Journal of Biomolecular NMR, 3 (1993) 355–360
ESCOM

J-Bio NMR 118

SESAME: A least-squares approach to the evaluation of protein structures computed from NMR data

Ju-xing Yang and T.F. Havel*

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, U.S.A.

Received 7 January 1993
Accepted 16 February 1993

Keywords: Protein structure; Nuclear magnetic resonance; Nuclear Overhauser effect; Structure refinement

SUMMARY

A method is proposed for defining a probability distribution on an ensemble of protein conformations from a 2D NOE spectrum, while at the same time back-calculating the experimental spectrum from the ensemble. This enables one to assess the relative quality and significance of the conformations, and to test the consistency of the ensemble as a whole with the experimental spectrum. The method eliminates the need to integrate the cross-peak intensities and is surprisingly insensitive to random noise in the spectrum. In this communication, these advantages are demonstrated by applying the method to simulated data, for which the correct result is already known.

The determination of biological conformation from NMR data is generally accomplished by distance geometry algorithms (see Wagner et al., 1991 and references therein). As a rule, these computations yield an ensemble of conformations that are consistent with the data. With sufficiently complete and precise data, the distribution of conformations in these ensembles appears to be correlated with the actual distribution of conformations in solution (Hyberts et al., 1992). Unfortunately, this correlation is a weak one, and hence the conformations are generally regarded as being equally likely fits to the data. Most attempts at refinement have attempted to identify a single conformation as being 'the' solution conformation, even though some conformational heterogeneity is inevitable in solution.

Several attempts to fit multiple conformations to NMR data have been made (Kessler et al., 1988; Brüschweiler et al., 1991; Kim et al., 1991). Most recently, Landis and Allured (1991) have taken a set of reference conformations and assigned statistical 'weights' $\{w_i\}$ to the conformations such that $w_i \geq 0$, $\Sigma_i w_i = 1$ and the mean square difference between the observed and the weighted

*To whom correspondence should be addressed.

Abbreviations. NOE, nuclear Overhauser effect; BPTI, basic pancreatic trypsin inhibitor; rmsd, root- mean-square deviation.

average NOE cross-peak intensities is minimal. In principle, these weights constitute a maximum likelihood estimate of the relative populations of the reference conformations.

We have explored a similar approach, but with the primary goal of estimating the relative quality and subjective probability of the conformations found in distance geometry ensembles. Such a measure of 'goodness' is theoretically better justified and hence more meaningful than difference measures such as 'R-factors'. Unfortunately, the strong correlations between the NOE cross-peak intensities in the various conformations of the ensemble renders the least-squares equations extremely ill-conditioned, so that small changes in the observed intensities (on the order of the combined expected signal-to-noise and peak integration errors) can completely change the solution – most of the weights in fact come out equal to zero. By means of a somewhat ad hoc clustering approach, Landis and Allured (1991) were evidently able to overcome this problem. We have, however, found a more rigorous approach that also works well.

In our approach, we treat the individual 'pixels' in an experimental 2D proton NOESY spectrum as observations, and fit the spectra computed from the conformations in the distance geometry ensemble to them directly. In this way, the ratio of the number of equations to unknowns can be dramatically increased while at the same time entirely eliminating the errors due to integration of the cross peaks to obtain their intensities. Thus the 'probabilities', $\mathbf{p} = (p_i)$ of the conformations in the ensemble are estimated by solving the following least-squares problem:

$$\min(\|\mathbf{Ap}-\mathbf{z}\|^2) \text{ subject to } \mathbf{p} \geq 0 \text{ and } \sum_{i=1}^{N} p_i = 1,$$

where N is the number of conformations, the ij-th element a_{ij} in the matrix **A** is the value of the i-th pixel in the spectrum of the j-th conformation, and the i-th element z_i of the vector **z** is the value of the corresponding pixel in the experimental spectrum. This problem can be readily solved by the methods described by Lawson and Hanson (1974).

Because the scale used to measure pixel intensities is arbitrary, we rescale the vector **z** by $\alpha \geq 0$ to get the best possible fit between the observed and calculated spectra. This is easily done by solving the following modified least-squares problem:

$$\min(\|\mathbf{Ap}-\alpha\mathbf{z}\|^2) \text{ subject to } \mathbf{p},\alpha \geq 0 \text{ and } \sum_{i=1}^{N} p_i = 1,$$

(in effect, making $-\mathbf{z}$ a column of **A** and fitting the corresponding homogeneous linear system). Although a more sophisticated scaling procedure may be advisable (Nibedita et al., 1992), for the present we prefer to keep our approach as simple as possible.

To calculate the least-squares matrix **A** itself, we first calculate the relaxation matrix R of each conformation in the distance geometry ensemble by standard methods (see e.g. Borgias et al., 1991), treating methyl and phenyl group rotations by jump models. The matrix of NOE (cross)-peak intensities **I** at the chosen mixing time τ_m is then obtained from $-\tau_m$ **R** by matrix exponentiation, using an efficient new algorithm (Najfeld, I and Havel, T.F., unpublished results). Then the digital 2D spectra are calculated with each peak scaled so that its intensity is equal to the intensity of the corresponding entry in **I**, and the matrix **A** is built from the pixels therein. In order to reduce the size of the matrix **A**, only those pixels are included in **A** for which the intensity (of the pixel) is above a given cutoff value in at least one of the calculated spectra.

The method by which we actually calculate the 2D NOESY spectrum is outside the scope of this paper (Wagner, G., Beeson, N. and Hyberts, S., personal communication); nonetheless, we provide the following brief account for the sake of completeness. Assuming that the chemical shifts of the protons have been assigned, the individual cross peaks are split into multiplets according to the J-coupling constants implied by the Karplus relation, using the empirical relation given by Hyberts (1992). Although basic theory implies a Lorentzian line shape, we have found that a better fit can be obtained as a rule by using a circular 2D Gaussian model. Thus each peak of each multiplet is given a line shape of the form $\exp(-((\upsilon_1-\upsilon_1^c)^2 + (\upsilon_2-\upsilon_2^c)^2)/(2\sigma^2))$, where υ_1, υ_2 are the two frequency variables, $\upsilon_1^c,\upsilon_2^c$ are the frequencies (chemical shifts) of the two protons involved, and σ is the peak width. We call the package of programs we have written which implements all of the above calculations SESAME, meaning Structure Evaluation by Simulation And Minimization of Error.

To test the robustness of our estimation procedure in the presence of signal-to-noise error, we generated a set of 25 conformations for the 58-residue protein BPTI by distance geometry methods, using a set of previously published simulated NMR distance constraints (denoted as B-I in Havel, 1991). The rmsd among these conformations averaged 1.76 Å (0.94 Å among the α-carbons alone). Thirty 'experimental' spectra were simulated from the NOESY spectra calculated for each of these 25 conformations by choosing 30 random 'probability' vectors $\bar{\mathbf{p}}$ and computing the corresponding average spectra $\bar{\mathbf{z}} = \mathbf{A}\bar{\mathbf{p}}$ for each, where $\mathbf{A}$ is the least-squares matrix constructed from these 25 spectra as described above. The size of these spectra was 1024 × 1024, spanning the range of −1 through 11 ppm, and they were calculated using a correlation time of 2 ns, a mixing time of 50 ms, and a peak width of 10 Hz at a spectrometer frequency of 500 MHz. The peaks were assigned their published chemical shifts (Berndt et al., 1992) where available; missing resonances were omitted from the spectra. These omissions, as well as all the other approximations made in this paper, will not affect its main conclusions because we are evaluating our procedure on simulated data, and the simulated and calculated spectra are based on the same approximations.

In order to simulate the noise that would be expected in an actual NOESY spectrum, we simply added random numbers to the elements of $\bar{\mathbf{z}}$, where each random number had the same normal distribution. The variance of these random numbers was set to be equal to the square of the calculated intensity at the maximum of a cross peak between two single protons at distances ranging from 2.2 to 6.0 Å (we call this distance the noise level in 'effective Ångstroms'). The cutoff used to compute the matrix $\mathbf{A}$ was taken as the value of a pixel at 5σ for a cross peak between two single protons at 6.0 Å, which eliminated approximately 77% of the pixels, leaving a total of about 240 000 observations to fit. This should be compared to the about 1000 cross peaks that could be observed in the experimental spectrum. In Figs. 1 through 3, we plot the average for all 30 simulated spectra of the 'R-factor' $\Sigma_i(z_i-\bar{z}_1)^2/\Sigma_i\bar{z}_1^2$, the relative deviation $\Sigma_i(p_i-\bar{p}_i)^2/\Sigma_i\bar{p}_1^2$, and the number of zero probabilities $p_i = 0$ versus the noise level in effective Ångstroms. It should be noted that the calculated probabilities showed very little correlation with the R-factors, even in the presence of no noise.

As can readily be seen, with a realistic noise level of 4.0 effective Ångstroms we are able to reproduce our target probabilities to within 1.9 ± 0.4% relative precision. This shows that by basing our fit on the pixels instead of the integrated peak intensities, we obtain enough observations to be able to distinguish and accurately rank the conformations in a typical distance geometry ensemble – assuming that our spectral simulation is accurate and that only random

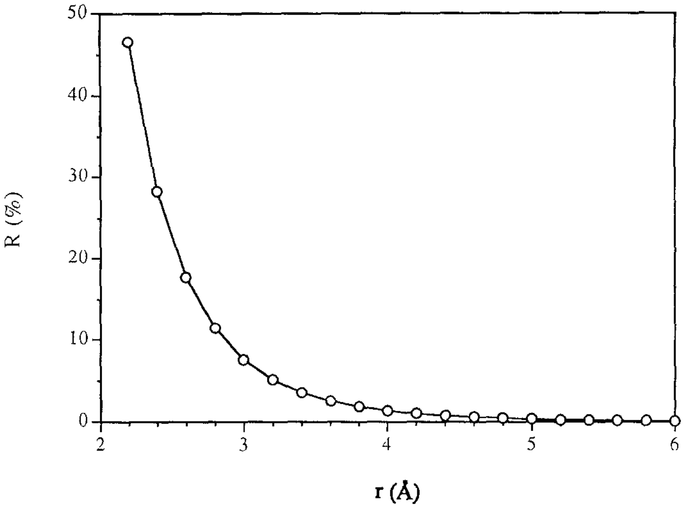

Fig. 1. The average for 30 simulated spectra plus noise of the 'R-factor' difference between the simulated and calculated spectra versus the simulated noise level ın effective Ångstroms (see text). The error bars cannot be shown because they are too small.

errors are present in the data. Of course, neither of these assumptions is entirely valid. For example, in order to render these calculations computationally tractable we have assumed that a finite distance geometry ensemble contains examples of all conformations present in significant concentration, that these conformations are sufficiently rigid to enable their spectral density functions to be approximated using a single correlation time, and that the line shapes in their

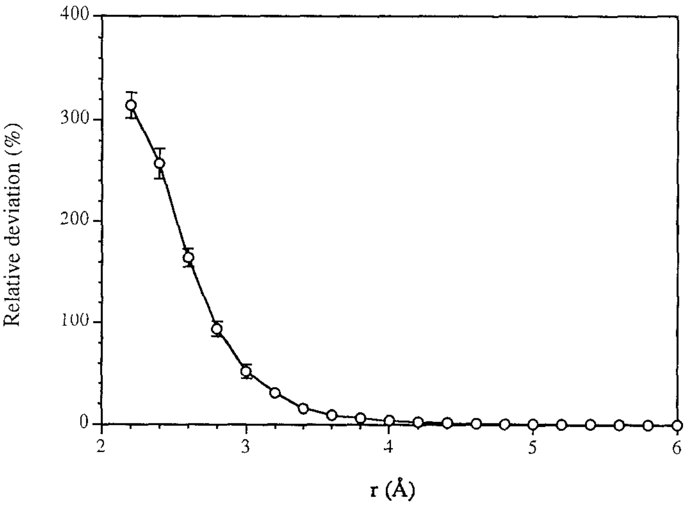

Fig. 2. The average for 30 simulated spectra plus noise of the relative deviation in percent between the assumed $\bar{\mathbf{p}}$ and calculated **p** probabilities versus the simulated noise level in effective Ångstroms (see text).

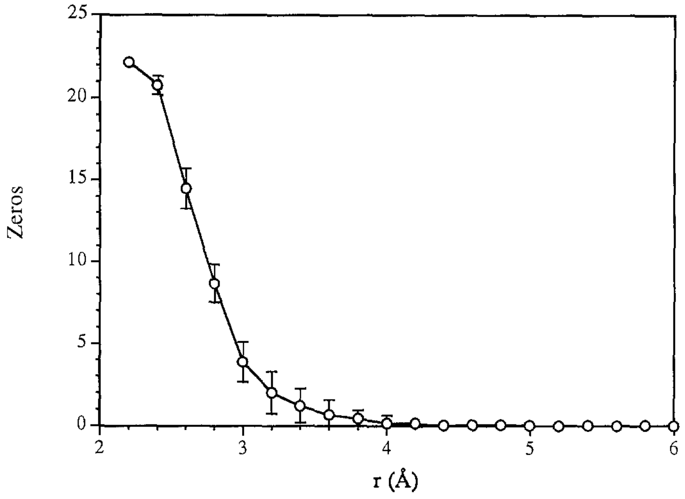

Fig. 3. The average for 30 simulated spectra plus noise of the number of calculated probabilities p_i equal to zero versus the simulated noise level in effective Ångstroms (see text).

spectra can be modeled by a uniform Gaussian. In addition, inherent in our averaging procedure is the assumption that the conformations in the ensemble are all sufficiently different that they interchange slowly on the T_1 time scale. Finally, the data themselves inevitably contain other systematic errors due to inaccurate phasing, chemical shift differences, window functions, and water suppression.

We point out that these problems are also present in earlier work wherein the fit was based on the integrated peak intensities (Landis and Allured, 1991). In contrast, the approach presented in this paper bypasses the tedious and error-prone process of peak integration entirely: it should not even be seriously affected by peak overlap (once the assignment problem has been solved). In addition, our approach makes a much larger number of observations available, thereby greatly improving the accuracy of the fit in the presence of random errors. Indeed, as we pointed out earlier in this paper, the seemingly straightforward approach of least-squares fitting the integrated cross-peak intensities directly is seriously affected by the noise levels present in typical experiments.

We further point out that, if one were to similarly base the fit on the spectra rather than the integrated cross-peak intensities in the usual sorts of NOE refinements, wherein the coordinates of a single conformation are modified to fit the calculated intensities to the data (see e.g. Case and Yip, 1989), then one should be able to obtain these same advantages in those refinements as well. In contrast to those refinements, however, the approach we have taken here enables one to back-calculate the data from an entire ensemble of conformations while at the same time defining a probability distribution on the ensemble which should be correlated with the actual solution populations. At the very least, it will certainly be an improvement over the 'equiprobability' assumption that is generally made in distance geometry calculations, and this probability distribution also provides an objective criterion by which one can rank and classify the members of distance geometry ensembles.

Of course, our method could be combined with the refinement of the individual conformations in the ensemble to obtain an even more accurate back-calculation, and we are presently in the process of evaluating such an approach with experimental data (Yang, J., Hyberts, S. and Havel, T.F., unpublished results). The fact that our method is insensitive to random errors should also make it possible to explore the significance of the above-mentioned systematic errors, and hence to develop better methods of dealing with them.

ACKNOWLEDGEMENTS

We thank G. Wagner, S. Hyberts, and N. Beeson of H.M.S. and C. Lawrence of the N.Y. Dept. of Public Health for useful discussions. This work was supported by NIH grants GM-38221 and GM-47467, and by a grant from BIOSYM Technologies, Inc.

REFERENCES

Berndt, K.D., Güntert, P., Orbons, L.P.M. and Wüthrich, K. (1992) *J Mol. Biol.*, **227**, 757–775.

Borgias, B.A., Gochin, M., Kerwood, D.J. and James, T.L. (1991) *Prog. NMR Spectrosc.*, **22**, 83–100.

Brüschweiler, R., Blackledge, M. and Ernst, R.R. (1991) *J. Biomol. NMR*, **1**, 3–11.

Case, D.J. and Yip, P. (1989) *J Magn. Reson.*, **83**, 643–648.

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

Hyberts, S. (1992) *Ph.D. Dissertation*, ETH, Zürich.

Hyberts, S G., Goldberg, M.S., Havel, T.F. and Wagner, G. (1992) *Protein Sci.*, **1**, 736–751.

Kessler, H., Griesinger, C., Laut z, J., Müller, A., van Gunsteren, W.F. and Berendsen, H.J.C. (1988) *J Am. Chem. Soc.*, **110**, 3393–3396.

Kim, Y., Ohlrogge, J.B. and Prestegard, J.H. (1991) *Biochem. Pharm.*, **40**, 7–13.

Landis, C. and Allured, V.S. (1991) *J. Am. Chem. Soc*, **113**, 9493–9499.

Lawson, C.L. and Hanson, R.J. (1974) *Solving Least-Squares Problems*, Prentice-Hall, New York.

Nibedita, R., Kumar, R.A., Majumdar, A. and Hosur, R.V. (1992) *J. Biomol. NMR*, **2**, 467–476.

Wagner, G., Hyberts, S. and Havel, T.F. (1991) *Annu Rev. Biophys. Biomol Struct.*, **21**, 167–198.