	0.66	0.63
After second IRMA cycle	0.53	0.38	0.69	0.56
After third IRMA cycle	0.52	0.42	0.79	0.52
X-ray structure	0.47	0.41	0.56	0.49
Number of NOEs	264	66	66	132

<sup>a</sup>  $R$  factors are computed using the  $R_r$  definition (Eq. [8]) and summing only over the indicated subsets of interresidue contacts.

precision for the side chains. Another source of inaccuracy is the neglect of local motions in the theoretical calculations. Again this will in general influence side chains more than backbone atoms. In agreement with this, the partial  $R$  factor for NOEs involving only backbone atoms is always lower than that for contacts involving side chain atoms. The side chain–side chain  $R$  factor varies strongly between different optimized structures (4).

We conclude that the X-ray structure exhibits a low  $R$  factor because of a correct positioning of side chains (except for a few near the surface). Introduction of a number of stereospecific assignments in the third structure refinement cycle improves the overall agreement between calculated and experimental NOE values, although individual contributions, especially from contacts between side chain atoms, still vary considerably. Measuring over all contacts (all definitions) or over interresidue contacts ( $R_i$  definition) yields an  $R$  factor below that of the crystal structure. Work on including mobility corrections using parameters derived from MD simulations (15) is in progress. Preliminary results show that the  $R$  factor drops even further.

We conclude that  $R$  factors are useful indicators of the agreement between calculated structures and experimental NOE data. They compare directly theoretical predictions with experimental quantities without interpretation of the latter in a theory-dependent manner.

Definitions [6] and [7] with the weighting factors equal to 1 have the virtue of simplicity and also correspond to the  $R$  factors as used in X-ray crystallography. However, these definitions measure the absolute error, which is strongly influenced by short-range contacts. The actual structural information and the accuracy of the procedures may be reflected better by definition [9], or by partial  $R$  factors including only a subset of the observed NOEs.

A definition in terms of relative errors  $\Delta A_{ij}/A_{ij}$  (arising from the general definition [4] by setting  $W_{ij} = 1/A_{ij}$ ) might seem to be appropriate, but is in our experience strongly sensitive to experimental noise. A more refined expression for the weights  $W_{ij}$  is needed, based upon an estimate of error contributions (1). Such a definition is currently being developed in our group.

Finally, the variation of partial  $R$  factors in subsequent phases of the structure refinement supports attempts to minimize the  $R$  factor directly in an RMD calculation (6, 16). This offers the possibility of integrating IRMA and RMD directly, circumventing the need for the back-transformation step.

#### ACKNOWLEDGMENTS

J.A.C.R. is the recipient of a fellowship from the Netherlands Organization for Scientific Research (NWO). We thank Joost van Opheusden for discussions of error definitions. This work was supported by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO).

#### REFERENCES

1. J.-F. LEFÈVRE, A. N. LANE, AND O. JARDETZKY, *Biochemistry* **26**, 5076 (1987).
2. G. GUPTA, M. H. SARMA, AND R. H. SARMA, *Biochemistry* **27**, 7909 (1988).
3. E. P. NIKONOWICZ, R. P. MEADOWS, AND D. G. GORENSTEIN, *Biochemistry* **29**, 4193 (1990).
4. J. A. C. RULLMANN, R. M. J. N. LAMERICH, C. GONZALEZ, T. M. G. KONING, R. BOELEN, AND R.

- KAPTEIN, in "Modelling of Molecular Structures and Properties" (J.-L. Rivail, Ed.), Studies in Physical and Theoretical Chemistry, Vol. 71, p. 703, Elsevier, Amsterdam, 1990.
5. A. N. LANE, *Biochim. Biophys. Acta* **1049**, 189 (1990).
  6. J. D. BALEJA, J. MOULT, AND B. D. SYKES, *J. Magn. Reson.* **87**, 375 (1990).
  7. B. A. BORGAS, M. GOCHIN, D. J. KERWOOD, AND T. L. JAMES, *Prog. NMR Spectrosc.* **22**, 83 (1990).
  8. R. BOELEN, T. M. G. KONING, AND R. KAPTEIN, *J. Mol. Struct.* **173**, 299 (1988).
  9. R. BOELEN, T. M. G. KONING, G. A. VAN DER MAREL, J. H. VAN BOOM, AND R. KAPTEIN, *J. Magn. Reson.* **82**, 290 (1989).
  10. J. A. W. H. VERMEULEN, R. M. J. N. LAMERICH, L. J. BERLINER, A. DE MARCO, M. LLINÁS, R. BOELEN, J. ALLEMAN, AND R. KAPTEIN, *FEBS Lett.* **219**, 426 (1987).
  11. R. M. J. N. LAMERICH, L. J. BERLINER, R. BOELEN, A. DE MARCO, M. LLINÁS, AND R. KAPTEIN, *Eur. J. Biochem.* **171**, 307 (1988).
  12. R. M. J. N. LAMERICH, Ph.D. thesis, University of Utrecht, The Netherlands, 1989.
  13. J. A. C. RULLMANN, R. BOELEN, R. M. J. N. LAMERICH, G. W. VUISTER, AND R. KAPTEIN, in "Modelling of Molecular Structures and Properties" (J.-L. Rivail, Ed.), Studies in Physical and Theoretical Chemistry, Vol. 71, p. 661, Elsevier, Amsterdam, 1990.
  14. W. A. HENDRICKSON AND M. M. TEETER, *Nature (London)* **290**, 107 (1981).
  15. M. M. G. KONING, Ph.D. thesis, University of Utrecht, The Netherlands, 1990.
  16. P. YIP AND D. A. CASE, *J. Magn. Reson.* **83**, 643 (1989).