https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Bootstrap Aggregation for Model Selection in the Model-free Formalism

Timothy Crawley[1] and Arthur G. Palmer, III[1]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, United States

Correspondence: Arthur G. Palmer, III (agp6@columbia.edu)

Abstract. The ability to make robust inferences about the dynamics of biological macromolecules using NMR spectroscopy depends heavily on the application of appropriate theoretical models for nuclear spin relaxation. Data analysis for NMR laboratory-frame relaxation experiments typically involves selecting one of several model-free spectral density functions using a bias-corrected fitness test. Here, advances in statistical model selection theory, termed bootstrap aggregation or bagging, are applied to ^{15}N spin relaxation data, developing a multimodel inference solution to the model-free selection problem. The approach is illustrated using data sets recorded at four static magnetic fields for the bZip domain of the *S. cerevisiae* transcription factor GCN4.

1 Introduction

Since the original publications in the early 1980's, the model-free formalism of Lipari and Szabo (Lipari and Szabo, 1982a, b) and the related two-step approach of Halle and Wennenström (Halle and Wennerström, 1981) have served as starting points for extracting dynamical information about macromolecules from NMR spin relaxation data. In the ensuing decades, increasingly complex models have offered a more refined understanding of internal and overall molecular motions (Clore et al., 1990; Lemaster, 1995; Tugarinov et al., 2001; Khan et al., 2015; Hsu et al., 2018, 2020; Smith et al., 2019). However, the availability of "extended" model-free models has created a further dilemma: should a data analysis protocol extract the most exacting information justified by the data, or employ the model most robust to experimental variation.

Several authors have addressed the model selection problem by employing the principle of parsimony or Occam's Razor (Palmer et al., 1991; Stone et al., 1992; Mandel et al., 1995; d'Auvergne and Gooley, 2003; Chen et al., 2004). These approaches seek to identify the simplest model that explains the data within experimental uncertainties by applying various bias-correcting penalties to the fitness statistic, e.g. F-statistic, AIC or BIC. However, these corrections alone often fall short of producing robust inferences and may yield parameter values susceptible to instability in both simulated and real-world replicates. In these situations, the model selection process has failed the principle of 'worrying selectively'. This criterion suggests, "Since all models are wrong the scientist must be alert to what is importantly wrong." (Box, 1976).

To illustrate the issue more concretely, a typical data analysis protocol uses a non-linear weighted least-squares algorithm to fit experimental spin relaxation data with a set of model-free spectral density functions (Mandel et al., 1995; Gill et al.,

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

2016). The resulting χ^2 residual sum-of-squares variables are penalized for the number of adjustable parameters in each model function, the model with the lowest penalized residual sum-of-squares is selected as optimal, and the best-fit parameters of the model reported. However, this procedure is subject to model-selection error: random statistical variation in the experimental data may lead to one model chosen as optimal for a given data set, but another model, with different set of parameters, may be selected if the experimental data were replicated, with consequent different random variation. The joint problem of model-selection and parameter estimation has been explored elegantly by d'Auvergne and Gooley (d'Auvergne and Gooley, 2007, 2008a, b).

The present paper addresses model-selection error by using the approach of bootstrap aggregation or bagging. This concept originated from a desire to improve the performance of machine learning algorithms. Thus, Breiman showed that predictor accuracy and stability improved when averaging predictor values obtained from bootstrap replicates of the original training set (Breiman, 1996). Buja and Stuetzle subsequently extended the use of bagging to generalized statistical analysis and showed sampling with and without replacement yield equivalent improvements (Buja and Stuetzle, 2006). The approach and notation of Efron is used in the following (Efron, 2014).

Bootstrap aggregation improves parameter stability; consequently, the resulting variations in model-free parameter values, for example between atomic sites or functional states in a given macromolecule, are more likely to be biologically or chemically meaningful. Although applicable to most model selection situations, bootstrap aggregation exhibits the most pronounced benefits when the data justify two distinct models with similar degrees of certainty.

Bootstrap aggregation for model-free analysis of NMR spin relaxation relaxation rate constants is illustrated by application to amide backbone ^{15}N spin relaxation data that have been recorded at ^{1}H magnetic fields of 600, 700, 800, and 900 MHz for the bZip domain of the *S. cerevisiae* transcription factor GCN4 by Gill and coworkers (Gill et al., 2016).

2 Theory

In the following, the notation used by Efron is rephrased in terms appropriate for NMR spin relaxation data (Efron, 2014). Laboratory-frame nuclear spin relaxation rate constants for backbone ^{15}N spins can be transformed into sets of spectral density function values, $J(\omega)$, (with experimental uncertainties), in which ω is an eigenfrequency of the spin system (Farrow et al., 1995; Gill et al., 2016). Laboratory-frame ^{15}N relaxation rate constants recorded at a single static magnetic field yield estimates of $J(0)$, $J(\omega_N)$, and $J(0.87\omega_H)$, in which ω_N and ω_H are the ^{15}N and ^{1}H Larmor frequencies. Thus, the number of spectral density values $N = 3G$ in which G is the number of static magnetic fields utilized. In the present application, $G = 4$. The set of experimental spectral densities are:

$$\mathbf{y} = \{y_j\} = \{y_1, y_2, ..., y_N\} \tag{1}$$

in which the values are ordered in increasing values of ω. The experimental data sets utilized in the present work are not affected by chemical exchange contributions to spin relaxation, but in general, such contributions can be taken into account from the field dependence of transverse relaxation rate constants prior to the model-free analysis (Kroenke et al., 1998).

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

The model-free spectral density function used to fit ^{15}N data is:

$$J(\omega) = \frac{2}{5}[S_f^2 S_s^2 \tau_m/(1+\omega^2\tau_m^2) + S_f^2(1-S_s^2)\tau_1/(1+\omega^2\tau_1^2) + (1-S_f^2)S_s^2\tau_2/(1+\omega^2\tau_2^2) + (1-S_f^2)(1-S_s^2)\tau_3/(1+\omega^2\tau_3^2)] \quad (2)$$

in which $1/\tau_1 = 1/\tau_m + 1/\tau_s$, $1/\tau_2 = 1/\tau_m + 1/\tau_f$, $1/\tau_3 = 1/\tau_m + 1/\tau_s + 1/\tau_f$, and $\tau_f < \tau_s$. The set of possible model parameters in this function are:

$$\mu = \{\mu_k\} = \{\tau_m, S_f^2, S_s^2, \tau_f, \tau_s\} \quad (3)$$

in which τ_m is the (effective) overall rotational correlation time, S_f^2 is the square of the generalized order parameter for internal motions on a fast ($\tau_f \leq 150$ ps) time scale, and S_s^2 is the square of the generalized order parameter for internal motions on a slow ($\tau_s > 150$ ps) time scale (vide infra). The square of the generalized order parameter $S^2 = S_f^2 S_s^2$. Overall rotational diffusion has been assumed to be isotropic for simplicity; this assumption can be relaxed as needed (Lee et al., 1997). The spectral density data are fit with a set of nested models. The full model, 'Model 5', contains all five parameters, while simpler models, Models 1-4, are generated by fixing the value of one or more parameters, effectively removing such parameters from the model. Thus:

Model 1: $\mu = \{\tau_m, S_f^2, 1, 0, 0\}$
Model 2: $\mu = \{\tau_m, S_f^2, 1, \tau_f, 0\}$
Model 3: $\mu = \{\tau_m, 1, S_s^2, 0, \tau_s\}$
Model 4: $\mu = \{\tau_m, S_f^2, S_s^2, 0, \tau_s\}$
Model 5: $\mu = \{\tau_m, S_f^2, S_s^2, \tau_f, \tau_s\}$

Then:

$$\hat{\mu} = \{\hat{\mu}_k\} = t_m(\mathbf{y}) \quad (4)$$

represents the optimal model t_m and associated parameter values μ obtained using the lowest penalized residual sum-of-squares as described above. In the present work, the small-sample AIC_c criterion was used for model selection (Hurvich and Tsai, 1989).

In general, a non-parametric bootstrap sample:

$$\mathbf{y}_i^* = \{y_{ij}^*\} = \{y_{i1}^*, y_{i2}^*, ..., y_{iN}^*\} \quad (5)$$

in which $i = 1, ...B$ and B is the total number of bootstrap samples, is generated by draws with replacement from the original data $\mathbf{y}$. The nature of spectral density data requires care in generating bootstrap samples and the particular procedure employed in the present work is described in Methods.

A conventional bootstrap determination of the standard deviations of the paramaters $\hat{\mu}$ begins by determining fitted parameters for the ith bootstrap sample as:

$$\hat{\mu}_i^* = \{\hat{\mu}_{ik}^*\} = t_m(\mathbf{y}_i^*) \quad (6)$$

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021


Open Access MAGNETIC RESONANCE Discussions

in which the fitting model is fixed to the optimal model selected in fitting the original spectral density values and only model parameter values are optimized. The bootstrap estimate of the standard deviation for the kth parameter is given by:

$$\hat{\sigma}_k^* = \left[\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\mu}_{ik}^* - \hat{\mu}_k^*)^2 \right]^{1/2} \tag{7}$$

in which,

$$\hat{\mu}_k^* = \frac{1}{B} \sum_{i=1}^{B} \hat{\mu}_{ik}^* \tag{8}$$

In the conventional approach, the reported results of the data analysis would be $\{\hat{\mu}_k\}$ and $\{\hat{\sigma}_k^*\}$. Model-selection error is not assessed.

In contrast to the conventional procedure, bootstrap aggregation determines both the optimal fitted model and associated model parameters for each bootstrap sample. Thus, for the ith bootstrap sample:

$$\tilde{\mu}_i^* = \{\tilde{\mu}_{ik}^*\} = t_i(\mathbf{y}_i^*) \tag{9}$$

in which the optimal model t_i is determined for the ith bootstrap sample using the same model selection strategy as for the original data. Unlike the conventional bootstrap procedure, the different members of the set $\tilde{\mu}_i^*$ obtained by bootstrap aggregation may represent different models as well as different sets of optimized parameters. The aggregated, or smoothed, estimator of the kth model parameter is given by:

$$\tilde{\mu}_k = \frac{1}{B} \sum_{i=1}^{B} \tilde{\mu}_{ik}^* \tag{10}$$

To make the above formalism concrete, suppose that for a given set of spectral density values, model selection and parameter optimization for B bootstrap samples yields B_2 samples in which model 2 is optimal and B_3 samples in which model 3 is optimal, with $B = B_2 + B_3$. Then,

$$\tilde{S}_f^2 = \frac{1}{B} \left[\sum_{i \in B_2} \tilde{S}_{fi}^{2*} + \sum_{i \in B_3} 1 \right] \tag{11}$$

and

$$\tilde{\tau}_f = \frac{1}{B} \left[\sum_{i \in B_2} \tilde{\tau}_{fi}^* + \sum_{i \in B_3} 0 \right] \tag{12}$$

As another example, suppose that for a given set of spectral density values, model selection and parameter optimization for B bootstrap samples yields B_4 samples in which model 4 is optimal and B_5 samples in which model 5 is optimal, with $B = B_4 + B_5$. Then,

$$\tilde{S}_f^2 = \frac{1}{B} \sum_{i=1}^{B} \tilde{S}_{fi}^{2*} \tag{13}$$

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021


Open Access MAGNETIC RESONANCE Discussions

because both models 4 and 5 fit S_f^2 as a parameter, but

$$\tilde{\tau}_f^2 = \frac{1}{B}\left[\sum_{i\in B_4} 0 + \sum_{i\in B_5} \tilde{\tau}_{fi}^{2*}\right] \tag{14}$$

Additionally, a smoothed standard deviation for $\tilde{\mu}$ can be obtained using the plug-in-principle (Efron, 2014). Here, the cumulative distribution function for the parameters of interest are estimated using the empirical distribution function of the bootstrap replicates. Using the above notation,

$$Y_{ij}^* = \#\{y_{ik}^* = y_j\} \tag{15}$$

is defined as the number of times that the ith bootstrap replicate, $\mathbf{y}_i^*$, contains the spectral density value y_j and

$$\mathbf{Y}_i^* = \{Y_{i1}^*, Y_{i2}^*, ..., Y_{iN}^*\} \tag{16}$$

is defined as a vector enumerating the representation of each original data point in the ith bootstrap replicate. Further, the average representation of original spectral density value y_j across the B bootstrap replicates is:

$$\bar{Y}_j^* = \frac{1}{B}\sum_{i=1}^{B} Y_{ij}^* \tag{17}$$

The covariance between the representation of the jth spectral density value and the kth model-free parameter value across B bootstrap replicates is:

$$\hat{cov}_{jk} = \frac{1}{B}\sum_{i=1}^{B}\left(Y_{ij}^* - \bar{Y}_j^*\right)\left(\tilde{\mu}_{ik}^* - \tilde{\mu}_k\right) \tag{18}$$

Finally, the smoothed estimate of the standard deviation for the kth model-free parameter is calculated as:

$$\tilde{\sigma}_k = \left[\frac{1}{N}\sum_{j=1}^{N} \hat{cov}_{jk}^2\right]^{1/2} \tag{19}$$

In bootstrap aggregation, the reported results consist of the smoothed estimators $\{\tilde{\mu}_k\}$ and $\{\tilde{\sigma}_k\}$ incorporating the effects of model-selection uncertainty. As noted by Efron, $\tilde{\sigma}_k \leq \hat{\sigma}_k^u$, in which $\hat{\sigma}_k^u$ is obtained using Eq. 7 naively applied to the bootstrap aggregated data (rather than to data analyzed with a fixed model as above) (Efron, 2014).

3 Methods

Backbone amide ^{15}N spin relaxation data have been recorded at $G = 4$ ^{1}H static magnetic fields of 600, 700, 800, and 900 MHz for the bZip domain of the *S. cerevisiae* transcription factor GCN4 (Gill et al., 2016). Experimental values of R_1, R_2, and the steady-state NOE measured at each magnetic field for each residue were converted to spectral density values as follows

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021


Open Access MAGNETIC RESONANCE Discussions

(Farrow et al., 1995; Gill et al., 2016):

$$J(0.87\omega_H) = \frac{4}{5d_{NH}^2}\sigma_{NH} \tag{20}$$

$$J(\omega_N) = \frac{7}{20}\left(\frac{0.87}{0.921}\right)^2 \sigma_{NH}/(3 * d_{NH}^2 + 4c_{NH}^2) \tag{21}$$

$$J(0) = \frac{6}{3d_{NH}^2 + 4c_{NH}^2}\Gamma \tag{22}$$

in which:

$$\sigma_{NH} = (NOE - 1) * R_1 * (\gamma_N/\gamma_H) \tag{23}$$

$$\Gamma = R_2 - 0.5R_1 - 0.454(\sigma_{NH}) \tag{24}$$

$d_{NH} = (\mu_0/4\pi)\hbar\gamma_H\gamma_N/r_{NH}^3$, $c_{NH} = \omega_N\Delta\sigma/3^{1/2}$, $r_{NH} = 0.102$ nm is the N-H bond length, and $\Delta\sigma = -172$ ppm is the ^{15}N chemical shift anisotropy. For each residue a single value of $J(0)$ was obtained as the weighted mean (using propagated experimental uncertainties) of the G values obtained at each magnetic field. The uncertainty in the mean $J(0)$ was obtained by jackknife simulations. For each residue, the spectral density values $\mathbf{y}$ used for model fitting consist of the mean $J(0)$, G values of $J(\omega_N)$ and G values of $J(0.87\omega_H)$, for a total of $N = 2G + 1 = 9$ data points.

As noted above, the ^{15}N spectral density values for each backbone amide consist of $G = 4$ values of each of $J(0)$, $J(\omega_N)$ and $J(0.87\omega_H)$. Random sampling with replacement from the $N = 12$ values to generate bootstrap samples, as normally applied, could result in samples in which the relative numbers of spectral density values from each class were highly skewed. For example, a bootstrap sample could be generated without any $J(0)$ values, leading to very anomalous fitted parameters. At the other extreme, random sampling with replacement could result in samples in which a single value was highly over-represented. For example, a bootstrap sample could be generated in which one particular $J(0)$ value is represented exclusively.

To avoid such highly unrepresentive possibilities, bootstrap samples were generated by enumerating the $19^3 = 6859$ possible arrangements in which at most two spectral density values from each set of $J(0)$, $J(\omega_N)$ and $J(0.87\omega_H)$ were duplicated. The 19 possible arrangements of the $G = 4$ indices $\{1,2,3,4\}$ and corresponding Y_{ij} for selecting bootstrap samples of $J(0)$, $J(\omega_N)$ and $J(0.87\omega_H)$ are shown in Table 1. In this Table, p_{ij} is a pointer selecting data from a particular set of spectral density values. For example $p_{4j} = [1,2,3,1]$; applying this pointer to the set of $J(0)$ values would select the $J(0)$ values obtained at 600 (×2), 700, and 800 MHz. The corresponding $Y_{4j} = [2,1,1,0]$ is the numbers of times $J(0)$ values recorded at the different fields were sampled. The process would be repeated for the other sets of spectral density values. For example, one particular bootstrap sample might use p_{4j} to select $J(0)$, p_{10j} to select $J(\omega_N)$, and p_{6j} to select $J(0.87\omega_H)$. The full vector Y_{ij} of length $N = 12$ is obtained by concatenating the individual Y_{4j}, Y_{10j}, and Y_{6j} vectors from the table.

The data were analyzed by three procedures. First, a conventional analysis was performed in which optimal models t_m and model parameters $\{\hat{\mu}_k\}$ were determined for each amino acid residue (for which data were available) using AICc. The uncertainties in model parameters, denoted $\{\hat{\sigma}_k\}$, were determined by 500 Monte Carlo simulations using the measured experimental uncertainties in the spectral density values (Gill et al., 2016). Second, the optimal model was determined as in the first procedure, but the uncertainties in model parameters, $\{\hat{\sigma}_k^*\}$, were determined by the conventional bootstrap, using Eq. (7).

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Table 1. Bootstrap Selections

i	p_{ij}	Y_{ij}	i	p_{ij}	Y_{ij}
1	[1,2,3,4]	[1,1,1,1]	11	[4,2,3,4]	[0,1,1,2]
2	[1,1,3,4]	[2,0,1,1]	12	[1,4,3,4]	[1,0,1,2]
3	[1,2,1,4]	[2,1,0,1]	13	[1,2,4,4]	[1,1,0,2]
4	[1,2,3,1]	[2,1,1,0]	14	[1,1,2,2]	[2,2,0,0]
5	[2,2,3,4]	[0,2,1,1]	15	[1,1,3,3]	[2,0,2,0]
6	[1,2,2,4]	[1,2,0,1]	16	[1,1,4,4]	[2,0,0,2]
7	[1,2,3,2]	[1,2,1,0]	17	[2,2,3,3]	[0,2,2,0]
8	[3,2,3,4]	[0,1,2,1]	18	[2,2,4,4]	[0,2,0,2]
9	[1,3,3,4]	[1,0,2,1]	19	[3,3,4,4]	[0,0,2,2]
10	[1,2,3,3]	[1,1,2,0]			

In both of these approaches, error estimates were obtained while fixing the model as the optimal model selected against the original data. Third, the smoothed model parameters $\{\tilde{\mu}_k\}$ and uncertainties $\{\tilde{\sigma}_k\}$ were determined by bootstrap aggregation using Eqs. (10) and (19), respectively. In this approach, the optimal model was determined individually for each bootstrap sample.

4 Results

The results of the conventional analysis using AICc for model-selection and Monte Carlo error estimation are shown in Figure 1. Each of the Monte Carlo simulations was analyzed using the optimal model determined from the original data. The optimal fitted parameters differ slightly from those reported by Gill and coworkers because AICc was used as the model-selection protocol, rather than AIC (Gill et al., 2016). The results of the conventional analysis using AICc for model-selection and bootstrap resampling for error estimation are shown in Figure 2. Each of the bootstrap data sets was analyzed using the optimal model determined from the original data. The results for bootstrap aggregation are shown in Figure 3 using AICc to determine the optimal model for each bootstrap sample. The smoothed model-free parameters were calculated using Eq. (10) and the smoothed parameter uncertainties were calculated using Eq. (19).

Bootstrap simulations in which a single optimal model is utilized is an alternative to Monte Carlo simulations for unsmoothed parameter estimation. The uncertainties in S^2 obtained by these two approaches are compared in Figure 4a. The uncertainties have approximately the same range, but are uncorrelated with each other. These results suggest the non-parametric bootstrap samples simulate the actual data distribution in comparable manner as the parametric Monte Carlo simulations, but without assuming a normal distribution of spectral density values. The smoothed parameter uncertainty obtained from Eq. (19) is compared to the uncertainties from Monte Carlo simulations in Fig. 4b. The increase in $\tilde{\sigma}(S^2)$ compared to $\hat{\sigma}(S^2)$ reflects the effect of model-selection uncertainty. As noted by Efron, the estimate of smoothed parameter uncertainty obtained from Eq.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

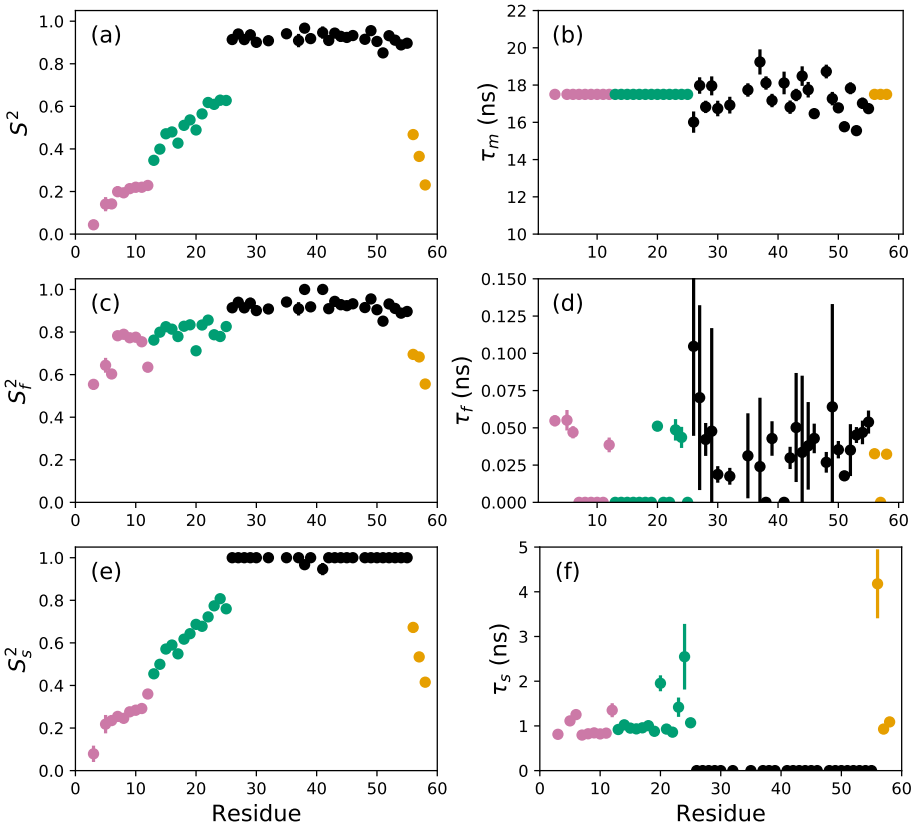

Figure 1. Model-free parameters from conventional model selection using AICc and 500 Monte Carlo simulations to determine parameter uncertainties. Values of S^2, τ_m, S_f^2, τ_f, S_s^2, and τ_s are plotted vs. residue number. Overall correlation times (τ_m) were determined individually for residues in the coiled-coil region (black), while τ_m was fixed at 17.5 ns for residues in the basic region and C-terminal residues, as denoted by the horizontal line. Regions of the protein are colored as (pink) basic region 1 (residues 3–12), (green) basic region 2 (residues 13–25), (black) coiled-coil (residues 26–55), and (orange) disordered C-terminus (residues 56–58) (Gill et al., 2016).

(19) is smaller than the naive estimate obtained by applying Eq. (6) to the aggregated bootstrap samples (Efron, 2014). To illustrate the advantage of Eq. (19), Fig. 4c compares $\hat{\sigma}^u(S^2)$ obtained from Eq. (6) and $\tilde{\sigma}(S^2)$ obtained from Eq.(19). Similar behavior is observed for other model-free parameters (not shown).

The performance of the conventional analysis, in which a single optimal model is chosen, and bootstrap aggregation, in which parameter values are smoothed over all models, are illustrated for particular residues Arg 11, Arg 26, and Asp 32. Tables 2, 3, and 4 show the values of AICc for each model fit to the original spectral density and the percentage that each model was chosen in the bootstrap aggregation. The conventional analysis chooses the optimal model as that model with the smallest AICc. Note that model selection using AICc selects model 2 if $\tau_f \leq 0.15$ ns and model 3 if $\tau_f > 0.15$ ns (vide infra). Tables 5, 6, and 7 show the optimized model-free parameters for each model fit to the original spectral density data and the smoothed

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

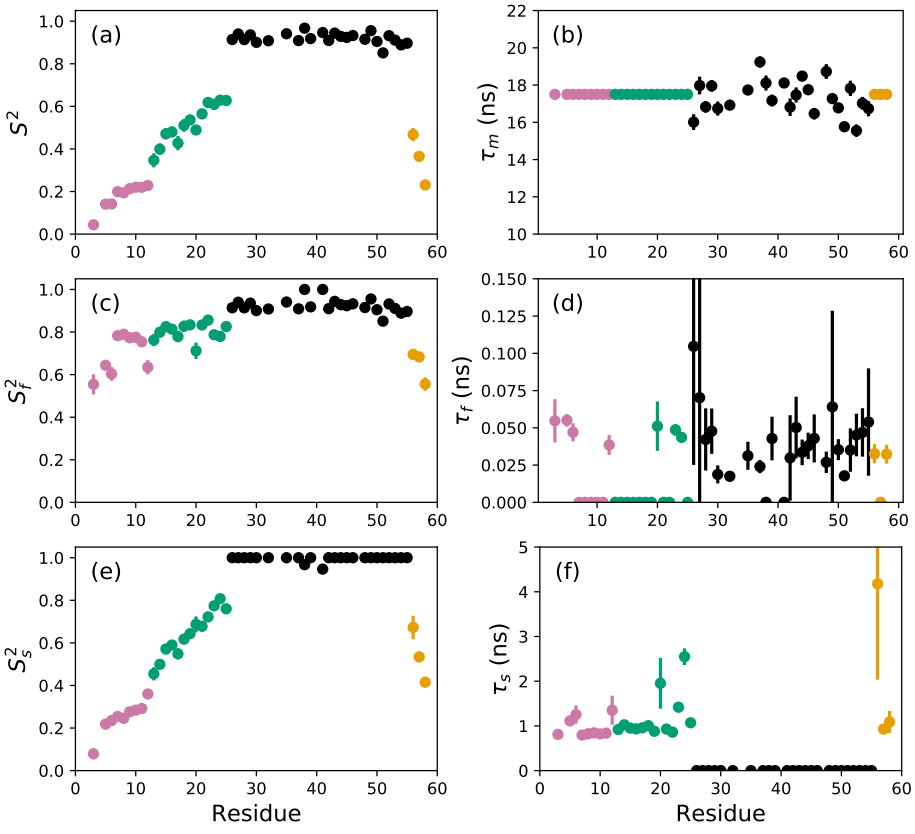

Figure 2. Model-free parameters from conventional model selection using AICc and bootstrap resampling to determine parameter uncertainties. Values of S^2, τ_m, S_f^2, τ_f, S_s^2, and τ_s are plotted vs. residue number. Parameter values are identical as in Figure (1), but the uncertainty estimates differ.

model-free parameters obtained by bootstrap aggregation. The optimal single model selected by AICc is highligthed with an asterisk.

To further illustrate bootstrap aggregation for Arg 11, Arg 26, and Asp 32, Figures 5, 6, and 7 show the distributions of model-free parameters determined from the optimal model for each bootstrap sample. The calculated spectral density function for bootstrap aggregation is compared to the fitted spectral density functions for each model in Figures 8, 9, and 10.

5 Discussion

The difficulties posed by conventional model-selection strategies, in which a single optimal model is chosen using AICc or other fitness statistic, are illustrated for the bZip domain of GCN4 in Fig. 1. In particular, some residues in the basic region (residues 3-25) are analyzed using Model 4, in which $\tau_f = 0$ and other residues are analyzed with model 5, in which $\tau_f > 0$. The resulting values of the other model-free parameters are systematically affected depending on whether or not $\tau_f = 0$. These

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

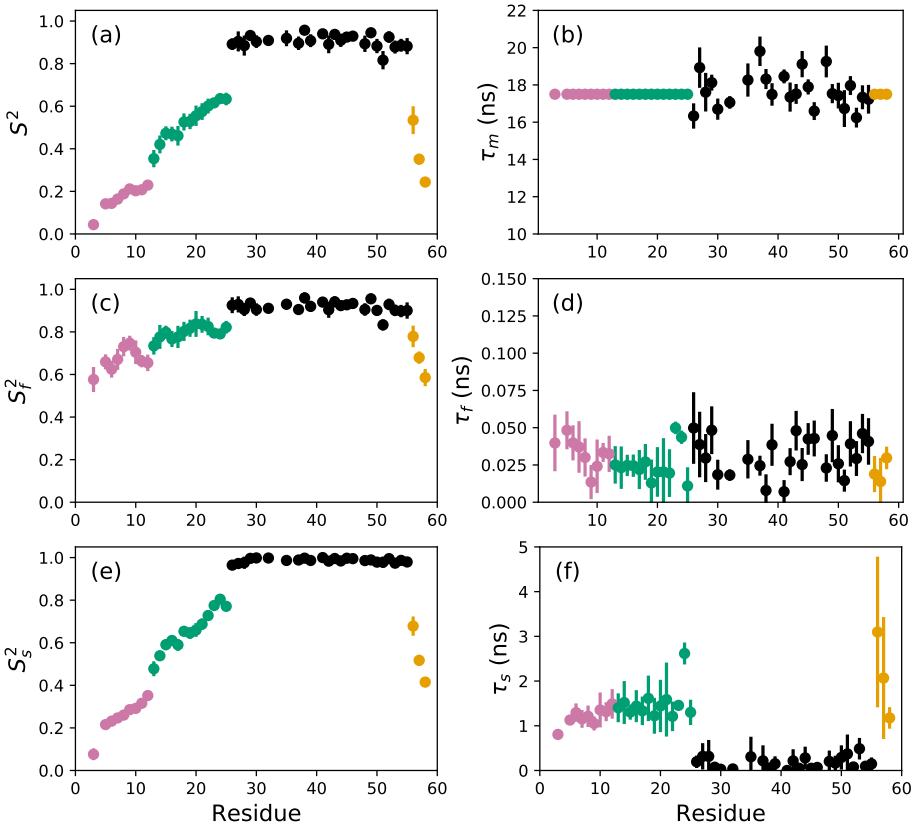

Figure 3. Model-free parameters from bootstrap aggregation to determined smoothed parameter estimates and uncertainties. Values of S^2, τ_m, S_f^2, τ_f, S_s^2, and τ_s are plotted vs. residue number.

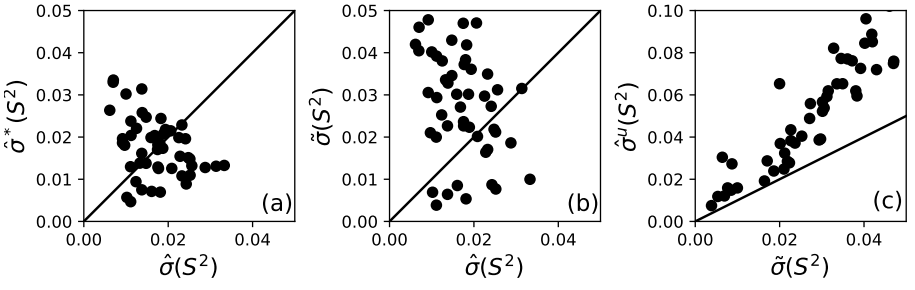

Figure 4. Comparison of model-free parameter uncertainties. (a) Uncertainties for S^2 calculated from Monte Carlo, $\hat{\sigma}_k$ and bootstrap simulations, $\hat{\sigma}_k^*$, for a single optimal model. (b) Uncertainties for S^2 calculated from Monte Carlo simulations for a single optimal model and smoothed $\tilde{\sigma}_k$ calculated from bootstrap aggregation. (c) Uncertainties $\hat{\sigma}_k^u$ and $\tilde{\sigma}$ for S^2 calculated from bootstrap aggregation, illustrating the smaller variability obtained using Eq. (19) for calculation of parameter sample deviations.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Table 2. Model Selection for Residue Arg 11

Fit	Model 1	Model 2	Model 3	Model 4	Model 5
AICc	67.9	NA	57.2	33.3	34.2
Smooth	0.000	0.000	0.000	0.243	0.757

Table 3. Model Selection for Residue Arg 26

Fit	Model 1	Model 2	Model 3	Model 4	Model 5
AICc	39.2	23.4	NA	33.5	56.6
Smooth	0.000	0.566	0.316	0.096	0.022

Table 4. Model Selection for Residue Asp 32

Fit	Model 1	Model 2	Model 3	Model 4	Model 5
AICc	18.4	10.3	NA	22.3	46.2
Smooth	0.000	0.970	0.000	0.019	0.011

systematic effects are evident most clearly in the scatter in S_f^2 and τ_s for residues in the basic region. The advantages of bootstrap aggregation in smoothing over variablity in model selection is evident in Fig. 3. The residue-to-residue variability of the model-free parameters is reduced. Thus, the distributions of τ_f and τ_s are much more uniform within the four regions of the protein, suggesting rather uniform time-scale processes in each sub-domain.

The results shown for residue Arg 11 in Tables 2 and 5 and Figs. 5 and 8 illustrate the mechanics behind bootstrap aggregation. The original optimization against the measured data yielded AICc values of 33.3 for model 4 and 34.1 for model 5. The conventional analysis then selects model 4 (with $\tau_f = 0$) as optimal, even though AICc for model 5 is only slightly larger. In contrast the bootstrap analysis suggests that model 4 would be optimal for 24% and model 5 would be optimal for 76% of randomly chosen data; under the assumption that the bootstrap samples represent the underlying distribution of spectral density values. Bootstrap smoothing then averages each model-free parameter over the empirical distributions shown in Fig. 5, with resulting optimized spectral density curves compared to the original experimental data in Fig. 8. The results for model 4 in Table 5 and the corresponding vertical orange line in Fig. 8 shows that the selection of model 4 in the conventional analysis results in an estimate for τ_s that is skewed toward the lower boundary of the τ_s bootstrap distribution.

The results shown for residue Arg 26 in Tables 3 and 6 and Figs. 6 and 9 illustrate another advantage of bootstrap aggregation. In this case, the original optimization against the measured data yielded an AICc value 23.4 of for model 2, substantially smaller than for any other model, implying a single model might be an adequate description for this residue. However, the bootstrap distribution for the internal correlation times is bimodal. The conventional choice of model 2 reults in an estimate

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Table 5. Model-free Parameters for Residue Arg 11

Model	S^2	S_f^2	S_s^2	τ_f	τ_s
1	0.886 ± 0.015	0.886 ± 0.015	1	0	0
3	0.480 ± 0.006	1	0.480 ± 0.006	0	0.761 ± 0.011
4*	0.220 ± 0.017	0.754 ± 0.015	0.292 ± 0.018	0	0.838 ± 0.014
5	0.211 ± 0.017	0.646 ± 0.022	0.326 ± 0.020	0.036 ± 0.004	1.13 ± 0.09
Smooth	0.208 ± 0.005	0.662 ± 0.029	0.316 ± 0.013	0.033 ± 0.006	1.31 ± 0.21

* indicates the model selected by AICc.

Table 6. Model-free Parameters for Residue Arg 26

Model	τ_m	S^2	S_f^2	S_s^2	τ_f	τ_s
1	14.55 ± 0.48	0.954 ± 0.031	0.954 ± 0.031	1	0	0
2*	16.01 ± 0.55	0.914 ± 0.024	0.914 ± 0.024	1	0.105 ± 0.054	0
4	16.00 ± 0.73	0.878 ± 0.038	0.935 ± 0.037	0.939 ± 0.013	0	0.274 ± 0.165
5	17.28 ± 2.69	0.812 ± 0.103	0.871 ± 0.070	0.932 ± 0.057	0.030 ± 0.020	0.93 ± 1.15
Smooth	16.33 ± 0.68	0.891 ± 0.027	0.925 ± 0.037	0.972 ± 0.015	0.050 ± 0.024	0.19 ± 0.14

* indicates the model selected by AICc.

Table 7. Model-free Parameters for Residue Asp 32

Model	τ_m	S^2	S_f^2	S_s^2	τ_f	τ_s
1	16.28 ± 0.39	0.944 ± 0.022	0.944 ± 0.022	1	0	0
2*	16.92 ± 0.46	0.908 ± 0.025	0.908 ± 0.025	1	0.017 +/- 0.016	0
4	16.92 ± 1.31	0.908 ± 0.053	1.000 ± 0.060	0.908 ± 0.040	0	0.02 ± 0.48
5	19.58 ± 3.45	0.756 ± 0.138	0.853 ± 0.074	0.887 ± 0.091	0.010 ± 0.009	8.33 ± 3.18
Smooth	17.06 ± 0.34	0.909 ± 0.017	0.911 ± 0.015	0.998 ± 0.005	0.018 ± 0.004	0.035 ± 0.074

* indicates the model selected by AICc.

of τ_f roughly centered in the distrubution, but the smoothed bootstrap estimates identify the presence of two separable time scales for internal motions, one with a mean 0.052 ± 0.019 and the other with mean 0.13 ± 0.08. Residue 26 is at the juncture between the basic region and coiled-coil motif of the GCN4 bZip domain; consequently, the latter effective internal correlation time might represent a vestige of the more pronounced motions evident in the basic region. The critical value of 0.15 ns used to separate fast from slow motions in the present work was chosen empirically to distinguish the two distributions observed for residue 26 (and used for all other residues). More sophisticated clustering algorithms could be used to make this distinction between models 2 and 3.

The results shown for residue Asp 32 in Tables 4 and 7 and Figs. 7 and 10 illustrate a case of strong agreement between the conventional analysis and bootstrap aggregation when a single motional model is strongly favored by the experimental data.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

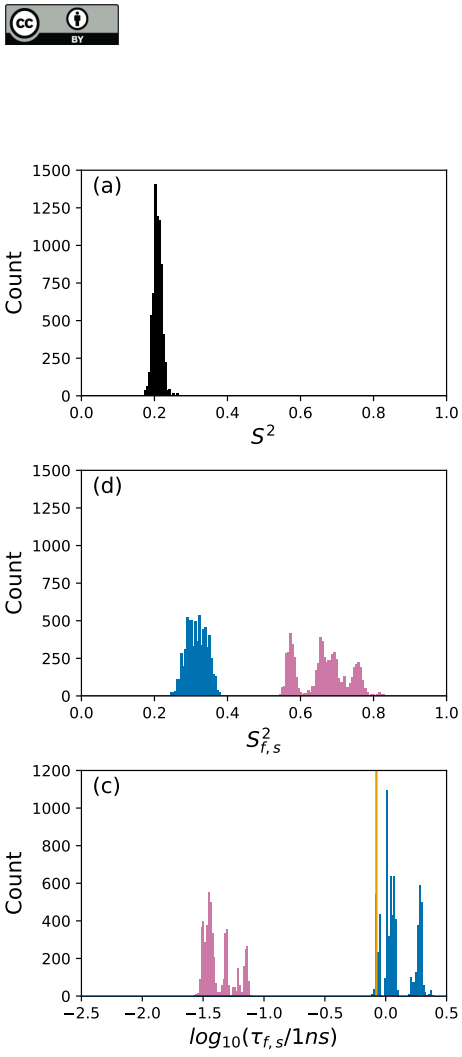

Figure 5. Distribution of model-free parameters from bootstrap aggregation for residue Arg 11. Color coding is (pink) S_f^2 or τ_f and (blue) S_s^2 or τ_s. The orange line in (c) indicates the value of τ_s obtained for the optimal single (unsmoothed) model 4. For clarity, null values of 1 for generalized order parameters and 0 for internal effective correlation times are not shown in the graphs; $\tau_f = 0$ is observed 1664 times.

The distributions shown in Fig. 7 then represent the variability in model-free parameters across the bootstrap samples. These results would be comparable to results obtained in Fig. 2, in which the bootstap samples were used to estimate model-free parameter uncertainties $\hat{\sigma}_k^*$ for a single fixed optimal model.

6 Conclusions

Model-selection error is a classical problem in statistics and has been recognized as a concern in the model-free analysis of NMR spin relaxation data since the work of d'Auvergne and Gooley (d'Auvergne and Gooley, 2007, 2008a, b). Bootstrap aggregation has emerged as a powerful approach for incorporating selection error into statistical model-building (Buja and Stuetzle, 2006; Efron, 2014). However, bootstrap aggregation requires sufficient numbers of data points to allow faithful re-

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

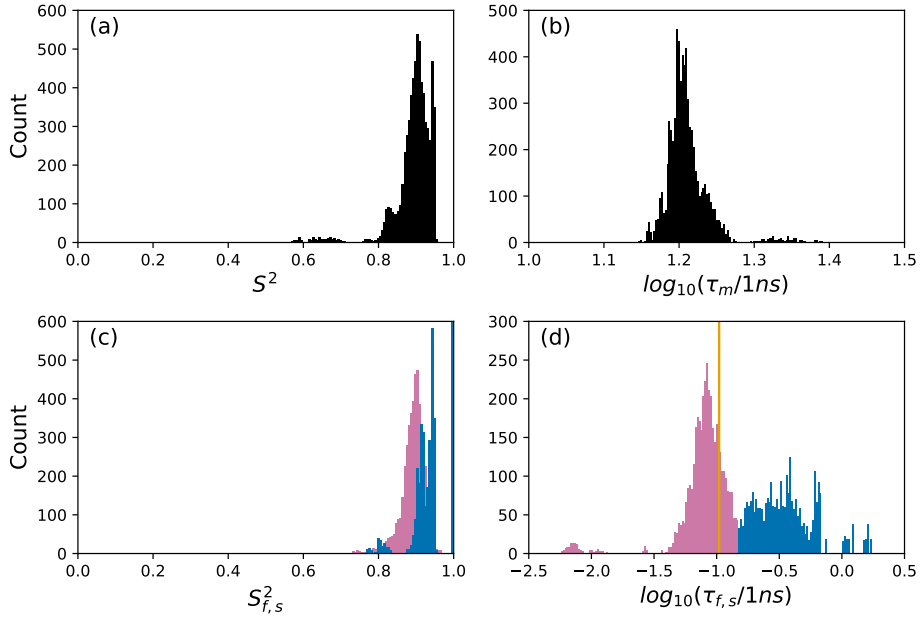

Figure 6. Distribution of model-free parameters from bootstrap aggregation for residue Arg 26. Color coding is (pink) S_f^2 or τ_f and (blue) S_s^2 or τ_s. The orange line in (c) indicates the value of τ_f obtained for the optimal single (unsmoothed) model 2. For clarity, null values of 1 for generalized order parameters and 0 for internal effective correlation times are not shown in the graphs; $S_f^2 = 1$ is observed 2167 times, $S_s^2 = 1$ is observed 3884 times, $\tau_f = 0$ is observed 2823 times, and $\tau_s = 0$ is observed 3884 times.

sampling of the distribution of the data. This issue is made more serious by the nature of nuclear spin relaxation data: spectral density values for $J(0)$, $J(\omega_N)$ and $J(0.87\omega_H)$ are very different and should not be interchanged by resampling. As shown in the present work, resampling within blocks of spectral density values clustered as $J(0)$, $J(\omega_N)$ and $J(0.87\omega_H)$ recorded at four static magnetic fields is sufficient to enable bootstrap aggregation.

Aggregation improves parameter stability by averaging over all models represented in the bootstrap sample. As applied to ^{15}N spin relaxation data for the bZip domain of GCN4, bootstrap aggregation reduces residue-to-residue variations in optimal model-free parameters, particularly in the partially disordered basic region. Consequently, trends in the conformational dynamics along the polypeptide backbone that reflect actual physical properties of the protein become more evident. NMR spin relaxation spectroscopy is a powerful approach for interogating conformional dynamics of biological macromolecules. Bootstrap aggregation, coupled with experimental measurements at multiple static magnetic fields, promises to advance efforts to understand the interplay between conformation and function in biology.

Code and data availability. A Jupyter notebook (Python 3.6) is provided that contains code for performing all data analyses reported in the publication. The NMR data analyzed in the publication are available at Mendeley Data (http://dx.doi.org/10.17632/vpwz6mrynr.1).

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

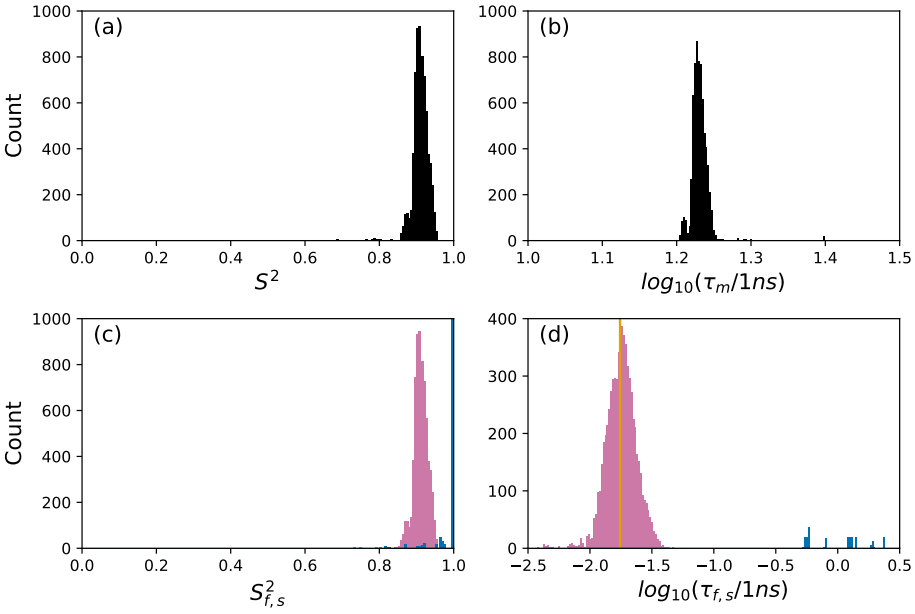

Figure 7. Distribution of model-free parameters from bootstrap aggregation for residue Asp 32. Color coding is (pink) S_f^2 or τ_f and (blue) S_s^2 or τ_s. The orange line in (c) indicates the value of τ_f obtained for the optimal single (unsmoothed) model 2. For clarity, null values of 1 for generalized order parameters and 0 for internal effective correlation times are not shown in the graphs; $S_s^2 = 1$ was observed 6650 times.

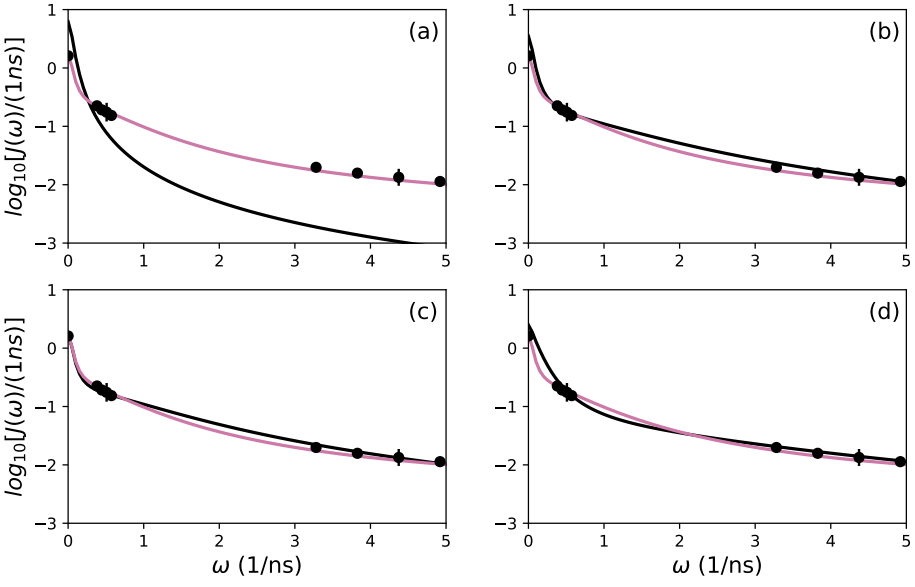

Figure 8. Comparison of (black lines) individual fits for Arg 11 of (a) model 1, (b) model 3, (c) model 4, and (d) model 5 or (pink) the bootstrap aggregation smoothed fit to (circles) experimental spectral density values.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

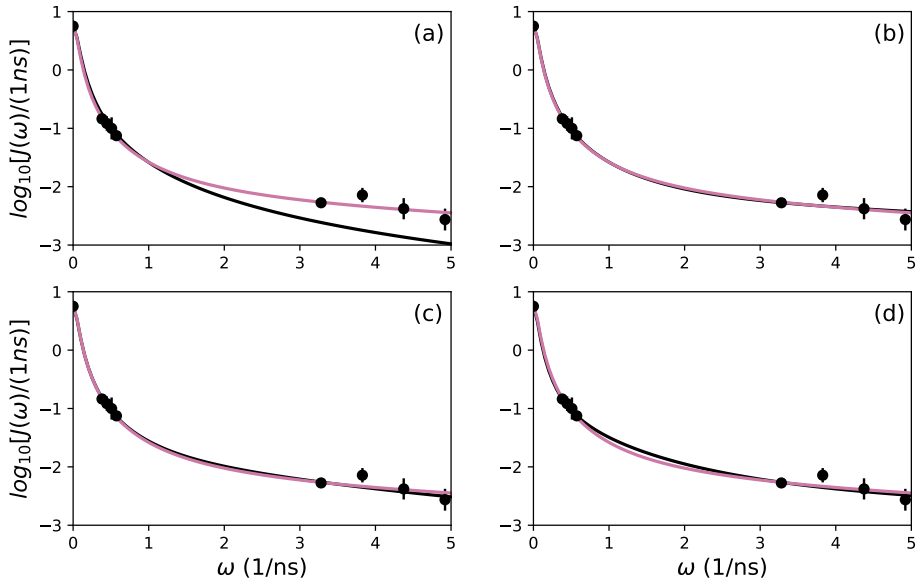

Figure 9. Comparison of (black lines) individual fits for Arg 26 of (a) model 1, (b) model 2, (c) model 4, and (d) model 5 or (pink) the bootstrap aggregation smoothed fit to (circles) experimental spectral density values.

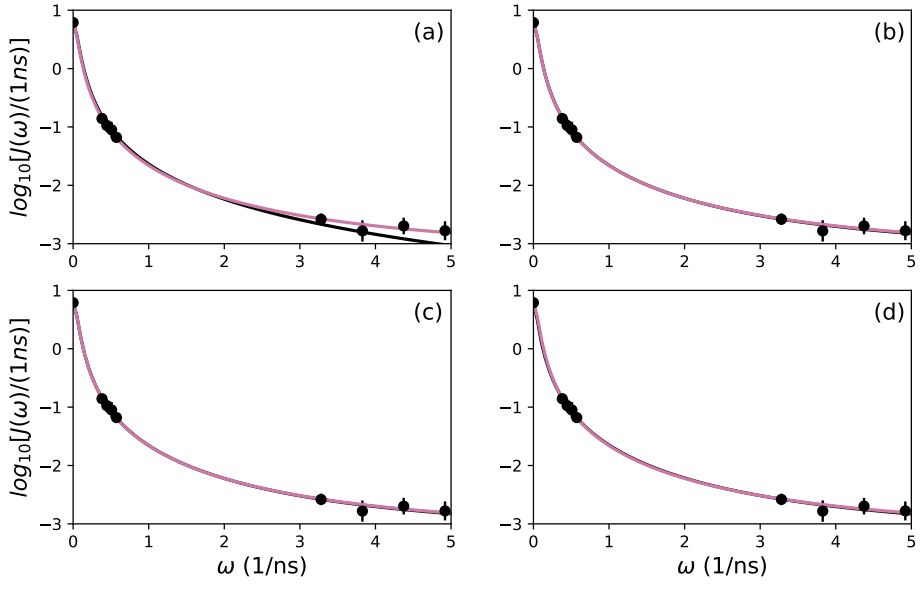

Figure 10. Comparison of (black lines) individual fits for Asp 32 of (a) model 1, (b) model 2, (c) model 4, and (d) model 5 or (pink) the bootstrap aggregation smoothed fit to (circles) experimental spectral density values.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Author contributions. A.G.P. conceived the project. Calculations and writing of the paper were performed by T.C. and A.G.P.

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was supported by National Institutes of Health grant R35 GM130398 (A. G. P.). Some of the work presented here was conducted at the Center on Macromolecular Dynamics by NMR Spectroscopy located at the New York Structural Biology Center, supported by a grant from the NIH National Institute of General Medical Sciences (P41 GM118302). A.G.P. is a member of the New York Structural Biology Center. This paper is dedicated to Prof. Geoffrey Bodenhausen on the occasion of his 70th birthday.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

References

Box, G. E. P.: Science and statistics, J. Am. Stat. Assoc., 71, 791–799, 1976.

Breiman, L.: Bagging predictors, Mach. Learn., 24, 123–140, 1996.

Buja, A. and Stuetzle, W.: Observations on bagging, Stat. Sin., 16, 323–351, 2006.

Chen, J., Brooks C.L., 3rd, and Wright, P. E.: Model-free analysis of protein dynamics: assessment of accuracy and model selection protocols based on molecular dynamics simulation, J. Biomol. NMR, 29, 243–57, 2004.

Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C., and Gronenborn, A. M.: Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins, J. Am. Chem. Soc., 112, 4989–4991, 1990.

d'Auvergne, E. J. and Gooley, P. R.: The use of model selection in the model-free analysis of protein dynamics, J. Biomol. NMR, 25, 25–39, 2003.

d'Auvergne, E. J. and Gooley, P. R.: Set theory formulation of the model-free problem and the diffusion seeded model-free paradigm, Mol. Biosyst., 3, 483–94, 2007.

d'Auvergne, E. J. and Gooley, P. R.: Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces, J. Biomol. NMR, 40, 107–19, 2008a.

d'Auvergne, E. J. and Gooley, P. R.: Optimisation of NMR dynamic models II. A new methodology for the dual optimisation of the model-free parameters and the Brownian rotational diffusion tensor, J. Biomol. NMR, 40, 121–33, 2008b.

Efron, B.: Estimation and accuracy after model selection, J. Am. Stat. Assoc., 109, 991–1007, 2014.

Farrow, N., Zhang, O., Szabo, A., Torchia, D., and Kay, L.: Spectral density function mapping using ^{15}N relaxation data exclusively, J. Biomol. NMR, 6, 153–162, 1995.

Gill, M. L., Byrd, R. A., and Palmer, A. G.: Dynamics of GCN4 facilitate DNA interaction: a model-free analysis of an intrinsically disordered region, Phys Chem Chem Phys, 18, 5839–49, 2016.

Halle, B. and Wennerström, H.: Interpretation of magnetic resonance data from water nuclei in heterogeneous systems, J. Chem. Phys., 75, 1928–1943, 1981.

Hsu, A., Ferrage, F., and Palmer, A. G.: Analysis of NMR spin-relaxation data using an inverse Gaussian distribution function, Biophys. J., 115, 2301–2309, 2018.

Hsu, A., Ferrage, F., and Palmer, A. G.: Correction: Analysis of NMR spin-relaxation data using an inverse Gaussian distribution function, Biophys. J., 119, 884–885, 2020.

Hurvich, C. M. and Tsai, C.-L.: Regression and time series model selection in small samples, Biometrika, 76, 297–307, 1989.

Khan, S. N., Charlier, C., Augustyniak, R., Salvi, N., Déjean, V., Bodenhausen, G., Lequin, O., Pelupessy, P., and Ferrage, F.: Distribution of pico- and nanosecond motions in disordered proteins from nuclear spin relaxation, Biophys. J., 109, 988–99, 2015.

Kroenke, C., Loria, J. P., Lee, L., Rance, M., and Palmer, A. G.: Longitudinal and transverse ^{1}H-^{15}N dipolar/^{15}N chemical shift anisotropy relaxation interference: unambiguous determination of rotational diffusion tensors and chemical exchange effects in biological macromolecules, J. Am. Chem. Soc., 120, 7905–7915, 1998.

Lee, L., Rance, M., Chazin, W., and Palmer, A.: Rotational diffusion anisotropy of proteins from simultaneous analysis of ^{15}N and ^{13}Cα nuclear spin relaxation, J. Biomol. NMR, 9, 287–298, 1997.

Lemaster, D. M.: Larmor frequency selective model free analysis of protein NMR relaxation, J. Biomol. NMR, 6, 366–74, 1995.

https://doi.org/10.5194/mr-2021-28
Preprint. Discussion started: 15 March 2021

Open Access MAGNETIC RESONANCE Discussions

Lipari, G. and Szabo, A.: Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity, J. Am. Chem. Soc., 104, 4546–4559, 1982a.

Lipari, G. and Szabo, A.: Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results, J. Am. Chem. Soc., 104, 4559–4570, 1982b.

Mandel, A. M., Akke, M., and Palmer, A. G.: Backbone dynamics of Escherichia coli ribonuclease HI: Correlations with structure and function in an active enzyme, J. Mol. Biol., 246, 144–163, 1995.

Palmer, A. G., Rance, M., and Wright, P. E.: Intramolecular motions of a zinc finger DNA-binding domain from xfin characterized by proton-detected natural abundance ^{13}C heteronuclear NMR spectroscopy, J. Am. Chem. Soc., 113, 4371–4380, 1991.

Smith, A. A., Ernst, M., Meier, B. H., and Ferrage, F.: Reducing bias in the analysis of solution-state NMR data with dynamics detectors, J. Chem. Phys., 151, 034 102, 2019.

Stone, M. J., Fairbrother, W. J., Palmer, A. G., Reizer, J., Saier, M. H., and Wright, P. E.: The backbone dynamics of the Bacillus subtilis glucose permease IIA domain determined from ^{15}N NMR relaxation measurements, Biochemistry, 31, 4394–4406, 1992.

Tugarinov, V., Liang, Z. C., Shapiro, Y. E., Freed, J. H., and Meirovitch, E.: A structural mode-coupling approach to ^{15}N NMR relaxation in proteins, J. Am. Chem. Soc., 123, 3055–3063, 2001.