Journal of Biomolecular NMR, **13:** 133–137, 1999.
KLUWER/ESCOM

Determination of order parameters and correlation times in proteins: A comparison between Bayesian, Monte Carlo and simple graphical methods

Michael T. McMahon & Eric Oldfield*
Department of Chemistry, University of Illinois at Urbana-Champaign, 600 South Matthews Avenue, Urbana, IL 61801, U.S.A.

Received 8 September 1998; Accepted 2 October 1998

Key words: Bayesian, NMR relaxation

Abstract

We describe a novel approach to deducing order parameters and correlation times in proteins using a Bayesian statistical method, and show how likelihood contours, P(τ,S), and confidence levels can be obtained. These results are then compared with those obtained from a simple graphical method, as well as those from Monte Carlo simulations. The Bayes approach has the advantage that it is simple and accurate. Unlike Monte Carlo methods, it gives useful contour plots of probability (also not provided by the simple graphical method), and provides likelihood/confidence information. In addition, the Bayesian approach gives results in very good agreement with those obtained from Monte Carlo simulations, and as such use of Bayesian statistical methods appears to have a promising future for studies of order and dynamics in macromolecules.

Introduction

NMR relaxation and interference are very important tools for investigating structure and dynamics in proteins. Using such measurements, together with suitable analytical models, it has been possible to obtain pictures of molecular motion in proteins, but there is always the need to consider just what type of model should be used to analyze the results. The 'model-free' approach in one form or another (Lipari and Szabo, 1982a,b) is currently the most common method used to relate the experimental measurements to motional parameters: the order parameter (S), the overall molecular rotational correlation time (τ_c) and the internal rotational correlation time (τ_e). Typically, the relaxation equations are solved by using an optimization algorithm which gives a best fit to all of the data, and values for S^2 and τ_e are reported for each backbone atom investigated, together with a single value for the overall rotational correlation time, τ_c. Errors in S^2 and τ_e can then be estimated by performing Monte Carlo simulations.

More recently, Jin et al. (1997) have proposed an alternative, graphical approach in order to determine which regions of S^2, τ_c, τ_e space are permissible, based on estimated experimental uncertainties. They conclude that in some cases the order parameters (and correlation times) are rather inaccurate, since in this graphical method they may range over very large regions of permissible S^2, τ_e space.

It is, of course, possible to learn more about the effects of experimental errors on any conclusions which might be drawn through use of statistical techniques, and in this study, we extend the 'graphical' method proposed by Jin et al. (1997) to incorporate experimental errors in a more general way. In particular, we generate various likelihood/confidence contours which give direct information on the likelihoods that a particular S^2, τ_e solution is consistent with the experimental T_1, T_2 and NOE results, and their associated uncertainties. The results of a Bayesian statistics approach are then compared with graphical and Monte

*To whom correspondence should be addressed.

Carlo methods. The Bayesian approach appears particularly useful since it readily gives interesting relative probabilities, unlike the simple graphical method and the Monte Carlo method. The Bayesian method thus appears to be a useful extension of previous methods, and indeed it is quite possible that such methods may actually facilitate choices between different motional models (Hall, K., private communication), such as differentiating between stochastic diffusion and random-jump models for methyl rotation (Allerhand and Oldfield, 1973).

Results and discussion

In modeling relaxation data there are many possible alternatives for the spectral density function. One of the simplest is the well-known isotropic reorientation density function given in Lipari and Szabo (1982a,b), which is a function of the two internal parameters (S^2, τ_e) and one global rotational correlation time (τ_c). A more complex model, originally described by Woessner (1962), is the axially symmetric anisotropic diffusion function, which has four terms describing the overall rotation of the molecule. More complicated models include fully anisotropic diffusion, which includes 6 terms for the global motion, and the Brüschweiler et al. (1995) model which incorporates several diffusion tensors in a single protein (each having six terms). In what follows we will describe our approach using just the isotropic diffusion model, since this can be readily extended to the other models as well.

In most work to date, a χ^2-minimization has been used to deduce S^2, τ_e (Jin et al., 1997). In this approach, the χ^2-distribution function is given by (Palmer et al., 1991):

$$\chi^2(S^2, \tau_c, \tau_e) = \sum_{1\ldots N} \frac{(R_{1,obs} - R_{1,calc})^2}{\sigma_1^2} + \frac{(R_{2,obs} - R_{2,calc})^2}{\sigma_2^2} + \frac{(NOE_{obs} - NOE_{calc})^2}{\sigma_{NOE}^2} \quad (1)$$

where $R_1 = \frac{1}{T_1}$, $R_2 = \frac{1}{T_2}$ and σ = standard deviation.

Using T_1, T_2, and NOE measurements and the relaxation equations given in Lipari and Szabo (1982a,b) (and determining τ_c separately) enables expressions for χ^2 (S^2, τ_e) to be easily evaluated. Typically, this is accomplished by using a minimization approach such as the Levenburg-Marquardt algorithm (Press et al., 1986), although as pointed out by Jin et al. (1997) this does not permit a simple analysis of the precision in these parameters.

We therefore consider using Bayesian statistical inference to deduce the range of probable S^2, τ_e, τ_c solutions, given a limited number of experimental observations. For a given R_1 value determined in an experiment and having a known uncertainty, σ_T, the Bayes 'likelihood' function for the true $R_{1,actual}$ is given by:

$$\ell(R_{1,actual}\ \textit{given}\ R_{1,exp}) = \exp\left[-\frac{(R_{1,exp} - R_{1,actual})^2}{2\sigma_T^2}\right] \quad (2)$$

where $R_{1,exp}$ is the experimentally determined R_1 and σ_T is the uncertainty. In the case of *n* experimental measurements of this quantity, the likelihood function (or ^{1}Z-surface; Box and Tiao, 1992; Le et al., 1995; Heller et al., 1997) is then:

$$\ell(R_{1,actual}\ \textit{given}\ R_{1,exp}) = {}^1Z = \exp\left[-\frac{n(R_{1,actual} - \overline{R_{1,exp}})^2}{2\sigma_T^2}\right] \quad (3)$$

where $\overline{R_{1,exp}}$ is the average of $R_{1,exp}$ over the *n* experimental determinations. In the case of only one experimental measurement of R_1, one measurement of R_2, and one NOE measurement, the likelihood would be:

$${}^3Z(S^2, \tau_c, \tau_e) = \exp\left(\frac{-(R_{1,calc} - R_{1,obs})^2}{2\sigma_1^2}\right) \cdot \exp\left(\frac{-(R_{2,calc} - R_{2,obs})^2}{2\sigma_2^2}\right) \cdot \exp\left(\frac{-(NOE_{obs} - NOE_{calc})^2}{2\sigma_{NOE}^2}\right) \quad (4)$$

which is also equal to:

$${}^3Z(S^2, \tau_c, \tau_e) = \exp\left(\frac{-\chi^2}{2}\right) \quad (5)$$

as can be seen from Equations 1 and 4. The Bayesian ^{3}Z surface can be calculated as a function of S^2 and

τ_e for a given τ_c, which yields a maximum value corresponding to the most likely solution for S^2, τ_c and τ_e. The normalized Z surface (likelihood function) can then be integrated over regions of τ, S space to obtain the probability that the correct solution is within a given region.

Now, of the three parameters which need to be determined, only τ_c remains (generally) constant from site to site, and as such can be handled separately. The simplest way to do this is to use the ratio of T_1/T_2 (which is essentially only a function of τ_c (Kay et al., 1989)), then to minimize the difference between this and the experimental values for all sites. Determinate 'errors' – from residues which have appreciable flexibility, can be eliminated by discarding data for which the following inequalities apply:

$$\frac{T_1}{T_2} - \left\langle \frac{T_1}{T_2} \right\rangle > 1.5\sigma_{\frac{T_1}{T_2}} \tag{6}$$

$$\text{NOE} < 0.5 \tag{7}$$

In Figure 1 we show an example of a Bayesian (Figure 1A) plot for the human type-α (epidermal) transforming growth factor from relaxation measurements performed by Li and Montelione (1995) which can be compared with the results of a Monte Carlo simulation (Figure 1B). The 21 sites obeying the above inequalities were used to construct these plots, without taking into account R_{ex} contributions, and Figure 1A shows $\prod_{1-21}{}^{1}Z$ as a function of τ_c for these 21 residues, normalized such that $\int_0^{\infty} {}^{21}Z(\tau_c)d\tau_c = 1$. There is clearly a maximum in the ${}^{21}Z$ function around $\tau_c = 4.4$ ns. Also shown are the 25%, 50%, 75% and 95% confidence intervals. The results of the Monte Carlo approach shown in Figure 1B are virtually identical to the Bayesian results, Figure 1A, both in terms of τ_c and overall range in τ_c values.

We have also investigated the possible uses of conventional χ^2 methods in deducing confidence intervals. However, a χ^2 probability distribution function has 25%, 50%, 75% and 95% confidence intervals which are clearly much broader (data not shown) than those produced by the Bayesian/Monte Carlo methods, due to the non-linear dependence of τ_c on T_1/T_2, and in such cases the goodness-of-fit method is inappropriate (Press et al., 1986). Also, using the F-test to compare the χ^2 minimum with χ^2 at other points is not permissible since the χ^2 values being compared are not statistically independent. Thus, χ^2 methods are less suitable for generating confidence widths then are the Bayesian and Monte Carlo methods.

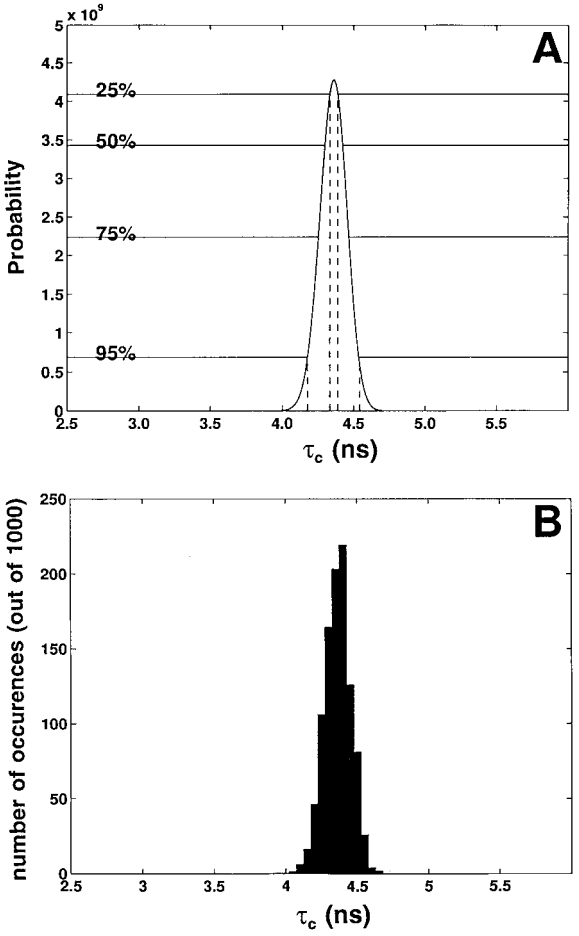

Figure 1. (A) ${}^{21}Z$ probability plot of T_1/T_2 vs. τ_c for 21 residues in the hTGFα protein using data from Jin et al., (1997) and Li and Montelione (1995), normalized such that $\int_0^{\infty} {}^{21}Z(\tau_c)d\tau_c = 1$. The maximum occurs at 4.4 ns. The 25%, 50%, 75%, 95% confidence intervals are shown (—). (B) Histogram of a Monte Carlo simulation of the same data. The Monte Carlo result is essentially identical to the Bayesian result (A).

Next, we consider S^2 and τ_e. We show in Figure 2 a comparison of the simple graphical method used previously (Jin et al., 1997) with the Bayesian approach. Figure 2A shows the overlapping regions generated from R_1, R_2, and NOE measurements for Ala^{31} in human type-α (epidermal) transforming growth factor (Li and Montelione, 1995; Jin et al., 1997). For the Bayesian analysis, the ${}^{3}Z$ likelihood surfaces can immediately be plotted as a function of S^2, τ_e, as shown in Figure 2B. Here, we show a contour plot using the same raw data as employed in Figure 2A. The maximum (the most probable solution) occurs at $\tau_e = 0.054$ ns, $S^2 = 0.825$ with the unnormalized likelihood (${}^{3}Z$ value) equal to 0.9, and the 25%, 50%,

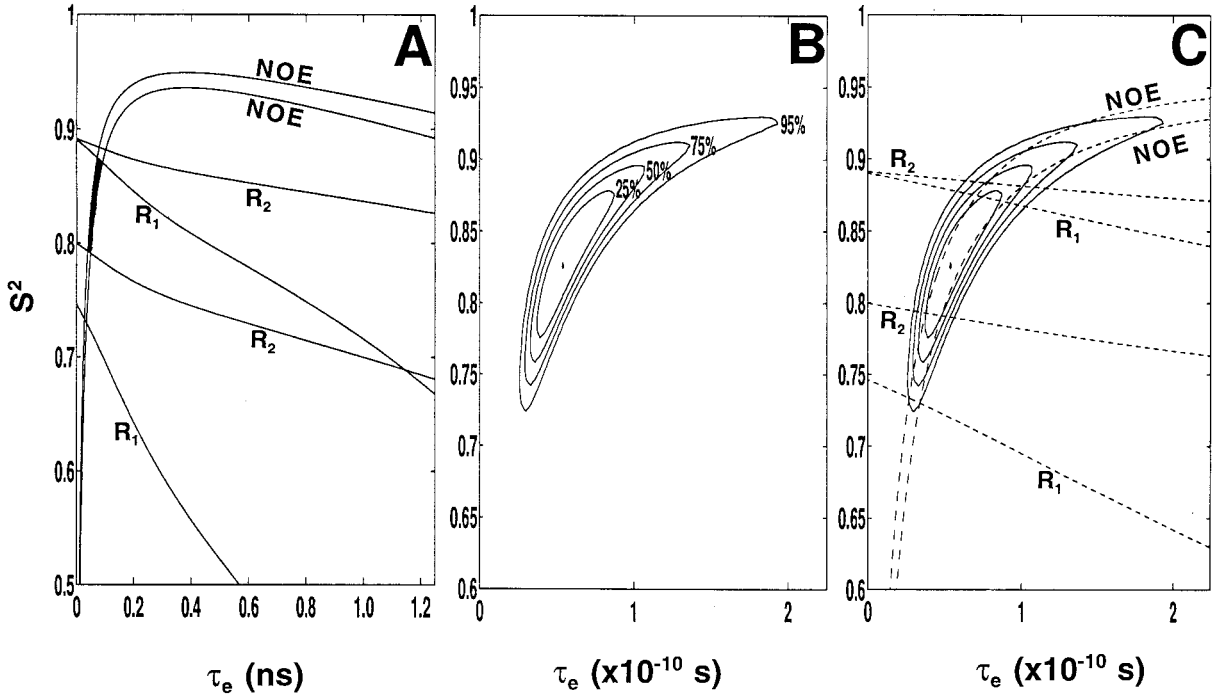

Figure 2. Comparison between different graphical approaches to determining S^2, τ_e. (A) Graphical approach to S^2, τ_e of Jin et al. (1997) shown for Ala^{31} in the hTGFα protein. (B) Contour plot of Bayesian likelihood for S^2 and τ_e for Ala^{31} in hTGFα, given the following: $\tau_c = 3.76$ ns (the value chosen in Jin et al. (1997) and Li and Montelione (1995), $R_1 = 2.5\ s^{-1}$, $R_2 = 5.36\ s^{-1}$, NOE = 0.62, $\sigma_{R1} = 0.22$, $\sigma_{R2} = 0.29$, $\sigma_{NOE} = 0.01$, ^{15}N CSA = 160 ppm, N-H bond length = 1.02 Å. The contours correspond to 25%, 50%, 75%, and 95% levels of confidence, respectively. (C) Comparison between Bayesian and simple graphical approach. R_1, R_2, NOE min/max lines are dashed, Bayesian contours are solid and as in B.

75% and 95% – confidence level contours are labeled. Including up to a 75% confidence level, Figure 2B, the parameter range covers $S^2 = 0.72$ to 0.85, and $\tau_e = 0.03$ to 0.13 ns. We also show for comparative purposes in Figure 2C the earlier graphical results of Jin et al. superimposed upon the Bayesian levels. The Bayesian approach is clearly more readily interpreted in terms of the most likely S^2, τ_e values. But how do these results compare with those obtained by using a Monte Carlo approach?

Figure 3 shows the results of a Monte Carlo simulation. In Figure 3A we show a plot in which 10 000 simulated points lie on an S^2, τ_e surface, while in Figure 3B the Bayesian contours are shown, for comparison. Of these points, 9 486 lie within the region shown in Figure 3B. There is near perfect agreement between the Monte Carlo and Bayesian results, as can be seen in Figure 3B, although there are some advantages to the Bayesian approach in terms of visualization.

In summary then, the results we have shown above help simplify the problem of extracting motional information from NMR spectra. We have presented the results of a Bayesian inferential treatment of likelihood to estimate S^2 and τ_e given τ_c. Moreover, the explicit incorporation of experimental error estimates in terms of Gaussian distributions enables more meaningful graphical representations of the results than use of simple cut-off or restriction plots (Jin et al., 1997). In the future, such graphical methods should also be of use in clarifying motional models (Allerhand and Oldfield, 1973; Hall, K., private communication), bond lengths, and chemical shielding anisotropies. In addition, other parameters, such as the ^2H electric field gradient (LiWang and Bax, 1997) as well as dipole-dipole/chemical shielding anisotropy cross-correlation results (Tjandra et al., 1996; Tjandra and Bax, 1997a,b; Tessari et al., 1997) can also be incorporated into motional models using a multiple Z-surface approach, such as we have recently used in another context to incorporate up to six spectroscopic observables in a formally similar Bayesian analysis, in heme proteins (McMahon et al., 1998). Overall, the Bayesian approach is simple and rapid and is ideally suited for representing dynamical information in a

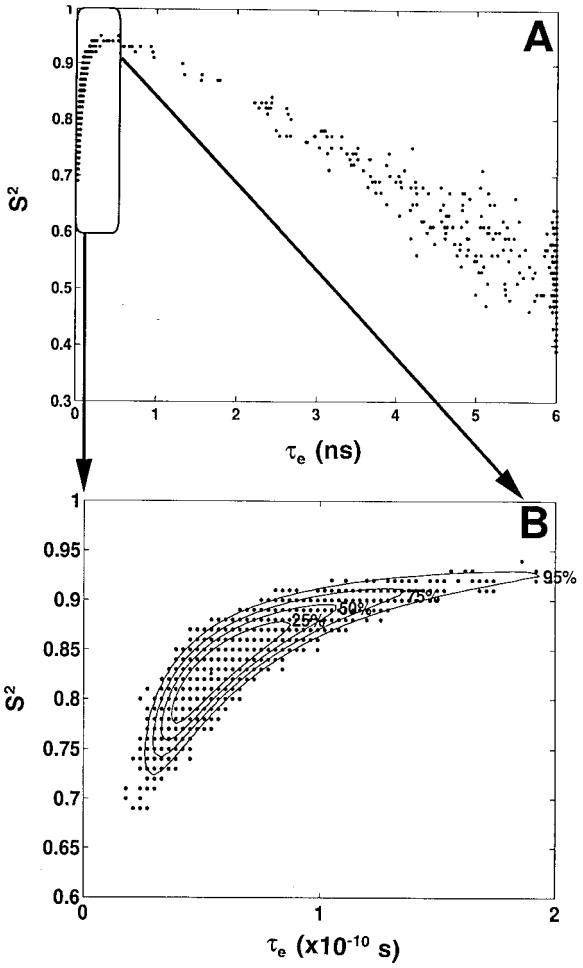

Figure 3. (A) Plot of a 10 000 point Monte Carlo simulation versus S^2 and τ_e for Ala^{31} in the hTGFα protein, using the same data as in Figure 2. Monte Carlo data points are represented by '·'. (B) Comparison of Monte Carlo results with the Bayesian approach, contours are solid, simulation points are dots. Of the 10 000 simulated points, 9 486 lie within the 95% contour region.

convenient graphical manner. In particular, our Monte Carlo results for τ_c, and for S^2, τ_e, are essentially indistinguishable from the Bayesian results, although the Bayesian method is more readily interpretable, and represents a considerable improvement over previous graphical methods.

Acknowledgements

We would like to thank Professor Barbara Bailey and Professor John Marden, Department of Statistics, University of Illinois, for valuable discussions. This work was supported by the United States Public Health Service (National Institutes of Health Grants HL-19481, GM-50694 and GM-08276).

References

Allerhand, A. and Oldfield, E. (1973) *J. Chem. Phys.*, **58**, 3115–3116.

Box, G.E.P. and Tiao, G.C. (1992) *Bayesian Inference in Statistical Analysis*, John Wiley and Sons, New York, NY.

Brüschweiler, R., Liao, X. and Wright, P.E. (1995) *Science*, **268**, 886–889.

Heller, J., Laws, D.D., Tomaselli, M., King, D.S., Wemmer, D.E., Pines, A., Havlin, R.H. and Oldfield, E. (1997) *J. Am. Chem. Soc.*, **119**, 7827–7831.

Jin, D., Figueirido, F., Montelione, G.T. and Levy, R.M. (1997) *J. Am. Chem. Soc.*, **119**, 6923–6924.

Kay, L.E., Torchia, D.A. and Bax, A. (1989) *Biochemistry*, **28**, 8972–8979.

Le, H., Pearson, J.G., deDios, A.C. and Oldfield, E. (1995) *J. Am. Chem. Soc.*, **117**, 3800–3807.

Li, Y.-C. and Montelione, G.T. (1995) *Biochemistry*, **34**, 2408–2423.

Lipari, G. and Szabo, A. (1982a) *J. Am. Chem. Soc.*, **104**, 4546–4559.

Lipari, G. and Szabo, A. (1982b) *J. Am. Chem. Soc.*, **104**, 4559–4570.

LiWang, A.C. and Bax, A. (1997) *J. Magn. Reson.*, **127**, 54–64.

McMahon, M.T., deDios, A.C., Godbout, N., Salzmann, R., Laws, D.D., Le, H., Havlin, R.H. and Oldfield, E. (1998) *J. Am. Chem. Soc.*, **120**, 4784–4797.

Palmer, A.G., Rance, M. and Wright, P.E. (1991) *J. Am. Chem. Soc.*, **113**, 4371–4380.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes in Fortran*; Cambridge University Press, Cambridge, U.K.

Tessari, M., Vis, H., Boelens, R., Kaptein, R. and Vuister, G.W. (1997) *J. Am. Chem. Soc.*, **119**, 8985–8990.

Tjandra, N., Szabo, A. and Bax, A. (1996) *J. Am. Chem. Soc.*, **118**, 6986–6991.

Tjandra, N. and Bax, A. (1997a) *J. Am. Chem. Soc.*, **119**, 9576–9577.

Tjandra, N. and Bax, A. (1997b) *J. Am. Chem. Soc.*, **119**, 8076–8082.

Woessner, D.E. (1962) *J. Chem. Phys.*, **37**, 647–654.