Journal of Biomolecular NMR, 7 (1996) 72–76
ESCOM

J-Bio NMR 331

Do NOE distances contain enough information to assess the relative populations of multi-conformer structures?

Alexandre M.J.J. Bonvin and Axel T. Brünger*

The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, P.O. Box 208114, New Haven, CT 06520, U.S.A.

Received 7 December 1995
Accepted 8 January 1996

Keywords: Occupancy refinement; Ensemble-averaged NOE restraints; Multi-conformer refinement; Complete cross-validation; Conformational variability; Solution structure

Summary

The feasibility of determining the relative populations of multi-conformer structures from NOE-derived distances alone is assessed. Without cross-validation of the NOE restraints, any population ratio can be refined to a similar quality of the fit. Complete cross-validation provides a less biased measure of fit and allows the estimation of the correct population ratio when used in conjunction with very tight distance restraints. With the qualitative distance restraints most commonly used in NMR structure determination, cross-validation is unsuccessful in providing the correct answer. Other experimental sources are therefore needed to determine relative populations of multi-conformer structures.

The important biological role of macromolecular motions has been highlighted by the increasing number of both crystallographic and NMR reports on structural changes linked to the functionality of biomacromolecules. A few recent examples include the partial folding and unfolding of Bam H1 endonuclease upon DNA binding (Newman et al., 1995), the activation mechanism of the cyclin dependent kinase2 (CDK2) involving conformational changes upon binding of cyclinA (Jeffrey et al., 1995) and the opening of flaps in HIV-1 protease, allowing access to the active site as measured by NMR relaxation experiments (Nicholson et al., 1995). Recognizing and identifying protein motions can be an important step toward a better understanding of their functional role (Gerstein et al., 1994). Often such motions are identified from different forms of a protein, e.g. free and bound conformations. However, information about motion can also be obtained directly from a single structure determined by NMR or X-ray crystallography, since the structure represents an ensemble and/or time average (Torda et al., 1989,1993; Gros et al., 1990; Kuriyan et al., 1991; Scheek et al., 1991; Burling and Brünger, 1994; Bonvin and Brünger, 1995). Methods need to be developed that allow the identification of conformational variability within a single crystal or NMR structure. In a few cases it has been possible to identify local variability in structures, e.g. in the solution structure of interleukin-8 (Bonvin and Brünger, 1995) or in the crystal structure of the mannose-binding protein A (MBP) (Burling et al., 1996).

In a previous paper we addressed the problem of identifying conformational variability in NMR solution structures, using ensemble-averaged NOE restraints in combination with complete cross-validation to avoid overfitting of the experimental NMR data (Bonvin and Brünger, 1995). Here we investigate whether, once multiple conformers have been identified, it is possible to assess their relative populations from the experimental NOE data alone. Our approach to determine the relative populations is similar to the work by Kim and Prestegard (1989,1990) for J-coupling restraints, who derived the populations of each conformer from the best fit to the experimental data, and in contrast to the work of Fennen et al. (1995), who used Boltzmann weights based on the potential energy of the system in the course of a molecular dynamics simulation. Complete cross-validation of the NOE restraints can be carried out to avoid overfitting (Brünger et al., 1993; Bonvin and Brünger, 1995). Using a synthetic test case, we investigate whether NOE distances contain enough information to assess the relative populations of multiple-conformation structures.

*To whom correspondence should be addressed.

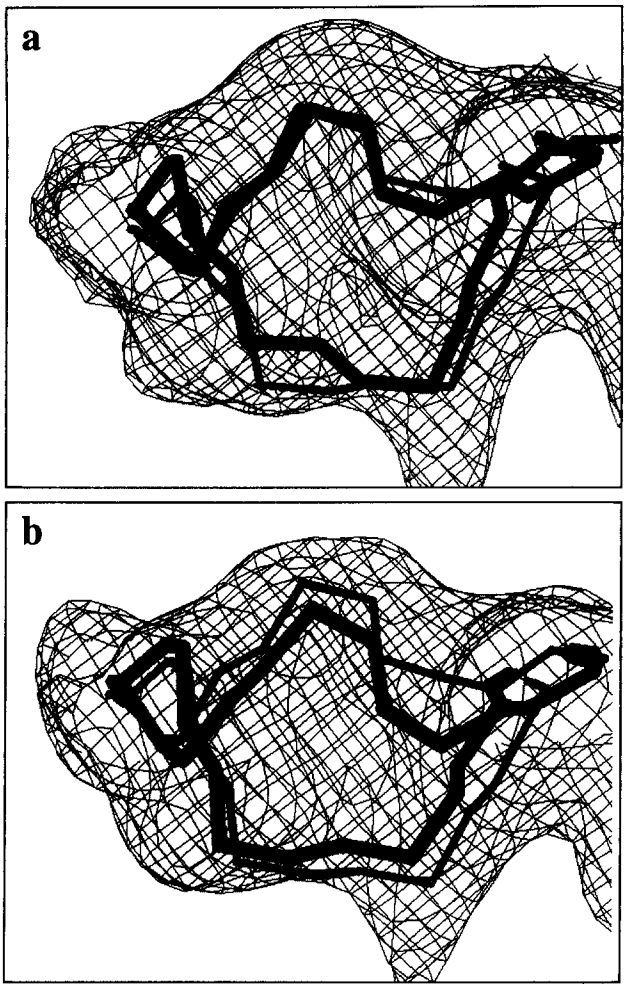

Fig. 1. View of the multiple-conformation loop region of the Ambt V structures refined by means of a probability map (thick dark gray lines) with equal populations for the 25/75% NOE data with (a) tight (± 10% error bounds) and (b) qualitative (2.7, 3.5 and 5.0 Å upper bounds) distance restraints. As a comparison, the reference structure is indicated in light gray lines. The corresponding probability maps are plotted at 1.25 standard deviations above the mean.

Synthetic NOE data were generated from the structure of the 40-residue protein ragweed allergen Ambt V (Metzler et al., 1992; G. Warren, Yale University, New Haven, CT, personal communication) with two alternate conformations for the loop defined by residues 21–27 (Bonvin and Brünger, 1995). From the two conformers, synthetic sets of 1031 NOE distances were calculated using r^{-6} averaging with a 5 Å cutoff, and with 25/75% and 75/25% population ratios, respectively. For each population ratio two sets of NOE distance restraints were generated: a tight restraint set by adding ± 10% error bounds to the average distances and a qualitative restraint set by assigning the restraints to qualitative distance ranges of 1.8–2.7 Å, 1.8–3.5 Å and 1.8–5.0 Å. These two restraint sets represent ideal cases in which all possible proton–proton distances within 5 Å were used, resulting in more than 250 restraints for the multi-conformer loop region. We chose to use such ideal data in order to assess the feasibility of determining the relative populations of multi-conformer structures from NOE data only in the best case possible. We first followed the protocol described in our previous paper (Bonvin and Brünger, 1995), i.e., each conformer in the ensemble contributing equally to the average distances, to make sure that we were still able to identify the correct conformers.

Complete cross-validation of the rms deviations from the NOE-derived distances was performed in order to find the minimum number of conformers that best fit the NOE data. For complete cross-validation, the NOE-derived distances were partitioned into 10 random subsets, each of which was, in turn, omitted during refinement. A minimum of 10 refinement runs were therefore performed for each particular case (different data or different number of conformers). Slow-cooling simulated annealing refinement (Nilges et al., 1988) with ensemble-averaged NOE restraints was repeated for increasing numbers of conformers (nconf). The rms deviations from the NOE-derived distances and the number of violations exceeding 0.2 Å were monitored for the (omitted) test sets and averaged. The twin-conformer (nconf = 2) model giving the best cross-validated measure of the fit was chosen. An average representation of the ensemble was then generated using a probability map refinement protocol (DeLano and Brünger, 1994; Bonvin and Brünger, 1995). Following this protocol, we were able to correctly reproduce the conformational variability in the loop region, even with the synthetic NOE data calculated from uneven populations. To minimize the number of parameters in all subsequent calculations, we treated only the loop region as being multi-conformer (it was identified from the rms deviations per residue for the twin-conformer structures using the average backbone rms deviation as a threshold (0.9 Å and 1.2 Å for the 25/75% and 75/25% NOE data, respectively)). Reducing the model in this way reduces potential problems in occupancy refinement. The force field typically used in X-PLOR (Brünger, 1993) for NMR refinement does not include any attractive nonbonded energy term other than the NOE term itself and therefore nothing prevents structures with very low occupancies from unfolding during high-temperature simulations. Using a single conformer outside the loop region reduces this problem. A detailed view of the loop regions of the structures refined by using a probability map, obtained both with the tight and the qualitative NOE restraints for the 25/75% NOE data, is shown in Fig. 1. These structures provided the starting point for occupancy refinement in order to determine the relative population of each conformer in the ensemble.

Occupancy refinement (constrained such that the sum of the occupancies equalled one) was performed using a grid search to find the relative population ratio that best fitted the NOE data. Occupancies were assigned separately to each conformer and a slow-cooling simulated an-

nealing protocol (5 ps at 1000 K, slow cooling to 1 K with a cooling rate of 50 K/0.05 ps and a final restrained energy minimization) was applied to the ensemble of structures (Bonvin and Brünger, 1995). Complete cross-validation was performed for each population ratio and was repeated several times with various random seeds to obtain error estimates. This is a simple procedure for a twin-conformer model, but it quickly becomes computationally intensive with increasing number of conformers. The averaged NOE distances were calculated from the ensemble of conformers as:

$$r_{ens} = \left[\sum_{k=1}^{nconf} q_k \, r^{-6} \right]^{-1/6} \quad \text{with} \quad \sum_{k=1}^{nconf} q_k = 1 \qquad (1)$$

where q_k gives the population (occupancy) of the conformer k. This type of averaging is appropriate, since magnetization transfer within a conformer is independent from the other conformers. To stress the importance of cross-validation, Fig. 2 presents the rms deviations from the tight NOE restraints used for refinement as a function of the population ratio at various stages. Clearly, the starting structures that were obtained using equal populations are biased toward a 50/50% ratio for both NOE data sets: varying the population ratio does not result in any improvement of the fit (Fig. 2a). Restrained energy minimization alone is unable to remove the bias (Fig. 2b). Only the slow-cooling simulated annealing protocol is able to remove the bias from the starting structures, but results in a new bias toward the population ratio used for refinement. It is indeed possible to refine almost any population ratio between 0.1 and 0.9 to similar low rms deviations! A less biased measure of the fit is needed to determine the best relative population ratio. Complete cross-validation can be used for this purpose. The cross-validated rms deviations from the NOE restraints as a function of the population ratio are shown in Fig. 3 for all four NOE data sets (25/75% and 75/25% with tight and qualitative distance restraints, respectively). The error estimates, which are on the order of 1 to 2%, are not indicated in the figures, since they would not show up at the plotting scale used. With tight NOE restraints, minima are found close to the target values. Although these minima are shallow, the forms of the curves clearly indi-

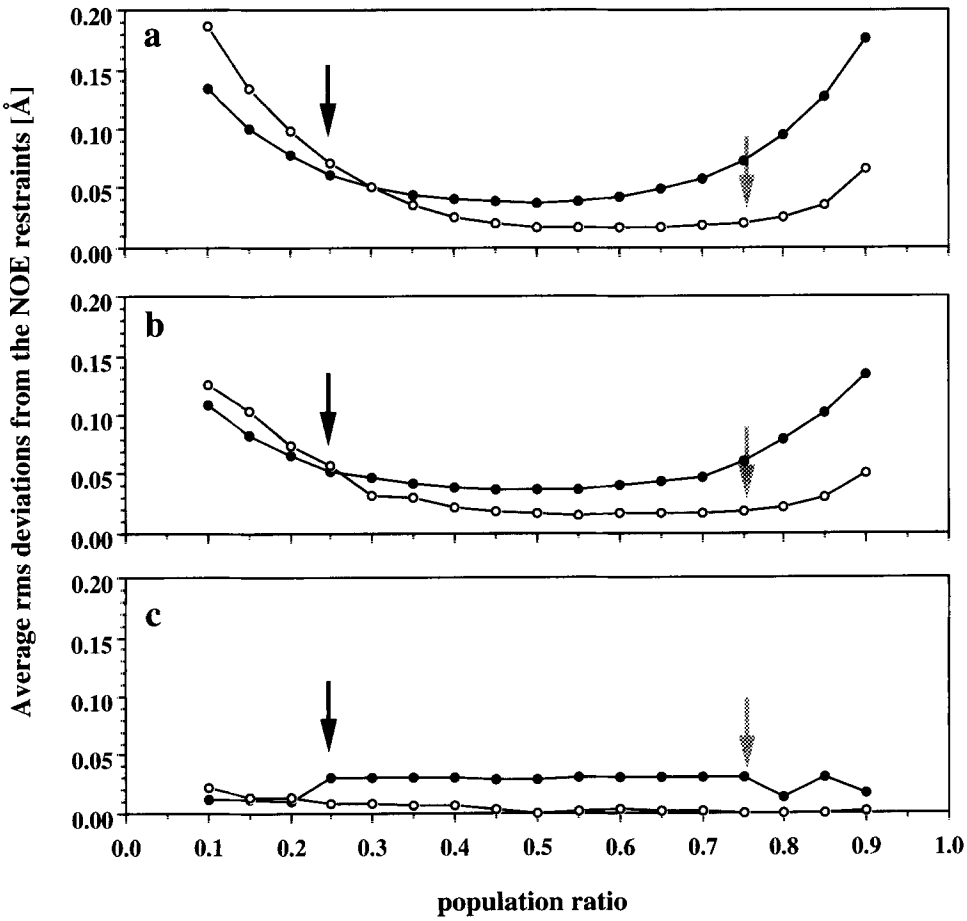

Fig. 2. Average rms deviations from the NOE distance restraints (tight restraints) as a function of the population ratio at various refinement stages. (a) Starting structures; (b) after restrained energy minimization; and (c) after slow-cooling simulated annealing. Deviations were calculated for the 25/75% (filled circles) and 75/25% (open circles) NOE data. The black and gray arrows indicate the correct population ratios for the 25/75% and 75/25% NOE data, respectively.

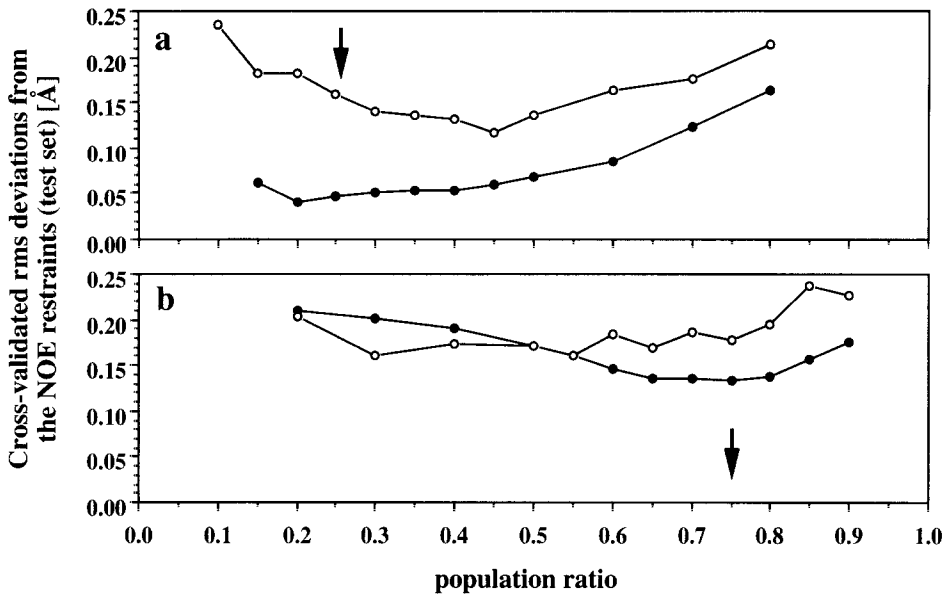

Fig. 3. Cross-validated rms deviations from the NOE distance restraints as a function of the population ratio for the 25/75% (a) and 75/25% (b) NOE data with tight (±10% error bounds, filled circles) and qualitative (2.7, 3.5 and 5.0 Å upper bounds, open circles) NOE restraints. The arrows indicate the correct population ratio.

cate which conformer is the most populated. With qualitative NOE distance restraints, however, incorrect minima are observed at a 45/55% ratio for the 25/75% data and a 30/70% ratio for the 75/25% data. These results indicate that qualitative NOE distance restraints do not contain enough information to assess the relative populations of multi-conformer structures, even when using complete cross-validation.

Our results show that only tight NOE restraints may contain enough information to assess the relative population ratio of multi-conformer structures. Such restraints, in principle, can be obtained from relaxation matrix calculations (Keepers and James, 1984; Olejniczak et al., 1986; Boelens et al., 1988; Borgias et al., 1990; Koehl and Lefèvre, 1990; Post et al., 1990; Madrid et al., 1991; Van de Ven et al., 1991; Edmonson, 1992; Leeflang and Kroon-Batenburg, 1992; Bonvin et al., 1993; Liu et al., 1995). However, even with such tight restraints, the minima found with complete cross-validation are shallow and must be interpreted with caution. Qualitative distance restraints, which are most commonly used for NMR structure determination, cannot be used for assessing population ratios since, even when complete cross-validation is used, they can result in wrong minima. In this study, even with ideal data including all possible NOEs, the exact minima could not be reproduced. With real experimental NMR data, a smaller number of NOEs will typically be available, making the situation even worse. Other experimental sources like J-coupling values and/or chemical shifts could provide the required additional information to unambiguously determine populations of multi-conformer structures. We should finally note that related work has been published previously for nucleosides and small organic molecules (Schirmer et al., 1972; Kruse et al., 1985), and more recently for a dipeptide (Landis et al., 1995). In all these cases, accurate determination of populations appears to be a difficult problem. These findings for small molecules only strengthen our conclusions for larger biomolecules, for which the complexity of the problem greatly increases.

Acknowledgements

The authors thank Paul Adams and Greg Warren for careful reading of this manuscript. A.M.J.J.B. thanks the Swiss National Foundation for Scientific Research for financial support. This work was funded in part by the National Science Foundation (A.T.B., BIR 9021975).

References

Boelens, R., Koning, T.M.G. and Kaptein, R. (1988) *J. Mol. Struct.*, **173**, 299–311.

Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1993) In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, Vol. 2 (Eds, Van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J.), ESCOM, Leiden, pp. 407–440.

Bonvin, A.M.J.J. and Brünger, A.T. (1995) *J. Mol. Biol.*, **250**, 80–93.

Borgias, B.A., Cochin, M., Kerwood, D.J. and James, T.L. (1990) *Prog. NMR Spectrosc.*, **22**, 83–100.

Brünger, A.T. (1993) X-PLOR v. 3.1: A system for X-ray crystallography and NMR, Yale University Press, New Haven, CT.

Brünger, A.T., Clore, M.G., Gronenborn, A.M., Saffrich, R. and Nilges, M. (1993) *Science*, **261**, 328–331.

Burling, F.T. and Brünger, A.T. (1994) *Isr. J. Chem.*, **34**, 165–175.

Burling, F.T., Weis, W.I., Flaherty, K.M. and Brünger, A.T. (1996) *Science*, **271**, 72–77.

DeLano, W.L. and Brünger, A.T. (1994) *Proteins*, **20**, 105–123.

Edmonson, S. (1992) *J. Magn. Reson.*, **98**, 283–298.

Fennen, J., Torda, A.E. and Van Gunsteren, W.F. (1995) *J. Biomol. NMR*, **6**, 163–170.

Gerstein, M., Lesk, A.M. and Chothia, C. (1994) *Biochemistry*, **33**, 6739–6749.

Gros, P., Van Gunsteren, W.F. and Hol, W.G.J. (1990) *Science*, **249**, 1149–1152.

Jeffrey, P.D., Russo, A.A., Polyak, K., Gibbs, E., Hurwitz, J., Massagé, J. and Pavletich, N.P. (1995) *Nature*, **376**, 313–320.

Keepers, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.

Kim, Y. and Prestegard, J.H. (1989) *Biochemistry*, **28**, 8792–8797.

Kim, Y. and Prestegard, J.H. (1990) *Proteins*, **8**, 377–385.

Koehl, P. and Lefèvre, J.-P. (1990) *J. Magn. Reson.*, **86**, 565–583.

Kruse, L.I., DeBrosse, C.W. and Kruse, C.H. (1985) *J. Am. Chem. Soc.*, **107**, 5435–5442.

Kuriyan, J., Ösapay, K., Burley, S.K., Brünger, A.T., Hendrickson, W.A. and Karplus, M. (1991) *Protein Struct. Funct. Genet.*, **10**, 340–358.

Landis, C.R., Luck, L.L. and Wright, J.M. (1995) *J. Magn. Reson. Ser. B*, **109**, 44–59.

Leeflang, B.R. and Kroon-Batenburg, L.M.J. (1992) *J. Biomol. NMR*, **2**, 495–518.

Liu, H., Banville, D.L., Basus, V.J. and James, T.L. (1995) *J. Magn. Reson. Ser. B*, **107**, 51–59.

Madrid, M., Llinás, E. and Llinás, M. (1991) *J. Magn. Reson.*, **93**, 329–346.

Metzler, W.J., Valentine, K., Roebber, M., Friedrichs, M.S., March, D.G. and Mueller, L. (1992) *Biochemistry*, **31**, 5117–5127.

Newman, M., Strzelecka, T., Dorner, L.F., Schildkraut, I. and Aggarwal, A.K. (1995) *Science*, **269**, 656–663.

Nicholson, L.K., Yamazaki, T., Torchia, D.A., Grzesiek, S., Bax, A., Stahl, S.J., Kaufman, J.D., Wingfield, P.T., Lam, P.Y.S., Jadhav, P.K., Hodge, N., Domaille, P.J. and Chang, C.-H. (1995) *Nature Struct. Biol.*, **2**, 274–280.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.*, **229**, 317–324.

Olejniczak, E.T., Gampe Jr., R.T. and Fesik, S.W. (1986) *J. Magn. Reson.*, **67**, 28–41.

Post, C.B., Meadows, R.P. and Gorenstein, D.G. (1990) *J. Am. Chem. Soc.*, **112**, 6796–6803.

Scheek, R.M., Torda, A.E., Kemmink, J. and Van Gunsteren, W.F. (1991) In *Computational Aspects of the Study of Biological Macromolecules by NMR* (Eds, Hoch, J.C., Poulsen, F.M. and Redfield, C.), Plenum Press, New York, NY, pp. 209–217.

Schirmer, R.E., Davis, J.P., Noggle, J.H. and Hart, P.A. (1972) *J. Am. Chem. Soc.*, **94**, 2561–2572.

Torda, A.E., Scheek, R.M. and Van Gunsteren, W.F. (1989) *Chem. Phys. Lett.*, **157**, 289–294.

Torda, A.E., Brunne, R.M., Huber, T., Kessler, H. and Van Gunsteren, W.F. (1993) *J. Biomol. NMR*, **3**, 55–66.

Van de Ven, F.J.M., Blommers, M.J.J., Schouten, R.E. and Hilbers, C.W. (1991) *J. Magn. Reson.*, **94**, 140–151.