Journal of Biomolecular NMR, **24:** 215–229, 2002.
KLUWER/ESCOM

Factors determining the reliable description of global tumbling parameters in solution NMR

Norma H. Pawley[a,**], Jason D. Gans[a] & Linda K. Nicholson[b,*]
[a]*Department of Applied and Engineering Physics, and* [b]*Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, U.S.A.*

Received 5 July 2002; Accepted 18 September 2002

Key words: anisotropy, dynamics, global tumbling, helical bundle, NMR

Abstract

An accurate description of global tumbling of a protein is essential for correct analysis and interpretation of internal dynamics and thermodynamics. The accurate fitting of global tumbling parameters is affected by the number of experimental relaxation data points available for analysis, the distribution of data points over the domain of the function describing the tumbling, the measurement error associated with the data, the error associated with use of an approximate functional form, and errors in the protein structure. We present an analysis of the influence of these factors on the error in global tumbling parameters and the corresponding error in the calculated T_1/T_2 values. We find that reduction of experimental and approximation error can compensate for a less-than-ideal quantity or distribution of data points, and that accurate parameters can be obtained for proteins with highly anisotropic distributions of bond vectors, as illustrated using the helical bundle protein G-CSF. This indicates that proteins with anisotropic distributions, such as the helical bundle class of proteins, should not summarily be excluded when selecting proteins for dynamic and thermodynamic analyses of ^{15}N backbone relaxation measurements.

Abbreviations: NORMAdyn, NMR Optimized Relaxation Modeling with Anisotropy, for dynamics analysis.

Introduction

^{15}N spin relaxation parameters have been widely used to characterize the motions of individual backbone N-H bonds within proteins, and have provided a powerful tool for probing the internal dynamics of proteins. In addition, it is increasingly common to quantitatively interpret protein dynamics in terms of thermodynamic parameters (Wand, 2001). Proper evaluation of site-specific internal motions and thermodynamic parameters requires careful consideration of a number of issues, including spectral overlap (Viles et al., 2001), anisotropic global tumbling (Osborne and Wright, 2001) and correlation between global and internal motions (Prompers and Bruschweiler, 2000; Tugarinov et al., 2002). For a standard model-free type of analysis (Lipari and Szabo, 1982; Mandel et al., 1995), once data processing is complete, the critical step is the separation of contributions from global and internal motions to the relaxation parameters. The key to this separation is an accurate description of the rigid body global tumbling of the protein. It has been widely recognized that correct characterization of global tumbling is hampered by the difficulty in distinguishing between the effects of anisotropy and conformational exchange on measured transverse relaxation rates (Tjandra et al., 1995; Schurr et al. 1994; Kroenke et al., 1998; Mandel et al., 1996; Andrec et al., 1999; de Alba et al., 1999), and a variety of methods for addressing this difficulty have been proposed (Kroenke et al., 1998; Tjandra et al., 1996; Osborne and Wright, 2001; Bernado et al., 2002; Pawley et al., 2001).

*To whom correspondence should be addressed. E-mail: lkn2@cornell.edu
**Present address: Los Alamos National Laboratory, Bioscience Division, Los Alamos, NM, U.S.A.

Accurate characterization of any function is affected by the number of available data points, their distribution across the domain of the function, the measurement error in the data points, and error associated with the use of an approximate functional form. The effect of these factors on reliability of tensor parameters in solid-state NMR and sampling strategies for spin-spin relaxation times have been explored (Hodgkinson and Emsley, 1997; Jones et al., 1996). For accurate characterization of a diffusion tensor using ^{15}N spin relaxation measurements, these factors translate to: number of N-H bond vectors, orientational distribution of N-H bond vectors, measurement error in T_1 and T_2 data, and use of the approximation that T_1/T_2 is unaffected by fast internal motions.

The effects of number and orientational distribution of bond vectors have been examined by others (Fushman et al., 2000, Lee et al., 1997), and by our own work (Pawley et al., 2001). It has been shown that when residues aligned with the unique axis of diffusion are removed from the data set used to determine the global tumbling, the result is skewed toward a more isotropic fit, since these residues hold the most information regarding the effects of anisotropic diffusion (Kroenke et al., 1998; Pawley et al., 2001). It has also been shown that simultaneous analysis of $^{13}C^{\alpha}$ and ^{15}N relaxation rate constants can improve the reliability of global tumbling parameters (Lee et al., 1997) by increasing the number of bond vectors available for analysis and improving the distribution of orientations. The effects of incomplete sampling of orientation space on determination of a second-rank tensor have been described (Fushman et al., 2000). In addition, it has been suggested that for accurate determination of global tumbling, the ideal distribution of bond vectors in a protein is isotropic in space, and that highly anisotropic distributions, such as those found in an α-helical bundle, will be unsuitable for accurate characterization of global tumbling parameters (Fushman et al., 2000).

We explore the effects of various bond vector distributions, errors, and sample sizes, on the accuracy of model parameters using both theoretical distributions of bond vectors, and a distribution obtained from a helical bundle, G-CSF (1rhg, (Hill et al., 1993)). Using the theoretical distributions we demonstrate that, in some cases, anisotropic distributions of bond vectors can yield more accurate tumbling parameters than an isotropic distribution. Using the actual N-H bond vector distribution of G-CSF we demonstrate that even highly anisotropic distributions such as those found in helical bundle proteins can be quite suitable for accurate characterization of global tumbling parameters. Returning to the theoretical distributions, we explore methods for increasing the accuracy of calculations for less-than-ideal distributions. Finally, we examine the differential impact of errors in the structure on the accuracy of model parameters for the four theoretical distributions of bond vectors.

Materials and methods

Overview

Synthetic data sets were constructed to address four specific questions:

(1) Is an isotropic distribution of bond vectors ideal for extracting accurate global tumbling parameters?
(2) Can accurate global tumbling parameters be determined for highly anisotropic distributions?
(3) How is accuracy affected by measurement error?
(4) How is accuracy affected by approximation error?

The characteristics of the data sets used to address these questions are outlined in Table 1, and are described in detail below. Data sets were constructed for all combinations in Table 1 (with the exception that 3.5% measurement error was not combined with small ($\tau_f < 20$ ps) approximation error) yielding 48 total combinations of theoretical distributions. Data sets mimicking G-CSF were analyzed for 9 different diffusion tensor orientations (*vide infra*).

For each data set, 500 subsets were generated by varying the internal motion parameters, bond vector orientations, and individual measurement errors, according to the appropriate distribution function (*vide infra*). Averaging over the subsets avoids bias that might be associated with a particular set of S^2, τ_f, α and error values. Noise-free values of T_1 and T_2 were calculated from the 'true' global tumbling parameters and randomly generated internal parameters of motion and bond vector orientations, then adjusted by an estimated noise factor. Once a synthetic data set was generated (*vide infra*), it was analyzed as if it were an experimental data set. Global tumbling parameters were extracted and compared with the 'true' parameters to determine the accuracy of the fits.

Construction of synthetic parameters of motion

A prolate, axially symmetric, anisotropic model of global tumbling was assumed. For the theoretical

Table 1. Characteristics of synthetic data sets

Distribution	Number of vectors (N_{tot})	Measurement error	Approximation error[a] (ps)
Theoretical $N_{\text{bounce}}(\alpha)$ $N_{\text{const.}}(\alpha)$ $N_{\cos}(\alpha)$ $N_{\sin}(\alpha)$	20, 50, 100, 200	2.5% or 3.5 %	$\tau_f < 600$ or $\tau_f < 20$
G-CSF mimic	112	2.5%	$\tau_f < 600$

[a]Error resulting from assuming that data is unaffected by fast internal motions, τ_f, *vide infra*.

distributions, the principal axes of the diffusion tensor were set to $D_\perp = 0.009$ (ns^{-1}) and $D_{||} = 0.018$ (ns^{-1}) to represent a relatively high degree of anisotropy ($D_{||}/D_\perp = 2.0$) that would potentially be misinterpreted if not properly characterized (Osborne and Wright, 2001). For the G-CSF mimic, $D_\perp$ and $D_{||}$ were set to the experimentally determined values (Fushman et al., 2000; Lee et al., 1997) $D_\perp = 0.012$ (ns^{-1}), $D_{||} = 0.016$ (ns^{-1}). To generate a single data set, N_{tot} pairs of S^2 and τ_f values were randomly drawn from the following distribution functions:

$$P_{S^2}\left(S^2\right) = e^{-\left(S^2-S_0^2\right)^2/2\sigma_S^2} \cdot \left[H\left(S^2\right) - H\left(S^2-1\right)\right] \Big/ \Gamma_1,$$

$$P_{\text{Gauss},\tau_f}\left(\tau_f\right) = e^{-\left(\tau_f-\tau_0\right)^2/2\sigma_\tau^2} \cdot \left[H\left(\tau_f\right) - H\left(\tau_f-\tau_{\text{cutoff}}\right)\right] \Big/ \Gamma_2$$

or

$$P_{\text{const.},\tau_f}\left(\tau_f\right) = \frac{1}{\tau_{\text{cutoff}}}\left[H\left(\tau_f\right) - H\left(\tau_f-\tau_{\text{cutoff}}\right)\right]$$

where Γ_i ($i = 1,2$) normalizes the probability, $H(x)$ is the Heaviside step function, $S_0^2 = 0.85$, $\sigma_S = 0.15$, $\tau_0 = 0.03$ ns, $\sigma_\tau = 0.03$ ns, $\tau_{\text{cutoff}} = 0.6$ ns for P_{Gauss,τ_f} (large allowable τ_f values lead to large approximation error) and $\tau_{\text{cutoff}} = 0.02$ ns for $P_{\text{const.},\tau_f}$ (small approximation error). Values of $\tau_f > 0.6$ ns were not considered, nor were any contributions from chemical exchange (R_{ex}), nor were any values resulting in nOe < 0.65, since residues with relaxation rates significantly influenced by internal motions are assumed to be removed by data filtering (Tjandra et al., 1996; Pawley et al., 2001; Kneller et al., 2002).

Construction of synthetic bond vector distributions

The angles, α, of the individual N-H bond vectors relative to the unique axis, z, of the diffusion tensor were randomly drawn from the following distribution functions using standard methods (Press et al., 1996):

$$N_{\text{const.}}(\alpha) = N_{\text{tot}}\frac{2}{\pi}\left[H(\alpha) - H\left(\alpha-\frac{\pi}{2}\right)\right]$$
uniformly distributed in α,

$$N_{\sin}(\alpha) = N_{\text{tot}}\sin(\alpha)\left[H(\alpha) - H\left(\alpha-\frac{\pi}{2}\right)\right]$$
uniformly distributed in space,

$$N_{\cos}(\alpha) = N_{\text{tot}}\cos(\alpha)\left[H(\alpha) - H\left(\alpha-\frac{\pi}{2}\right)\right],$$

$$N_{\text{bounce}}(\alpha) = N_{\text{tot}}\left(\frac{\pi}{2\theta_0} - B\left(\frac{\pi}{2\theta_0}-1\right)\right)\cos\left(\frac{\pi\alpha}{2\theta_0}\right)\left[H(\alpha) - H(\alpha-\theta_0)\right] + B\sin\left(\frac{\pi\left[\alpha-\theta_0\right]}{2\left[\frac{\pi}{2}-\theta_0\right]}\right)\left[H(\alpha-\theta_0) - H\left(\alpha-\frac{\pi}{2}\right)\right],$$

where $H(x)$ is the Heaviside step function, θ_0 is the magic angle (defined by $\theta_0 = \cos^{-1}\left(\sqrt{1/3}\right)$), and the constant, $B = 0.6$, determines the relative contributions from the sin and cos components of $N_{\text{bounce}}(\alpha)$. The distributions are not normalized;

rather, they integrate to N_{tot}, where N_{tot} varies from 20 to 200, as described in Table 1. The orientation of N-H bond vectors in the x–y plane were randomly drawn from a distribution of the form $P_{\text{const.}}(\phi) = \frac{2}{\pi}\left[H(\phi) - H\left(\phi - \frac{\pi}{2}\right)\right]$, resulting in an axially symmetric distribution of bond vectors. Only angles between 0 and $\pi/2$ are considered, due to the inherent symmetries of the problem (the relaxation rates depend only on the square of the bond vector components). These distributions were used for calculation of the synthetic relaxation data, as described below. In order to demonstrate the process of finding the correct diffusion tensor orientation, bond vectors were rotated out of the diffusion tensor frame, after the creation of the relaxation data, into a starting reference frame using Euler rotation matrices ($R(\psi, \theta, \phi)$ as defined in Arfken (1985), where ψ, θ, and ϕ define the relative orientation of the two frames, and ψ is defined to be zero for an axially symmetric diffusion tensor).

Construction of synthetic relaxation data

Values of T_1 and T_2 were calculated from the parameters of global and internal motion using standard equations (Woessner, 1962; Tjandra et al., 1996; Lee et al., 1997). The effects of measurement error were simulated by the addition of the Gaussian-distributed random variables δ_{T1} and δ_{T2}. These random variables were drawn from a Gaussian distribution with a width (i.e. standard deviation) of $2.5\% \times T_i$ or $3.5\% \times T_i$, to reflect the expected percent error in experimental data. When included, the effects of errors in the structure (i.e., errors in bond vector orientation) were simulated by the addition of Gaussian-distributed error of width 5 degrees (Skrynnikov et al., 2000).

Calculation of global tumbling parameters and errors

For each set of input data (i.e., a given vector distribution, N_{tot}, diffusion tensor orientation, and error size, summarized in Table 1) errors in global tumbling parameters were estimated by averaging over 500 subsets, each constructed from a different set of internal motions and bond vector orientations, selected from the appropriate distributions described above. This approach avoids bias associated with a particular set of internal motions and specific orientations. Global tumbling parameters for each subset were determined using NORMAdyn, as described in Pawley et al., (2001) with two exceptions. First, data was treated as 'post-filter' and filtering steps were not applied, since chemical exchange and large amplitude motions were not included when generating the data sets. Second, simulated annealing was performed only for the first of the 500 subsets in each data set. To reduce the computational expense of the remaining 499 subsets, a random jump away from the previous minimum was taken in each of the four dimensions ($D_\perp$, $D_{||}$, θ, ϕ), followed by conjugate gradient minimization. The random jump avoids initial condition bias for distributions in which insufficient sampling causes the solution space to be degenerate (i.e., the slope of χ^2 is zero along a particular direction), and was found to yield the same results as the more expensive simulated annealing protocol.

The following measures are used to describe the quality of a fit:

The F-statistic is a measure of how much additional fitted parameters associated with a given model improve χ^2, the 'goodness of fit' (Bevington and Robinson, 1992; Lee et al., 1997).

$$F_{n-m,N-n} = \frac{\left(\chi^2_{N-m} - \chi^2_{N-n}\right)(N-n)}{\chi^2_{N-n}(n-m)}$$

and

$$\chi^2_{N-m} = \sum_{i=1}^{N} \left[\frac{\frac{T_{1,i}^{\text{calc}}}{T_{2,i}^{\text{calc}}} - \frac{T_{1,i}^{\text{synth}}}{T_{2,i}^{\text{synth}}}}{\sigma_i} \right]^2,$$

where calc denotes the values calculated by fitting the model parameters, synth denotes the synthetic values (experimental values determined *in silico*), N is the number of data points (i.e., N-H bond vectors), n and m are the number of fitted parameters associated with each model ($n > m$), and σ_i is the error in the synthetic value of T_1/T_2. For an isotropic tumbling model, $m = 1$. For an axially symmetric tumbling model, $n = 4$. ε_d is a measure of the error in the principal values of the diffusion tensor (Fushman et al., 2000). For axially symmetric global tumbling,

$$\varepsilon_d = 100 \sqrt{\frac{2}{3}\left(\frac{D_\perp^{\text{calc}} - D_\perp^{\text{true}}}{D_\perp^{\text{true}}}\right)^2 + \frac{1}{3}\left(\frac{D_{||}^{\text{calc}} - D_{||}^{\text{true}}}{D_{||}^{\text{true}}}\right)^2},$$

ε_a is a measure of the error in the orientation of the diffusion tensor (Fushman et al., 2000).

$$\varepsilon_a = 100\left[1 - \frac{\text{Trace}\left|R^{\text{calc}}(\psi', \theta', \phi')\, R^{\text{true}}(\psi', \theta', \phi')^T\right|}{3}\right].$$

When the calculated orientation of the unique axis of the diffusion tensor is orthogonal to the true orientation, $\varepsilon_a = 100$.

$\tilde{\chi}$ is a measure of the error in the calculated values of T_1/T_2:

$$\tilde{\chi} = \sum_{\alpha_i=0}^{\alpha_i=90} \left[\frac{T_1^{\text{calc}}(\alpha_i)}{T_2^{\text{calc}}(\alpha_i)} - \frac{T_1^{\text{true}}(\alpha_i)}{T_2^{\text{true}}(\alpha_i)} \right]^2.$$

An approximate scale for $\tilde{\chi}$ can be obtained relative to the average value of T_1/T_2, which is protein dependent, varying with $D_{||}$, $D_{\perp}$, S^2 and τ_f. For the theoretical distributions, using $D_{\perp} = 0.009$ (ns^{-1}), $D_{||} = 0.018$ (ns^{-1}), $S^2 = 0.85$, $\tau_f = 30$ ps, $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$ when $\tilde{\chi} = 63$, where calc represents the calculated value of T_1/T_2, true represents the value of T_1/T_2 determined from error-free parameters, and σ represents the size of the synthetic error in T_1/T_2 ($\sigma = 0.035$ for 2.5% error in T_1 and T_2). For the G-CSF distributions, using $D_{\perp} = 0.012$ (ns^{-1}), $D_{||} = 0.016$ (ns^{-1}), $S^2 = 0.85$, $\tau_f = 30$ ps, $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$ when $\tilde{\chi} = 30$.

Finally, the generalized sampling parameter, Ξ, is defined as in Fushman et al. (2000)

$$\Xi = \frac{2}{3} \sum_{i,j=x,y,z} \frac{3\langle r_i r_j \rangle - \delta_{ij}}{2},$$

where r_i is the projection of a given unit vector r on the axis i (in an arbitrary reference frame) and δ_{ij} is the Kronecker delta. Ξ is used to quantify the distribution of vector orientations on a scale from 0 to 1. For a uniform distribution of vectors $\Xi = 0$; if all vectors are aligned, $\Xi = 1$.

Results and discussion

Revisiting the isotropic distribution of bond vectors

It has been suggested (Fushman et al., 2000; Lee et al., 1997) that a set of N-H bond vectors uniformly distributed in space ($\Xi \sim 0$) is ideal for accurate characterization of global tumbling parameters. Largely as a consequence of this suggestion, it has also been proposed that highly anisotropic distributions ($\Xi > 0.25$), such as those found in a helical bundle (e.g., G-CSF, $\Xi = 0.49$), will be unsuitable for accurate characterization of global tumbling parameters.

In order to examine the effects of bond vector distribution, four specific distributions were created and analyzed. The bond vector orientations for each distribution are shown over the first octant of a unit sphere in Figure 1a. These distributions can also be visualized as function of α, as shown in Figure 1b. For each distribution, the effects of varying N_{tot}, measurement error, and approximation error were evaluated as described (*vide supra*). The quality of the fits is shown in Figure 2.

The F-statistic (Figure 2a) is used to assess whether use of an axially symmetric diffusion tensor provides a significant improvement in the fit relative to isotropic diffusion. Large values of F indicate significant improvement. While axially symmetric tumbling is the correct model for all points in Figure 2a, data sets with few data points ($N_{\text{tot}} = 20$) and data sets with an isotropic distribution of bond vectors ($N_{\sin}(\alpha)$) display lower F-statistics than data sets with more data points or anisotropic distributions of bond vectors. In both cases, the cause of the decreased F-statistic (lower significance) is due to insufficient sampling of the axially symmetric functional form.

The error in the principal values of the diffusion tensor, the orientation of the diffusion tensor and the estimated T_1/T_2 values are shown in Figures 2b–d. For all three measures, the error decreases with increasing N_{tot}. For all three error measures, and all values of N_{tot}, the error is largest for the $N_{\sin}(\alpha)$ distribution, i.e., for an isotropic distribution of N-H bond vector orientations. The error in the principal values of the diffusion tensor (ε_d, Figure 2b) and in the estimated T_1/T_2 values ($\tilde{\chi}$, Figure 2d) is smallest for $N_{\text{bounce}}(\alpha)$ at all values of N_{tot}. This implies that the error in estimated T_1/T_2 values is dominated by the error in the principal values of the diffusion tensor, since the error in orientation of the diffusion tensor follows a different pattern (ε_a, Figure 2c). The distribution with the smallest error in the orientation of the diffusion tensor varies with N_{tot}. In summary, the results in Figure 2 illustrate that, by every measure, the worst distribution (largest error, smallest F-statistic) was the isotropic distribution ($N_{\sin}(\alpha)$, $\Xi = 0.02$). Overall, the best distribution was $N_{\text{bounce}}(\alpha)$, $\Xi = 0.38$. It is clear that a uniform distribution in space is not the ideal distribution of bond vectors for obtaining an accurate description of global tumbling.

These results are most easily explained by examining the functional form to which data is being fit, as shown in Figure 3. $N_{\text{bounce}}(\alpha)$ places the largest number of vectors (data points) at the extrema of the function, and in the regions where the value of T_1/T_2 for an anisotropically tumbling protein differs most strongly from the value of T_1/T_2 for an isotropically tumbling protein (thick dashed line in Figure 3). In contrast, $N_{\sin}(\alpha)$ places very few vectors at the $\alpha = 0$ extremum, where T_1/T_2 differs most strongly be-

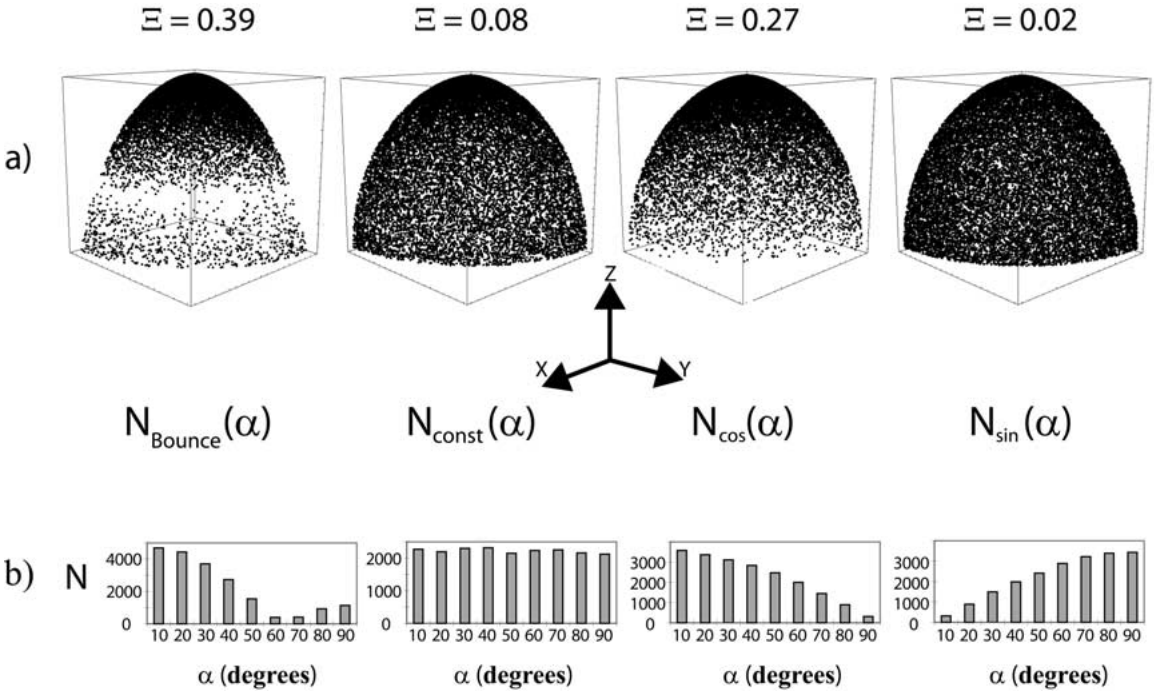

Figure 1. Synthetic bond vector distributions. (a) Distributions are shown as points on a unit sphere, with each point on the surface of the sphere representing the tip of a vector. The orientation of the coordinate axes and bond vectors is given with respect to the 'diffusion frame'. Only the first octant is shown. (b) The relative numbers of vectors (N) for each distribution are shown as a function of α, the angle between an N-H bond vector and the unique axis of the diffusion frame.

tween isotropically and anisotropically tumbling proteins. In summary, the performance of the distributions from best to worst is $N_{\text{bounce}}(\alpha)$, $N_{\text{const.}}(\alpha)$, $N_{\cos}(\alpha)$, $N_{\sin}(\alpha)$. The excellent performance of $N_{\text{bounce}}(\alpha)$ and $N_{\text{const.}}(\alpha)$ suggests that placing the largest number of points at the extrema is most advantageous, and a slight advantage is gained by emphasizing the $\alpha = 0$ extremum, as indicated by the smaller errors for $N_{\text{bounce}}(\alpha)$ versus $N_{\text{const.}}(\alpha)$. When data is available at only one extremum, as for $N_{\cos}(\alpha)$ and $N_{\sin}(\alpha)$, an advantage is still gained by emphasizing the $\alpha = 0$ extremum, as indicated by the smaller errors for $N_{\cos}(\alpha)$ versus $N_{\sin}(\alpha)$.

Accurate global tumbling parameters can be determined for highly anisotropic distributions

The distribution that is isotropic in space has the advantage that, while it is not the best, it is consistent. That is, the behavior of this distribution is the same, regardless of the orientation of the molecular frame with respect to the diffusion tensor. In contrast, the behavior of other distributions can change, depending on the relative orientation of the molecular and diffusion frames (Figure 5). Hence, an anisotropic distribution of bond vectors may perform well in one orientation, but not in another.

In order to explore the effects of orientation on the extraction of global tumbling parameters for an anisotropic distribution of vectors, synthetic data sets were created based on the distribution of bond vectors in G-CSF (Figure 4a; Hill et al., 1993) and its global tumbling parameters, as reported by Lee et al. (1997). The distribution of G-CSF bond vectors in the molecular (inertia) frame is depicted on the surface of a unit sphere in Figure 4b, and in a bar graph in Figure 4c. If the diffusion tensor is assumed to coincide with the inertia frame, this distribution has the largest number of bond vectors at $\alpha < 45$ deg (i.e., they lie close to the unique axis of the diffusion tensor). In contrast, the experimentally determined orientation of the diffusion tensor with respect to the molecular frame yields $\theta = 77°$, which places the largest number of bond vectors perpendicular to the unique axis of the diffusion tensor (as in highlighted distribution, Figure 5a). What size errors would we expect in the calculated parameters for this distribution, and how would the size of the errors change if the orientation

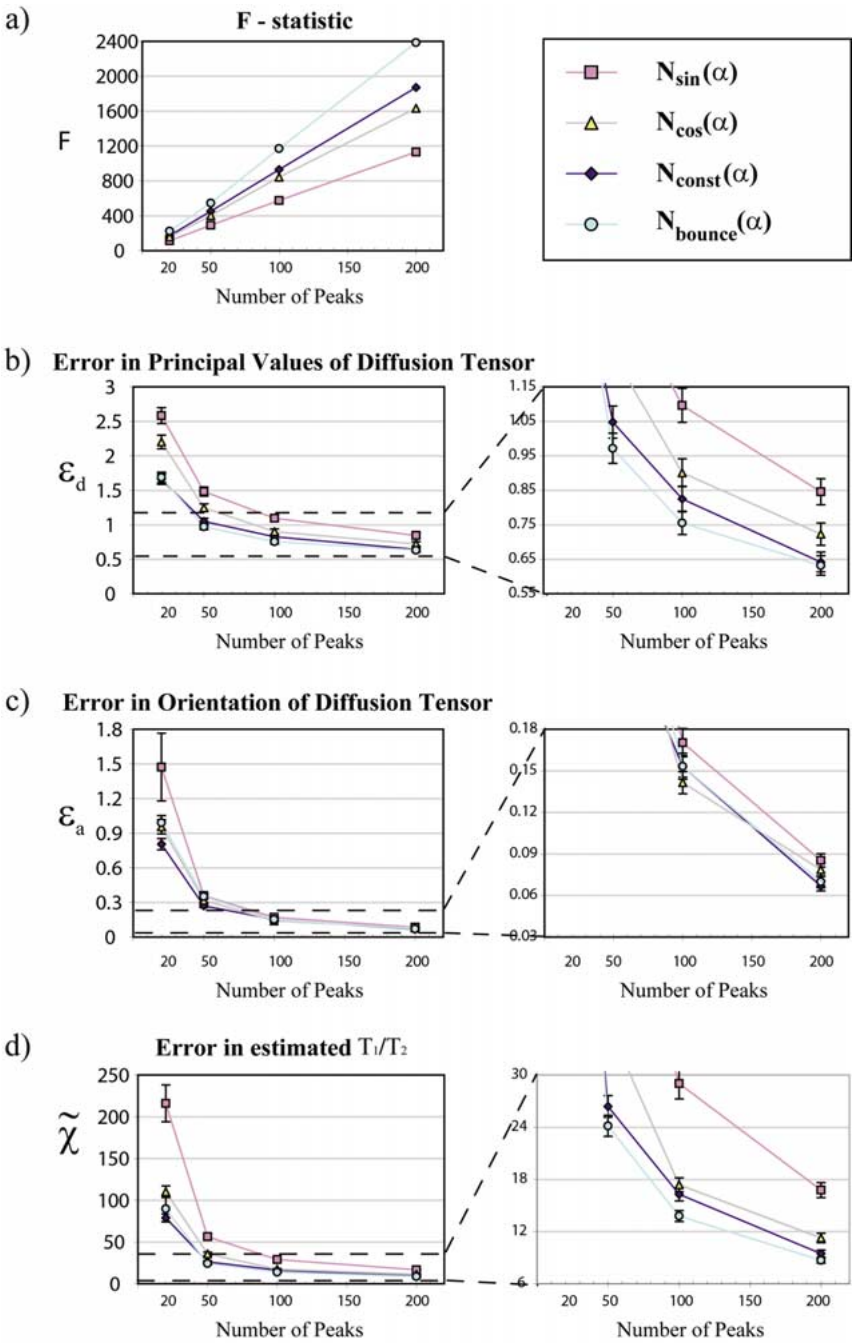

Figure 2. Effects of number and distribution of N-H bond vectors on calculated dynamic parameters. Statistics are computed for four distributions of N-H bond vectors. The number of data points in each distribution varies from 20 to 200. Reported values include the effects of 2.5% measurement error and large ($\tau_f < 600$ ps) approximation error. Reported values are the average over 500 data sets. Error bars are calculated from the standard deviation over 500 data sets. In general, the number of N-H bond vectors, N_{tot}, more strongly affects the accuracy of the calculation than the particular distribution. Overall, $N_{\text{bounce}}(\alpha)$ provides the most accurate estimate of global tumbling parameters, while $N_{\text{sin}}(\alpha)$ provides the least accurate estimate. (a) The F-statistic (isotropic model vs. axially-symmetric anisotropic model). (b) ε_d, the error in the principal values of the diffusion tensor. (c) ε_a, the error in the orientation of the diffusion tensor. (d) $\tilde{\chi}$, the error in the calculated values of T_1/T_2. Note that $\tilde{\chi} > 63$, the value for which $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$, for $N_{\text{tot}} = 20$, but $\tilde{\chi} < 63$ for all $N_{\text{tot}} \geqslant 50$. $\tilde{\chi}$ is more strongly affected by errors in the principal values of the diffusion tensor (ε_d) than by errors in the orientation of the diffusion tensor (ε_a).

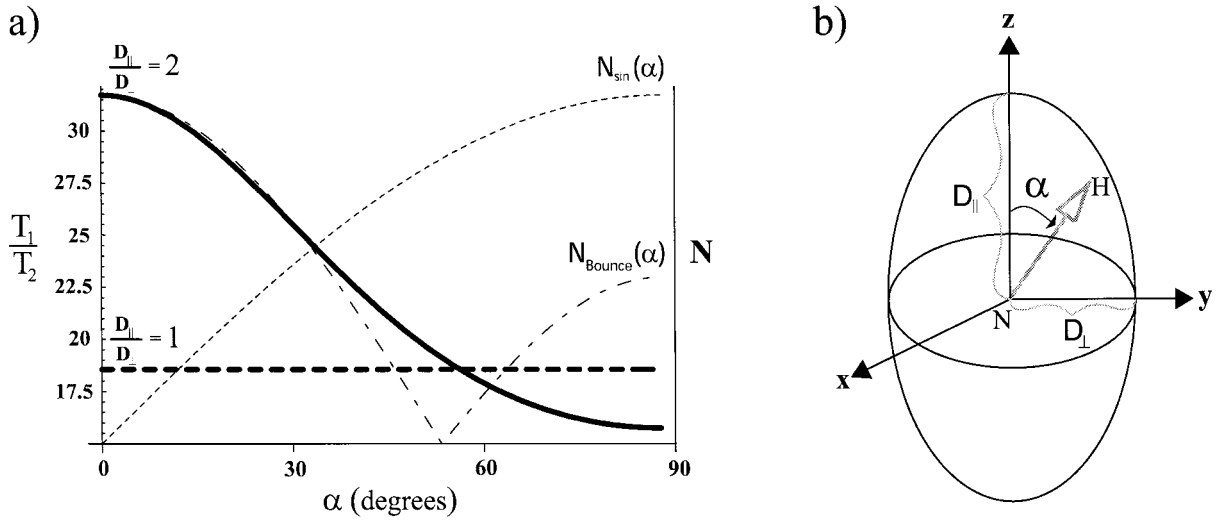

Figure 3. Dependence of functional form and synthetic distributions on α. (a) Thick lines show T_1/T_2 as a function of α, for $D_{||}/D_{\perp} = 2$ (solid) and $D_{||}/D_{\perp} = 1$ (dashed, isotropic tumbling). Note that the effect of anisotropic tumbling is more readily apparent in the T_1 and T_2 values of residues with $\alpha \approx 0°$ than in residues with $\alpha \approx 90°$. Thin lines show $N_{\text{bounce}}(\alpha)$ (dashed) and $N_{\sin}(\alpha)$ (dotted). $N_{\text{bounce}}(\alpha)$ places a larger number of data points at both extrema of the function, with more points in the region where anisotropic tumbling differs most significantly from isotropic tumbling. $N_{\sin}(\alpha)$ places a large number of points at only one extremum of the function, with few points in the region where anisotropic tumbling differs most significantly from isotropic tumbling. (b) Ellipsoid corresponding to a prolate diffusion tensor showing an N-H bond vector at an arbitrary angle, α, from the principal axis of diffusion.

of the diffusion tensor were altered with respect to the molecular frame? To answer this question, the orientation of the diffusion frame was rotated in discrete steps with respect to the molecular frame to produce the distributions shown in Figure 5a. While changes in orientation affect the accuracy of the parameters, the errors are small (ε_d and $\varepsilon_a < 1.6\%$, $\tilde{\chi} < 16$) for all orientations considered here (Figure 5b) and fall in the expected range for the number of available data points, $N_{\text{tot}} = 112$ (Figures 2b–d). In other words, although different distributions over α result from different relative orientations of the molecular and diffusion frames, even in the worst observed case (i.e., distribution maximum at 90°) the errors remain small. The conclusion is that even a highly anisotropic distribution, such as that found in the helical bundle, G-CSF, can indeed be suitable for accurate global tumbling analysis.

Increasing the accuracy of calculations for less-than-ideal distributions

In a PDB survey by Fushman et al. (2000) 85.7% of proteins surveyed were found to have acceptable (i.e., sufficiently isotropic) distributions of N-H bond vector orientations for accurate global tumbling analysis. We have demonstrated that this is the minimal set of acceptable proteins, since anisotropic distributions ($\Xi > 0.25$) may also yield accurate global tumbling parameters. Nevertheless, it is quite likely that distributions exist that are far from ideal, that would potentially cause difficulty in the analysis of global tumbling.

When data is limited, it is important to maximize the information content of each data point by maximizing the signal to noise ratio. Two types of noise (error) are easily addressed: Measurement error and approximation error. Measurement error is affected by a host of factors, including pulse calibration, sample concentration, number of scans performed, number of points sampled to determine the exponential decay, spectral overlap, intrinsic linewidths, and use of a cryoprobe. In a positive sense, this provides many points of intervention at which measurement error can be reduced. The quality of fits for our theoretical distributions, with measurement errors of 3.5% and 2.5% (i.e., $3.5\%/\sqrt{2}$) is shown in Figure 6. In all cases considered here, the benefit of decreasing the measurement error by $\sqrt{2}$ is equivalent to the benefit of increasing the number of data points by $\sim$75–90%. Note that this increase in data points is accompanied by increased sampling of orientational space. This result can be used to help guide experimental cost-benefit-analyses. For example, it has been proposed (Lee et al., 1997) that non-ideal distributions be improved by simultaneous analysis of ^{15}N and $^{13}C^{\alpha}$ relaxation.

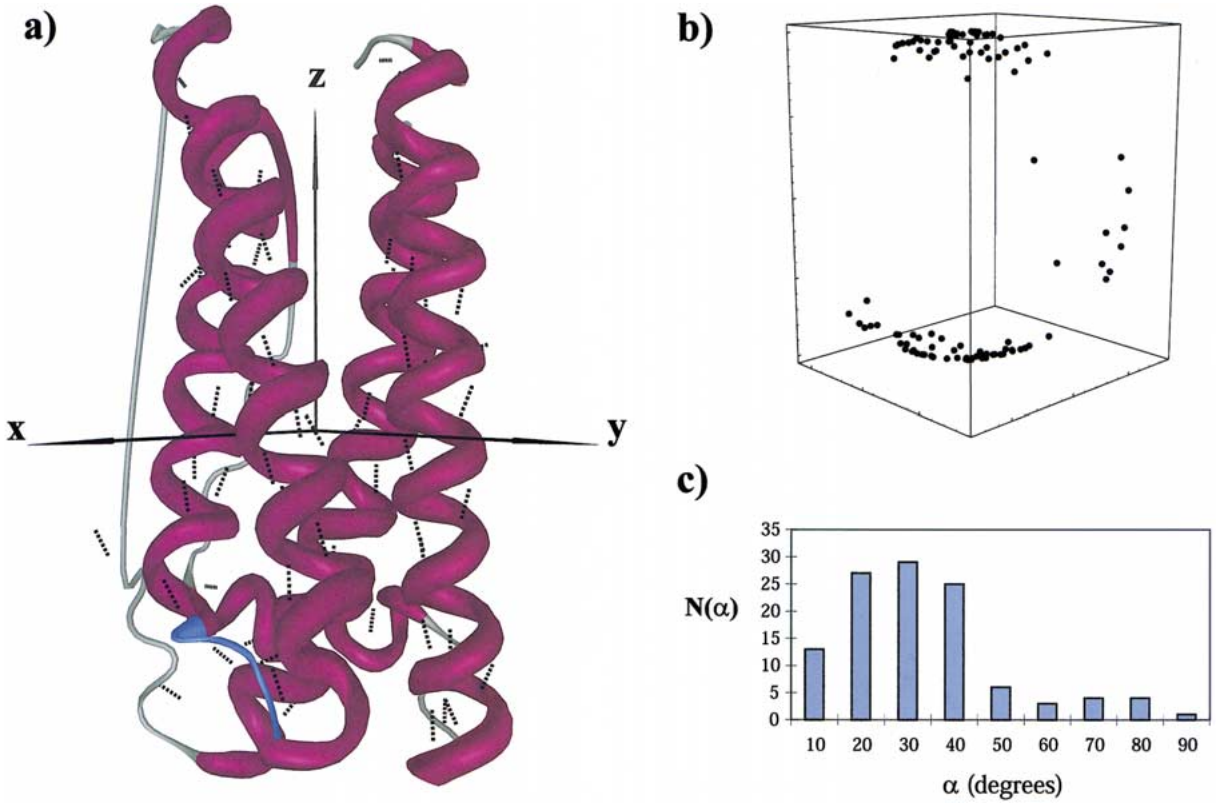

Figure 4. Structure and bond vector distribution of G-CSF. (a) The structure of G-CSF (1rhg) is shown oriented in the moment of inertia frame. N-H bond vectors are shown as dotted lines. (b) The distribution of bond vectors in alpha helices is shown as points on a unit sphere. (c) The distribution of bond vectors in alpha helices shown as a function of α_I, where α_I is the angle between an N-H bond vector and the unique axis of the inertia frame.

While this can indeed provide improvement, there are several drawbacks to this approach. First, the process is both time-consuming and expensive, requiring additional relaxation experiments and isotopic enrichment of the protein with ^{13}C. Second, direct fitting of T_1/T_2 data is difficult if measurements are made for different nuclear species, since the functions depend explicitly on the Larmor frequencies of the nuclear spins. The local diffusion approximation, which would allow one to work around this difficulty, breaks down for large anisotropy (Lee et al., 1997). Third, while this method will always increase the number of available data points, it will not always improve the range of sampled orientations. In particular, addition of $^{13}C^{\alpha}$ relaxation is not expected to improve the bond vector distribution for proteins with predominantly β-sheet secondary structure (Fushman et al., 2000). Finally, the measurement uncertainties tend to be larger (by a factor of ~ 2) for $^{13}C^{\alpha}$ relaxation data than for ^{15}N relaxation data. Consequently, a single $^{13}C^{\alpha}$ data point typically contains less information than a single ^{15}N data point. Hence, knowledge of the distribution of $^{13}C^{\alpha}-^{1}H^{\alpha}$ vectors with respect to the N-H vectors, the resolution of the ^{13}C NMR spectrum, and the expected effort required to acquire and analyze the $^{13}C^{\alpha}$ relaxation data can be weighed against the expected effort required to decrease the measurement error for N-H bond vectors alone.

Approximation error results from fitting experimental values of T_1/T_2 to a functional form which assumes that the ratio is unaffected by fast internal motions, τ_f. The expected deviation from this approximate form varies with the amplitude and time scale of the internal motion, as illustrated in Figure 7a. When internal motions reduce the values of T_1/T_2, but are ignored in the fitting, the global tumbling parameters must compensate, introducing systematic error. The primary means of removing this error is to fit the data using only those residues that are largely unaffected by τ_f ($\tau_f < 20$ ps, Model 1 as described by Mandel et al., 1995). This certainly reduces the noise, but unfortunately reduces the signal as well. In our expe-

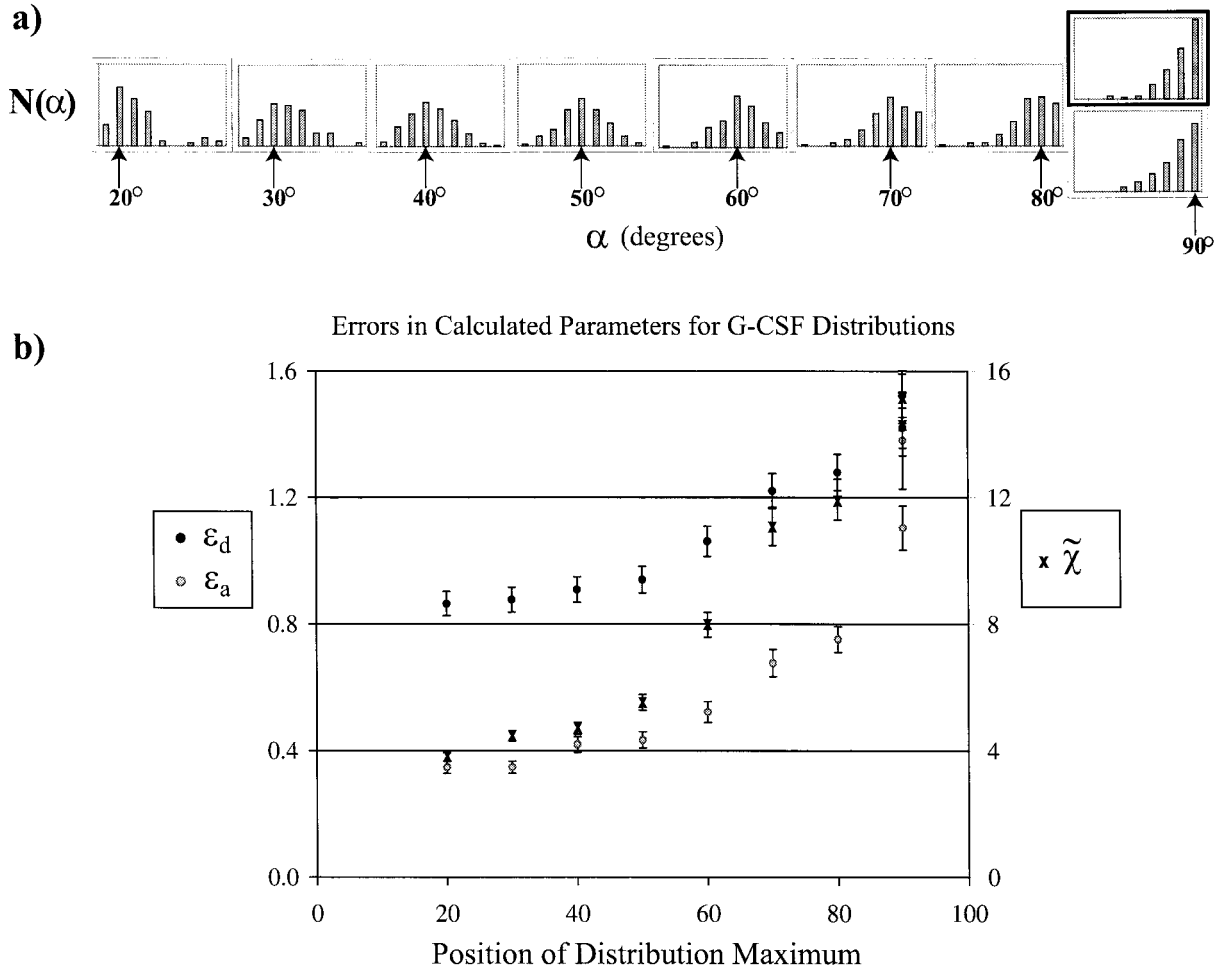

Figure 5. G-CSF bond vector distributions and errors in calculated tumbling parameters. (a) The orientation of the diffusion tensor was rotated with respect to the molecular frame to obtain the N-H bond vector distributions (with respect to the diffusion tensor) shown. Rotations were chosen to generate maximal sampling in each 10 deg bin between 0 and 90 deg. Insufficient surface area is subtended between 0 and 10 deg to obtain maximal sampling in this region. The rotations used to generate the distributions were, from left to right, $\theta = 165°$, $\theta = -30°$, $\theta = 20°$, $\theta = 30°$, $\theta = 45°$, $\theta = 55°$, $\theta = 60°$, $\theta = 75°$ (highlighted), $\theta = 90°$; $\phi = 45°$ for all. The orientation of the diffusion tensor calculated by Lee et al. ($\theta = 77$ deg) closely corresponds to the orientation ($\theta = 75$ deg) used to generated the highlighted distribution with maximal sampling at 90 deg. (b) Statistics computed for the nine distributions. Reported values include the effects of 2.5% measurement error and large ($\tau_f < 600$ ps) approximation error. Reported values are the average over 500 data sets. Error bars are calculated from the standard deviation over 500 data sets. The error measures are small for all orientations considered here. (ε_d and $\varepsilon_a < 1.6\%$, $\tilde{\chi} < 30$, the value for which $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$.)

rience, the number of residues that fit this criterion is relatively small. How much signal can be sacrificed to reduce the noise, and still produce the desire outcome of increased accuracy? To address this question we re-created our synthetic data sets, using values of τ_f restricted to the range $0 < \tau_f < 20$ ps (*vide supra*). The quality of fits for the two types of synthetic data ($\tau_f < 600$ ps, simulating current filtering as applied by NORMAdyn, or $\tau_f < 20$ ps, simulating use of only residues that fit Model 1) are compared in Figure 7b. Fitting Model 1 residues only is unlikely to be helpful for small proteins. For example, for a 50 residue protein, the cost of removing signal outweighs the benefit of reducing the approximation error when as few as ~10 residues are removed. In contrast, this could be a helpful strategy for very large proteins (e.g., 200 residues), in which the benefit of reducing the approximation error outweighs the cost of removing up to ~100 residues.

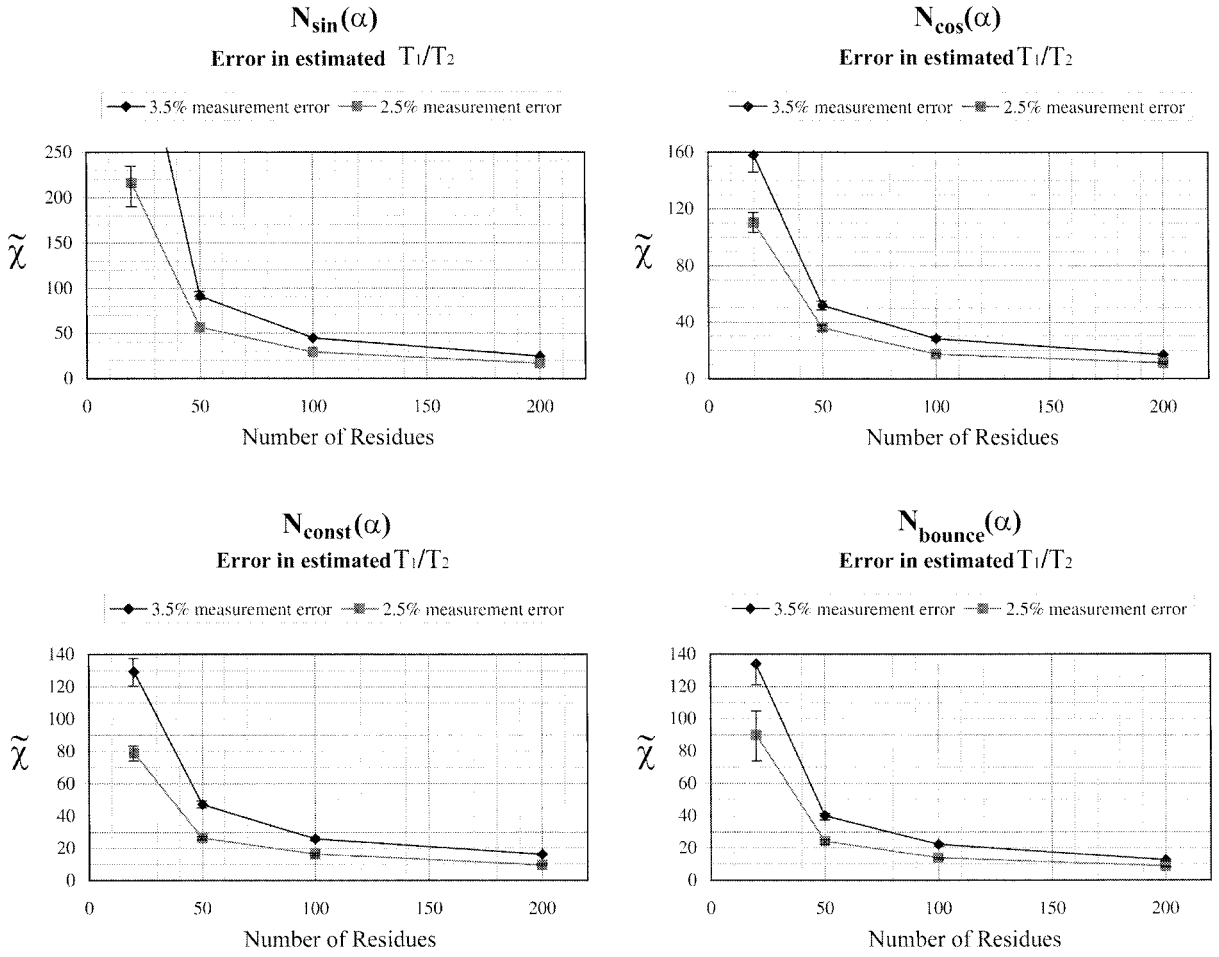

Figure 6. Effects of measurement error on dynamics calculations. The error in estimated T_1/T_2 is calculated using two sets of measurement error for four distributions of N-H bond vectors. The number of data points in each distribution varies from 20 to 200. Reported values include the effects of large (τ_f < 600 ps) approximation error. Reported values are the average over 500 data sets. For all four distributions, a $\sqrt{2}$ decrease in measurement error is equivalent to an increase in data points of ~ 75–90%.

If the experimental data can not uniquely determine the diffusion tensor for a given protein, theoretical options remain. The rotational diffusion tensor can be predicted from the protein structure using hydrodynamic calculations, as implemented in HYDRONMR (Bernado et al., 2002; de la Torre et al., 2000; de la Torre, 2001) or COPED (Osborne and Wright, 2001). These programs have been used to identify residues undergoing ns or μs-ms internal motions, based on a comparison between experimental relaxation data and estimated relaxation data calculated from global tumbling parameters. For proteins with sub-optimal distributions of bond vectors, the global tumbling parameters obtained from these programs could be used to complement analysis of experimental relaxation data, providing initial parameter estimates, constraints, or confirmation of results (Osborne and Wright, 2001).

Changes in the accuracy of calculations upon inclusion of errors in structure

Very recently, Zweckstetter and Bax reported the effects of errors in the structure on the accurate determination of alignment tensors (Zweckstetter and Bax, 2002). Here, we briefly examine the effect of structural error on the accurate determination of diffusion tensors. The four theoretical distributions shown in Figure 1 were employed to examine these effects for bond vector distributions of varying anisotropy. The quality of the fits was evaluated for inclusion of 5° structural error, which is generally a reasonable

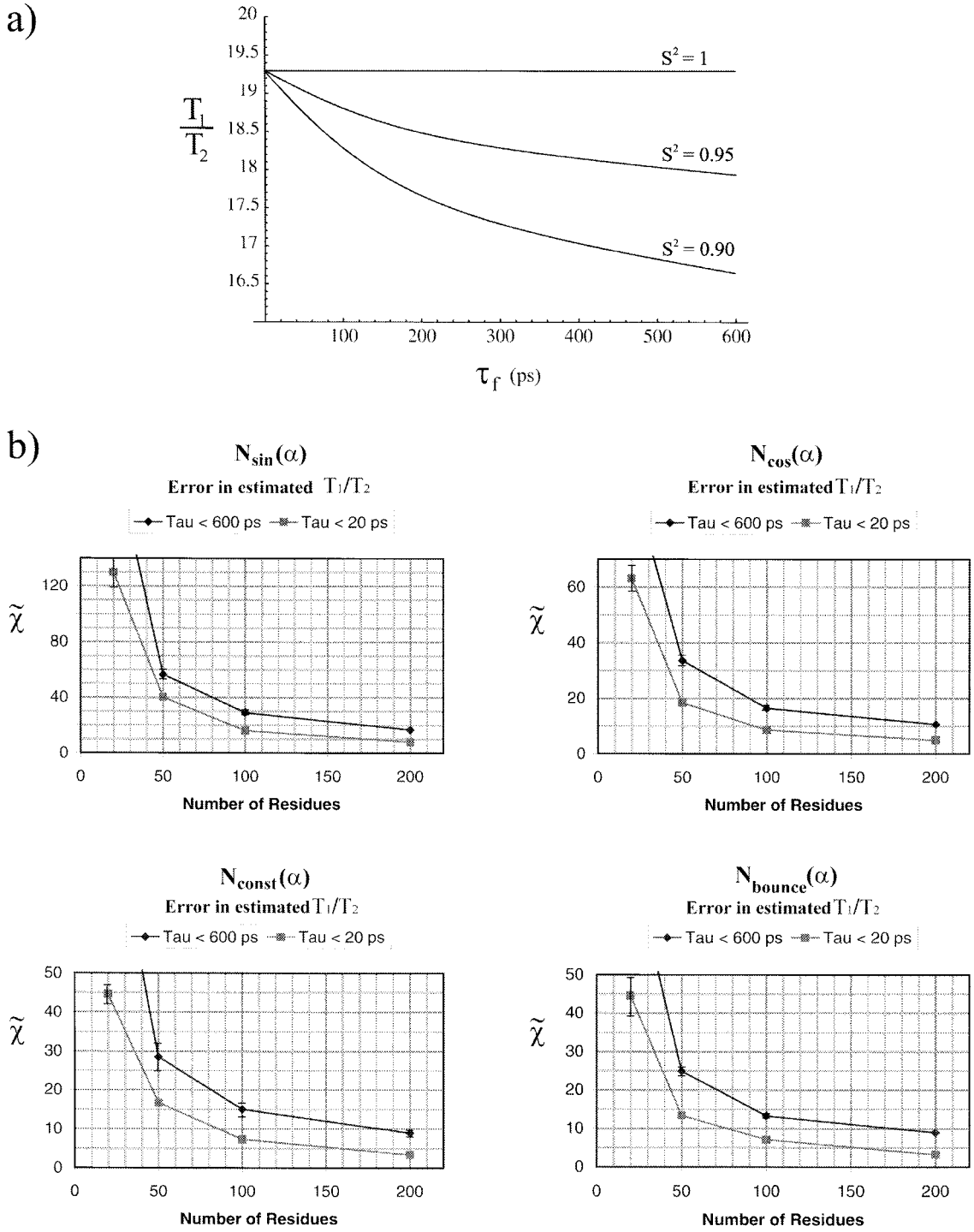

Figure 7. Effects of approximation error on T_1/T_2 values and estimated T_1/T_2 values. (a) The approximation used in fitting is depicted by the $S^2 = 1$ line. The expected deviation from this form varies with the amplitude and time scale of the internal motion, as illustrated by the $S^2 = 0.95$ and $S^2 = 0.9$ curves. Current filtering techniques are expected to remove residues with deviations in T_1/T_2 greater than ~10% from the approximate value (Pawley et al., 2001). (b) The error in estimated T_1/T_2 is calculated for four distributions of N-H bond vectors using two sets of approximation error ($\tau_f < 600$ ps, simulating current filtering as applied by NORMAdyn, or $\tau_f < 20$ ps, simulating use of residues that fit Model 1 only). The number of data points in each distribution varies from 20 to 200. Reported values are the average over 500 data sets, and include the effects of 2.5% measurement error.

representation of experimental errors in bond vector orientations (Zweckstetter and Bax, 2002). The results are shown for each distribution at four values of N_{tot} in Figure 8. Comparison of Figure 8 and Figure 2 shows the changes in fit quality resulting from the inclusion of structural errors. As expected, this additional source of error decreases F, while increasing the error in the principal values of the diffusion tensor, the orientation of the diffusion tensor, and the error in estimated T_1/T_2. In accordance with the results of Zweckstetter and Bax on determination of the alignment tensor, we find that the increase in error in the principal values of the diffusion tensor is directional: the inclusion of structural error results in underestimation of the diffusion tensor (Figure 1, Supplemental Information). While the isotropic distribution ($N_{sin}(\alpha)$, $\Xi = 0.02$) remains the worst overall (Figure 8d, $\tilde{\chi}$), for very large numbers of data points ($N_{tot} > 100$), inclusion of structural error causes a change in the best observed distribution, from $N_{bounce}(\alpha)$ to $N_{const.}(\alpha)$.

The sensitivity of each distribution to the addition of structural error is shown in Figure 9. The error in the principal values of the diffusion tensor (Figure 9a, ε_d) is relatively unaffected by the inclusion of structural error for the $N_{const.}(\alpha)$ distribution, and is most affected by structural error for the $N_{sin}(\alpha)$ distribution. The error in the orientation of the diffusion tensor (Figure 9b, ε_a) is noticeably affected by the inclusion of structural error for all distributions, and the increase in error is largest for the $N_{bounce}(\alpha)$ distribution. The error in estimated T_1/T_2 (Figure 9c, $\tilde{\chi}$) increases most sharply for the $N_{sin}(\alpha)$ distribution, followed by the $N_{bounce}(\alpha)$ distribution. The increase in $\tilde{\chi}$ observed for $N_{const.}(\alpha)$ and $N_{cos}(\alpha)$ upon the addition of structural error is very similar to the increase observed upon change in measurement error from 2.5% to 3.5% (for $N_{const.}(\alpha)$, $N_{tot} = 200$, $\Delta\tilde{\chi} = 6.5$ and

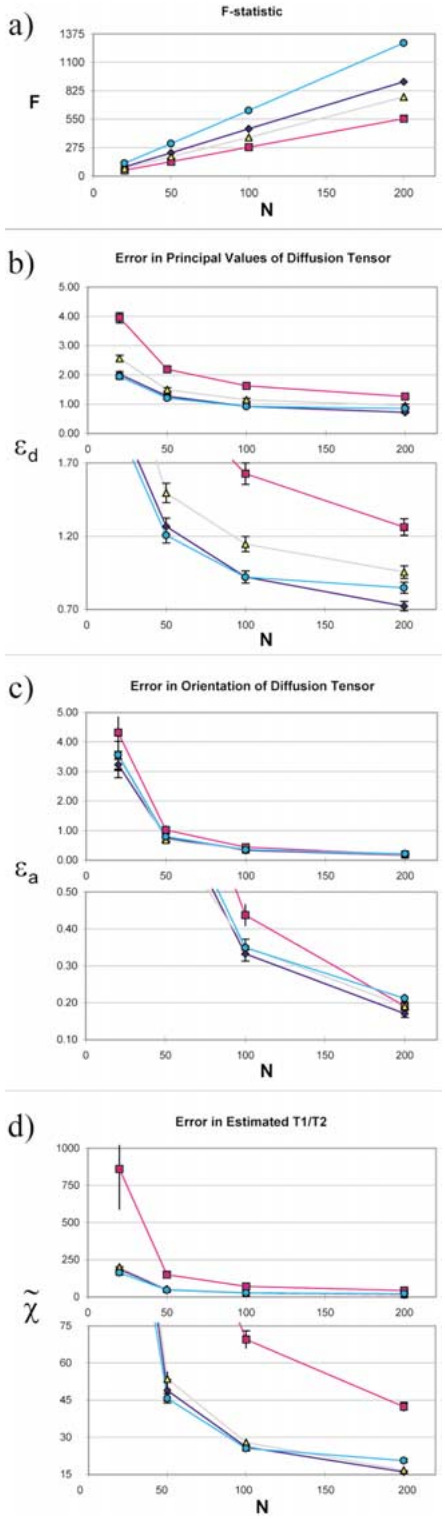

Figure 8. Effects of structural error on accuracy of calculated dynamic parameters. Statistics are computed for four distributions of N-H bond vectors: $N_{sin}(\alpha)$ (pink squares), $N_{cos}(\alpha)$ (yellow triangles), $N_{const.}(\alpha)$ (blue diamonds) and $N_{bounce}(\alpha)$ (cyan circles), as in Figure 2. The number of data points in each distribution varies from 20 to 200. Reported values include the effects of 2.5% measurement error, large ($\tau_f < 600$ ps) approximation error, and 5° structural error. Reported values are the average over 500 data sets. Error bars are calculated from the standard deviation over 500 data sets. (a) The F-statistic (isotropic model vs. axially symmetric anisotropic model). (b) ε_d, the error in the principal values of the diffusion tensor. (c) ε_a, the error in the orientation of the diffusion tensor. (d) $\tilde{\chi}$, the error in the calculated values of T_1/T_2. Note that, in general, $\tilde{\chi} < 63$ for $N_{tot} \gtrsim 50$. For the $N_{sin}(\alpha)$ distribution, $\tilde{\chi} < 63$ for $N_{tot} > 100$.

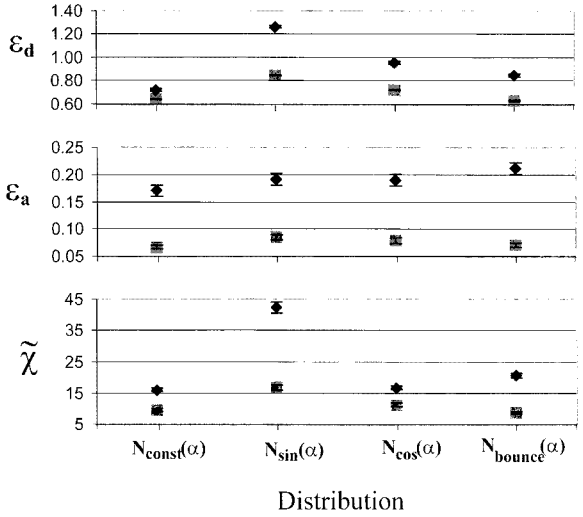

Figure 9. Sensitivity of theoretical distributions to the addition of structural error. Error measures in the absence of structural error (squares), and in the presence of 5° structural error (diamonds), are compared for the four distributions $N_{\text{const.}}(\alpha)$, $N_{\sin}(\alpha)$, $N_{\cos}(\alpha)$, $N_{\text{bounce}}(\alpha)$. Values are shown for $N_{\text{tot}} = 200$. Reported values include the effects of 2.5% measurement error and large ($\tau_f < 600$ ps) approximation error. Reported values are the average over 500 data sets. Error bars are calculated from the standard deviation over 500 data sets. (a) ε_d, the error in the principal values of the diffusion tensor. Values of ε_d increase most sharply for $N_{\sin}(\alpha)$, and increase only marginally for $N_{\text{const.}}(\alpha)$. (b) ε_a, the error in the orientation of the diffusion tensor. Values of ε_a increase most sharply for $N_{\text{bounce}}(\alpha)$. (c) $\tilde{\chi}$, the error in the calculated values of T_1/T_2. Values of $\tilde{\chi}$ increase most sharply for $N_{\sin}(\alpha)$ and $N_{\text{bounce}}(\alpha)$.

6.6, respectively; for $N_{\cos}(\alpha)$, $N_{\text{tot}} = 200$, $\Delta\tilde{\chi} = 5.4$ and 5.7, respectively). In other words, the benefit of decreasing the measurement error by $\sqrt{2}$ is similar in size to the benefit of removing error in the structure for $N_{\text{const.}}(\alpha)$ and $N_{\cos}(\alpha)$, and is significant for $N_{\text{bounce}}(\alpha)$ and $N_{\sin}(\alpha)$.

As noted in Materials and methods section, an error scale for $\tilde{\chi}$ can be set by observing when $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$. For the theoretical distributions, $(\text{calc} - \text{true})^2/(\sigma \cdot \text{true})^2 \approx 1$ when $\tilde{\chi} = 63$. Examination of Figure 8c shows that reasonably accurate estimates of T_1/T_2 can be obtained for $N_{\text{const.}}(\alpha)$, $N_{\text{bounce}}(\alpha)$ and $N_{\cos}(\alpha)$ when $N_{\text{tot}} \gtrsim 50$, and for $N_{\sin}(\alpha)$ when $N_{\text{tot}} > 100$.

Conclusions

The accurate description of protein dynamics and thermodynamics requires an accurate description of global tumbling, since this tumbling dominates relaxation rates in proteins. Based on the assumption that the best accuracy is obtained for proteins in which the distribution of bond vectors used in the relaxation analysis is isotropic in space, it has been estimated that approximately 15% of proteins are likely to yield inaccurate tumbling parameters (Fushman et al., 2000). Accurate determination of the diffusion tensor is expected to be limited to distributions with values of $\Xi < 0.25$.

In contrast, we have demonstrated here that anisotropic distributions of bond vectors do not necessarily yield inaccurate tumbling parameters, and, in some cases, may yield more accurate parameters than the isotropic distribution. In particular, three of the five distributions presented herein correspond to $\Xi > 0.25$. In two of these cases, the distribution is shown to provide a more accurate description of global tumbling than the isotropic distribution ($\Xi = 0.02$). In the case of the G-CSF helical bundle ($\Xi = 0.49$), the error measures (ε_d and ε_a) are less than 2% for all diffusion tensor orientations considered. This demonstrates that the important class of helical bundle proteins can be analyzed accurately. Based on the theoretical distributions examined, similar results are expected for other protein classes with anisotropic distributions of N-H bond vectors.

We have also demonstrated the efficacy of reducing measurement error as an alternative to simultaneous analysis of $^{13}C^{\alpha}$ relaxation data, and have described the trade-offs associated with filtering out approximation error. In all cases examined here, reduction of measurement error provides significant improvement in the calculated parameters, and in some cases improves the estimates as much as the removal of structural error. Reduction of measurement error in the relaxation data can more easily be accomplished than removal of intrinsic errors in the structure. The current pace of improvements in NMR technology, such as development of the cryoprobe, of TROSY-based pulse sequences, and the availability of higher field strengths indicates that the reduction of measurement error will be an increasingly effective avenue for improving the accuracy of global and internal parameters of motion derived from analysis of ^{15}N relaxation measurements.

Acknowledgements

This research was funded by the NSF (MCB-9808727). N.H.P. was supported by the National Physical Science Consortium and the Department of Energy through Lawrence Livermore National Laboratory.

References

Andrec, M., Montelione, G.T. and Levy, R.M. (1999) *J. Magn. Reson.*, **139**, 408–421.

Arfken, G. (1985) *Mathematical Methods for Physicists,* Academic Press, Boston, MA.

Bernado, P., de la Torre, J.G. and Pons, M. (2002) *J. Biomol. NMR*, **23**, 139–150.

Bevington, P.R. and Robinson, D.K. (1992) *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, Boston, MA.

de Alba, E., Baber, J.L. and Tjandra, N. (1999) *J. Am. Chem. Soc.*, **121**, 4282–4283.

de la Torre, J.G. (2001) *Biophys. Chem.*, **94**, 265–274.

de la Torre, J.G., Huertas, M.L. and Carrasco, B. (2000) *J. Magn. Reson.*, **147**, 138–146.

Fushman, D., Ghose, R. and Cowburn, D. (2000) *J. Am. Chem. Soc.*, **122**, 10640–10649.

Hill, C.P., Osslund, T.D. and Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 5167–5171.

Hodgkinson, P. and Emsley, L. (1997) *J. Chem. Phys.*, **107**, 4808–4816.

Jones, J.A., Hodgkinson, P., Barker, A.L. and Hore, P.J. (1996) *J. Magn. Reson. Ser. B*, **113**, 25–34.

Kneller, J.M., Lu, M. and Bracken, C. (2002) *J. Am. Chem. Soc.*, **124**, 1852–1853.

Kroenke, C.D., Loria, J.P., Lee, L.K., Rance, M. and Palmer, A.G. (1998) *J. Am. Chem. Soc.*, **120**, 7905–7915.

Lee, L.K., Rance, M., Chazin, W.J. and Palmer, A.G. (1997) *J. Biomol. NMR*, **9**, 287–298.

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

Mandel, A.M., Akke, M. and Palmer, A.G. (1995) *J. Mol. Biol.*, **246**, 144–163.

Mandel, A.M., Akke, M. and Palmer, A.G. (1996) *Biochemistry*, **35**, 16009–16023.

Osborne, M.J. and Wright, P.E. (2001) *J. Biomol. NMR*, **19**, 209–230.

Pawley, N.H., Wang, C.Y., Koide, S. and Nicholson, L.K. (2001) *J. Biomol. NMR*, **20**, 149–165.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1996) *Numerical Recipes in C: The Art of Scientific Computing*, Press Syndicate of the University of Cambridge, New York, NY.

Prompers, J.J. and Bruschweiler, R. (2000) *J. Phys. Chem. B*, **104**, 11416–11424.

Schurr, J.M., Babcock, H.P. and Fujimoto, B.S. (1994) *J. Magn. Reson. Ser. B*, **105**, 211–224.

Skrynnikov, N.R., Goto, N.K., Yang, D.W., Choy, W.Y., Tolman, J.R., Mueller, G.A. and Kay, L.E. (2000) *J. Mol. Biol.*, **295**, 1265–1273.

Tjandra, N., Feller, S.E., Pastor, R.W. and Bax, A. (1995) *J. Am. Chem. Soc.*, **117**, 12562–12566.

Tjandra, N., Wingfield, P., Stahl, S. and Bax, A. (1996) *J. Biomol. NMR*, **8**, 273–284.

Tugarinov, V., Shapiro, Y.E., Liang, Z.C., Freed, J.H. and Meirovitch, E. (2002) *J. Mol. Biol.*, **315**, 155–170.

Viles, J.H., Duggan, B.M., Zaborowski, E., Schwarzinger, S., Huntley, J.J.A., Kroon, G.J.A., Dyson, H.J. and Wright, P.E. (2001) *J. Biomol. NMR*, **21**, 1–9.

Wand, A.J. (2001) *Nat. Struct. Biol.*, **8**, 926–931.

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

Zweckstetter, M. and Bax, A. (2002) *J. Biomol. NMR*, **23**, 127–137.