Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Magnetic Resonance 177 (2005) 160–165

JMR
Journal of
Magnetic Resonance

www.elsevier.com/locate/jmr

Communication

Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar couplings constants

Michael Habeck [1], Wolfgang Rieping [1], Michael Nilges *

Unité de Bioinformatique Structurale, Institut Pasteur, CNRS URA 2185, 25-28, rue du Docteur Roux, 75015 Paris, France

Received 21 April 2005; revised 4 June 2005
Available online 8 August 2005

Abstract

We apply Bayesian inference to analyze three-bond scalar coupling constants in an objective and consistent way. The Karplus curve and a Gaussian error law are used to model scalar coupling measurements. By applying Bayes' theorem, we obtain a probability distribution for all unknowns, i.e., the torsion angles, the Karplus parameters, and the standard deviation of the Gaussian. We infer all these unknowns from scalar coupling data using Markov chain Monte Carlo sampling and analytically derive a probability distribution that only involves the torsion angles.

Keywords: Bayesian inference; Karplus parameters; Structure determination; Markov chain Monte Carlo; Marginalization

1. Introduction

Three-bond scalar coupling constants provide useful information on the local geometry of biomolecular structures [1]. The Karplus curve [2] relates the intervening dihedral angle φ to the three-bond scalar coupling constant:

$$^{3}J(\varphi) = A\cos^{2}\varphi + B\cos\varphi + C. \tag{1}$$

From a known Karplus curve, either bounds on torsion angles are derived [3,4], or the structure is directly refined against the measurements [5]. In both approaches the coefficients A, B, and C need to be known [6]. Since the Karplus coefficients are sensitive to the chemical environment of the macromolecule [7,8], use of empirically parametrized Karplus curves may introduce systematic errors. Therefore, the Karplus coefficients should, in principle, be calibrated for each molecular structure separately, for example, by estimating them from an X-ray structure of the molecule under investigation [3]. But then a previous structure determination must have been carried out. Furthermore, the chemical environment may differ for solution and crystal structure.

We use Bayesian inference to simultaneously determine protein torsion angles and the unknown Karplus parameters directly from experimental three-bond scalar coupling data without assuming prior knowledge of a pre-determined structure. Bayesian probability theory [9] provides a general framework for solving parameter estimation problems and has already been applied in NMR related data analysis to estimate coupling constants from antiphase multiplets [10], to analyze relaxation experiments [11], and for parameter estimation from time-domain data [12].

2. Methods and results

In a Bayesian structure determination framework [13,14], the problem of unknown Karplus coefficients can be treated in an elegant way. The observation of a

* Corresponding author.
E-mail addresses: habeck@pasteur.fr (M. Habeck), rieping@pasteur.fr (W. Rieping), nilges@pasteur.fr (M. Nilges).

[1] These authors contributed equally to this work.

1090-7807/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2005.06.016

scalar coupling constant of strength ${}^3J$ is described through a probability expressing the fact that, due to experimental and processing errors as well as theoretical shortcomings, measured and theoretically predicted scalar couplings will never match exactly. The least biasing error model assuming no systematic deviation and knowledge of the average discrepancy σ is a Gaussian [9]:

$$p({}^3J|\varphi,A,B,C,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}({}^3J-{}^3J(\varphi))^2\right\}. \tag{2}$$

This probability density function is conditioned on the actual value of the torsion angle φ, on the parameters A, B, C of the Karplus curve and on the global error σ. When we observe n couplings ${}^3J_1,\ldots,{}^3J_n$ independently, the likelihood function, i.e., the probability of all measurements, is

$$L(\boldsymbol{\varphi},A,B,C,\sigma)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\chi^2(\boldsymbol{\varphi},A,B,C)\right\} \tag{3}$$

with the goodness of fit

$$\chi^2(\boldsymbol{\varphi},A,B,C)=\sum_{i=1}^{n}({}^3J_i-A\cos^2\varphi_i-B\cos\varphi_i-C)^2. \tag{4}$$

Given n scalar coupling measurements only, the torsion angles $\boldsymbol{\varphi}=\{\varphi_i\}$, the Karplus parameters A, B, C, and the Gaussian error σ are unknown, but can be estimated from the data. The rationale is that any quantity entering the theoretical model to calculate the observable can be reconstructed from the measurements. Obviously, not only the torsion angles but also the Karplus coefficients determine the predicted values of scalar coupling constants; the error quantifies how close the fit can get. It should thus be possible to estimate these parameters from the data. Bayes' theorem [9] formalizes this rationale: upon modulation with a prior density $\pi(\boldsymbol{\varphi},A,B,C,\sigma)$, the likelihood determines the unknown parameters in terms of a posterior density

$$p(\boldsymbol{\varphi},A,B,C,\sigma)\propto L(\boldsymbol{\varphi},A,B,C,\sigma)\,\pi(\boldsymbol{\varphi},A,B,C,\sigma). \tag{5}$$

From the posterior density we can not only derive the most probable torsion angles but also the Karplus coefficients A, B, C, and the unknown error σ.

In case of the probabilistic models employed here, the parameters A, B, C, and σ can even be eliminated before structure calculation. This is made possible through the *marginalization* rule [9]: every uninteresting parameter must be integrated out in the posterior density. Assuming Jeffreys' prior [15], $\pi(\sigma)=\sigma^{-1}$, for the unknown scale parameter σ (i.e., a flat probability density in $\ln\sigma$), marginalization yields the integrated likelihood function:

$$\begin{aligned}L(\boldsymbol{\varphi},A,B,C)&=\int d\sigma\,L(\boldsymbol{\varphi},A,B,C,\sigma)\,\pi(\sigma)\\&\propto[\chi^2(\boldsymbol{\varphi},A,B,C)]^{-n/2}.\end{aligned} \tag{6}$$

Further integration over A, B, C (assuming a flat prior density $\pi(A,B,C)$) yields an integrated likelihood function that depends only on the torsion angles $\boldsymbol{\varphi}$

$$\begin{aligned}L(\boldsymbol{\varphi})&=\int dA\,dB\,dC\,L(\boldsymbol{\varphi},A,B,C)\,\pi(A,B,C)\\&\propto\det(\mathbf{A}(\boldsymbol{\varphi})^{\mathrm{T}}\mathbf{A}(\boldsymbol{\varphi}))^{-1/2}[\mathbf{j}^{\mathrm{T}}(\mathbf{I}-\mathbf{A}(\boldsymbol{\varphi})\mathbf{A}(\boldsymbol{\varphi})^{+})\mathbf{j}]^{-(n-3)/2}\end{aligned} \tag{7}$$

with the $n\times 3$ matrix

$$\mathbf{A}(\boldsymbol{\varphi})=\begin{pmatrix}\cos^2\varphi_1 & \cos\varphi_1 & 1\\ \vdots & \vdots & \vdots\\ \cos^2\varphi_n & \cos\varphi_n & 1\end{pmatrix} \tag{8}$$

and the data vector $\mathbf{j}=({}^3J_1,\ldots,{}^3J_n)^{\mathrm{T}}$; $\mathbf{A}^{+}=(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is the generalized inverse [16] of $\mathbf{A}$. If we assume the torsion angles to be known, $\mathbf{A}^{+}\mathbf{j}$ minimizes $\chi^2(\boldsymbol{\varphi},A,B,C)$ with respect to the Karplus parameters. This solution can also be given a probabilistic interpretation: $\mathbf{A}^{+}\mathbf{j}$ is the maximum likelihood estimate for the Karplus parameters.

Any quantity apart from the torsion angles has been eliminated via marginalization, thereby projecting the posterior density to conformational space

$$p(\boldsymbol{\varphi})\propto\exp\{-\beta E(\boldsymbol{\varphi})\}L(\boldsymbol{\varphi}). \tag{9}$$

Here, we used a Boltzmann distribution as prior density in torsion angle space [13] involving a potential energy E and the reciprocal temperature β. We can use the negative logarithm of $p(\boldsymbol{\varphi})$ to derive the most probable torsion angles.

The Karplus curve is highly degenerate, which complicates parameter estimation. A reflection of all torsion angles, $\varphi_i\to 2\pi-\varphi_i$, does not affect the calculated coupling constant. The transformations $\varphi_i\to\pi+\varphi_i$ and $\varphi_i\to\pi-\varphi_i$ can both be compensated by letting $B\to -B$. Since we are already facing an up to fourfold degeneracy in the Karplus curve with fixed coefficients, this degeneracy will be duplicated due to the invariance $B\to -B$ if we consider the Karplus parameters as free variables. Furthermore, if $|B|$ is close to zero, a $\pi/2$ shift in φ_i can be compensated by $A\to -A$, $C\to A+C$. These invariances make it impossible to uniquely determine torsion angles from a single data set; only by measuring data for different coupling types we can resolve the degeneracy in the torsion angles. However, without taking prior structural knowledge into account, there will still remain at least a twofold degeneracy due to the equality $\chi^2(\{\varphi_i+\pi\},A,-B,C)=\chi^2(\{\varphi_i\},A,B,C)$.

If different coupling types have been measured, we describe each data set with its own Karplus curve and its own error distribution. That is, data sets for m different coupling types necessitate $3m$ Karplus coefficients A_j, B_j, and C_j and m error parameters σ_j, and the joint posterior distribution for all unknowns is

$$p(\boldsymbol{\varphi}, \{A_j, B_j, C_j, \sigma_j\}) \propto \exp\{-\beta E(\boldsymbol{\varphi})\} \times \prod_{j=1}^{m} L(\boldsymbol{\varphi}, A_j, B_j, C_j, \sigma_j) \qquad (10)$$

with $L(\boldsymbol{\varphi}, A_j, B_j, C_j, \sigma_j)$ defined as in (3). Since the parameters for different data sets separate, marginalization is straightforward, and the marginal posterior density for the torsion angles only becomes

$$p(\boldsymbol{\varphi}) \propto \exp\{-\beta E(\boldsymbol{\varphi})\} \prod_{j=1}^{m} L_j(\boldsymbol{\varphi}), \qquad (11)$$

where each likelihood factor $L_j(\boldsymbol{\varphi})$ is of form (7) involving matrices $\mathbf{A}(\boldsymbol{\varphi} - \boldsymbol{\delta}_j)$ with $\boldsymbol{\delta}_j$ being the phase angles of the jth coupling type.

We analyzed ^3J coupling data measured on ubiquitin [17,18] (PDB restraint file 1d3z), comprising values for six scalar couplings that involve the main chain torsion angle φ. We removed two outliers in the data set for the C′–C′ coupling and simulated the joint posterior densities (10) and (11). Besides the 72 φ_i angles, we estimated six Karplus curves and six error parameters by posterior simulation [19] using a Gibbs sampling [20] scheme. Gibbs sampling is a Markov chain Monte Carlo technique which permits simulation of multidimensional probability distributions in an iterative fashion. For one cycle, every parameter is drawn, one after the other, from its conditional posterior distribution, while inserting the most recent values of the remaining parameters into the conditioning side of that distribution. Iteration of this rule generates a sequence of stochastic samples drawn from the target distribution in (10). Here, the conditional posterior density of the Karplus parameters is a three-dimensional Gaussian, and the inverse squared error follows a gamma distribution. Random number generators for these distributions exist. We sampled the torsion angles by approximating their conditional posterior densities with a histogram. These histograms were obtained by calculating slices in $\chi^2(\boldsymbol{\varphi}, A_j, B_j, C_j)$ and $\ln L_j(\boldsymbol{\varphi})$, respectively (setting $\beta = 0$), where only one torsion angle is varied in steps of 3.6°. To accelerate convergence of the calculation, we embedded the Gibbs sampler into a Replica-exchange Monte Carlo scheme as described in [21] with only one pseudo-temperature λ. Our setup comprises 50 copies of the Gibbs sampler, each simulating a "heated" target distribution, with inverse temperatures ranging from $\lambda_{\min} = 1.0$ for the target distribution to $\lambda_{\max} = 0.1$ for the high-temperature distribution. Provided $\lambda_{\max}$ is sufficiently small, this distribution is practically flat which ensures ergodic sampling. Stochastic exchanges of samples between neighboring copies allow the simulation to escape local modes.

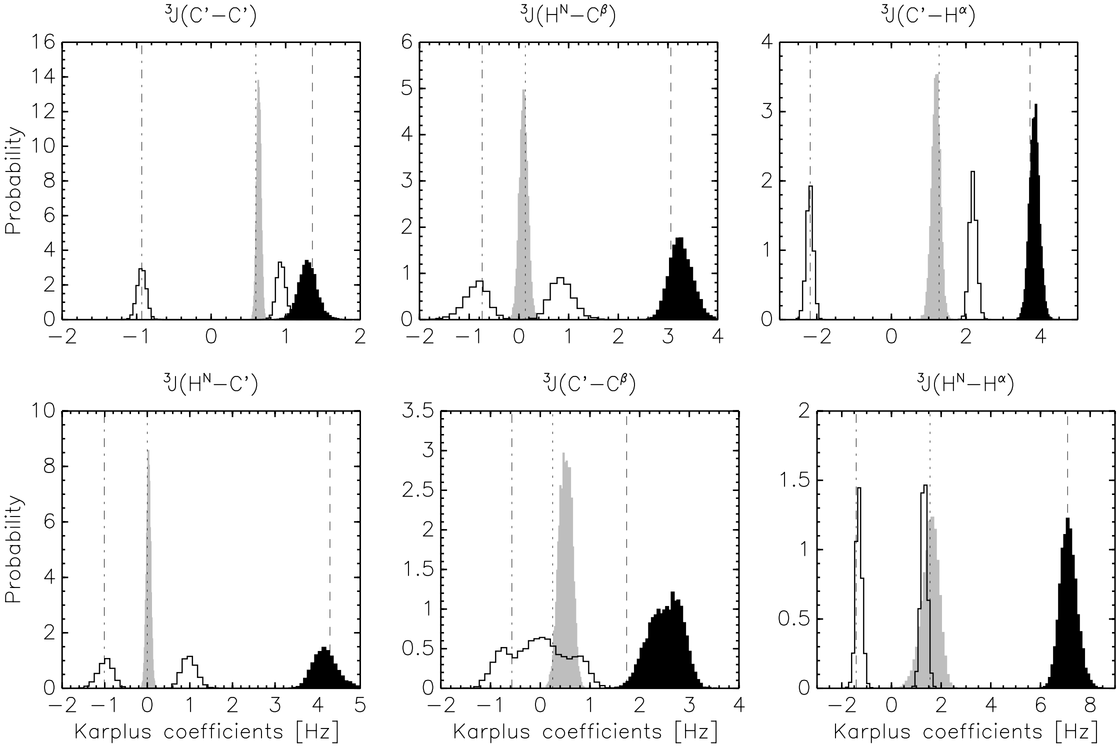

Fig. 1. Posterior histograms for the Karplus coefficients A (filled black histogram), B (black curve), and C (filled grey histogram) of the six scalar couplings. Vertical lines indicate the estimates reported in PDB restraint file 1d3z (A: dashed, B: dot-dashed, C: dotted).

Fig. 1 shows the posterior histograms of the Karplus coefficients. The estimated coefficients scatter around the values obtained by maximum likelihood using the known NMR structure [18]. The posterior density of the Karplus coefficient B is at least bimodal due to the abovementioned reflection symmetry that pertains also for multiple data sets. The posterior histograms can directly be used to derive error estimates for the Karplus coefficients; strategies relying on cross-validation, like those employed in [22,17], are superfluous. Table 1 lists the mean values and their precision for all Karplus parameters. In most cases the value reported in the PDB restraint file 1d3z is found to lie within the error interval of the Bayesian estimate. Only the C′–C$^{\beta}$ couplings are problematic.

The Bayesian approach allows us to estimate the parametrizations of the Karplus curves and to simultaneously reconstruct the torsion angles for which scalar couplings have been measured. Fig. 2 shows the posterior histograms for three representative torsion angles. All three angles scatter around the respective values found in the NMR structure 1d3z. The precision of the estimates depends on the number of scalar coupling measurements observed for a particular angle: all six coupling constants were measured for φ_{Lys6}, two measurements involve φ_{Gly10}, only one measurement involves φ_{Glu24}; the spread of the φ-histograms reflects this fact. If we are not interested in the parametrization of the Karplus curves, we can also use the marginalized posterior density (11) (with $\beta = 0$), instead of the joint posterior density, to estimate the torsion angles. A high overlap of the posterior histograms demonstrates the equivalence of both approaches (see Fig. 2).

Fig. 3 shows torsion angle samples generated from the joint and from the marginal posterior density. Most of the torsion angles of the NMR structure 1d3z are found to lie within the sampled region or in close proximity. Yet, it is difficult to determine macromolecular structures from ^{3}J data alone. This is due to a lack in precision of the reconstructed torsion angles as well as to missing information on other torsion angles. Fig. 3 once more demonstrates the equivalence of the joint and the marginal posterior density: the torsion angle samples obtained from both simulations are very similar. Again the degeneracy due to the reflection symmetry in B is observed. This symmetry can be broken by taking prior structural knowledge into account, i.e. by setting the reciprocal temperature β in the conformational prior density to a realistic value. The posterior densities of the torsion angles will then be unimodal (data not shown).

Table 1
Mean values and precision of the six Karplus curves (due to the twofold degeneracy in B we calculated the statistics for its absolute value)

	A (Hz)	$\|B\|$ (Hz)	C (Hz)
^{3}J(C′–C′)	1.30 ± 0.12 (1.36)	0.93 ± 0.06 (0.93)	0.64 ± 0.03 (0.60)
^{3}J(C′–H$^{\alpha}$)	3.84 ± 0.14 (3.72)	2.19 ± 0.10 (2.18)	1.20 ± 0.11 (1.28)
^{3}J(C′–C$^{\beta}$)	2.52 ± 0.33 (1.74)	0.49 ± 0.33 (0.57)	0.51 ± 0.12 (0.25)
^{3}J(H^{N}–C′)	4.19 ± 0.30 (4.29)	0.99 ± 0.18 (1.01)	0.03 ± 0.05 (0.00)
^{3}J(H^{N}–H$^{\alpha}$)	7.13 ± 0.34 (7.09)	1.31 ± 0.13 (1.42)	1.56 ± 0.34 (1.55)
^{3}J(H^{N}–C$^{\beta}$)	3.26 ± 0.23 (3.06)	0.87 ± 0.24 (0.74)	0.10 ± 0.08 (0.13)

The values reported in the PDB restraint file 1d3z are indicated in brackets.

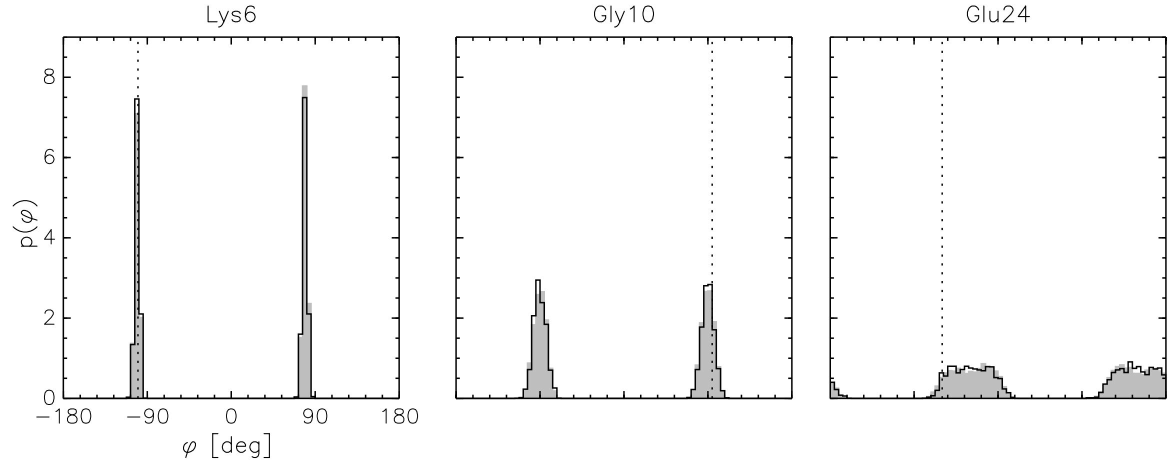

Fig. 2. Posterior histograms for three representative φ angles (in torsion angle degrees). The shaded histogram stems from a simulation of the joint posterior density, the black lines are the posterior histograms obtained for the marginal posterior density. The values found in the NMR structure 1d3z are indicated by dotted lines.

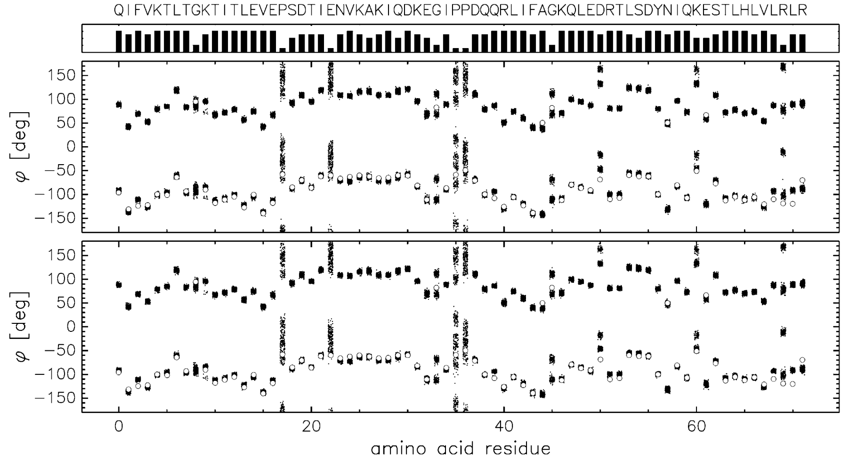

Fig. 3. Torsion angle samples from a simulation of the joint posterior density (top) and of the marginal posterior density (bottom). The header shows the amino acid types and a histogram of the number of ^3J measurements per φ angle (maximum is six, minimum is one). Circles indicate the values found in the NMR structure 1d3z.

3. Discussion

Schmidt et al. [23] proposed a "self-consistent" method that calculates torsion angles and Karplus coefficients simultaneously. Their algorithm is a special case of the Bayesian treatment. In the self-consistent approach the residual $\chi^2(\boldsymbol{\varphi}, A, B, C)$ with an additional term that breaks the symmetry in B is minimized. This procedure corresponds to maximum likelihood (ML) estimation [9]. The deficiencies of ML-based analyses are their inability to take prior knowledge into account and to fully exploit probability densities with sampling algorithms. As a consequence, no statements can be made about the precision of parameter estimates. Another advantage of our approach over the self-consistent method is that we estimate the unknown errors σ_j during structure calculation. Thus, errors are not fixed to a constant value (0.25 Hz in case of [23]) but are optimally adapted to the quality of the data.

The current model describes scalar coupling data as instantaneous measurements. It has been shown that three-bond scalar couplings are subject to motional averaging [24]. A model assuming Gaussian torsional fluctuations that can account for such averaging effects has been proposed in [25]. This model adds to each unknown torsion angle an unknown variance quantifying the magnitude of the local motion. The self-consistent method is able to determinate the additional variances, and it would be straightforward to estimate the motional variance in the Bayesian approach. Also, marginalization is still possible and results in a joint posterior defined on the torsion angles and their variances.

A Bayesian structure determination framework allows an objective interpretation of scalar coupling measurements. Parametrizations of Karplus curves can either be estimated directly from the data during structure calculation or eliminated beforehand by marginalization. Our method directly derives from probability theory. It does not introduce additional heuristics and is thus consistent and unbiased.

References

[1] D.S. Garrett, J. Kuszewski, T.J. Hancock, P.J. Lodi, G.W. Vuister, A.M. Gronenborn, G.M. Clore, The impact of direct refinement against three-bond HN–C$_\alpha$H coupling constants on protein structure determination by NMR, J. Magn. Reson. 104 (1) (1994) 99–103.

[2] M. Karplus, Vicinal proton coupling in nuclear magnetic resonance, J. Am. Chem. Soc. 85 (1963) 2870–2871.

[3] A. Pardi, M. Billeer, K. Wüthrich, Calibration of the angular dependence of the amide proton–C$_\alpha$ proton coupling constants, $^3J_{\mathrm{HN}\alpha}$, in a globular protein, J. Mol. Biol. 189 (1984) 383–386.

[4] P. Güntert, W. Braun, M. Billeter, K. Wüthrich, Automated stereospecific ^{1}H NMR assignments and their impact on the precision of protein structure determination in solution, J. Am. Chem. Soc. 111 (1989) 3997–4004.

[5] Y. Kim, J.H. Prestegard, Refinement of the NMR structures for acyl carrier protein with scalar coupling data, Proteins Struct. Funct. Genet. 8 (1990) 377–385.

[6] D.A. Case, H.J. Dyson, P.E. Wright, Use of chemical shifts and coupling constants in NMR structural studies on peptides and proteins, Methods Enzymol. 239 (1994) 392–416.

[7] C.A.G. Haasnoot, F.A.A.M. de Leeuw, C. Altona, Relationship between vicinal H–H coupling constants and substituent electronegativities, Tetrahedron 36 (1980) 2783–2792.

[8] C.A.G. Haasnoot, F.A.A.M. de Leeuw, H.P.M. de Leeuw, C. Altona, Relationship between proton–proton NMR coupling constants and substituent electronegativities. III. Conformational analysis of proline rings in solution using a generalized Karplus equation, Biopolymers 20 (1981) 1211–1245.

[9] E.T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK, 2003.

[10] M. Andrec, J.H. Prestegard, A metropolis Monte Carlo implementation of bayesian time-domain parameter estimation: application to coupling constant estimation from antiphase multiplets, J. Magn. Reson. 130 (1998) 217–232.

[11] M. Andrec, G.T. Montelione, R.M. Levy, Estimation of dynamic parameters from NMR relaxation data using the Lipari–Szabo model-free approach and Bayesian statistical methods, J. Magn. Reson. 139 (1999) 408–421.

[12] G.L. Bretthorst, Bayesian analysis I. Parameter estimation using quadrature NMR models, J. Magn. Reson. 99 (1990) 533–551.

[13] M. Habeck, W. Rieping, M. Nilges, A new principle for macromolecular structure determination, in: G. Erickson, Y. Zhai (Eds.), 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, American Institute of Physics, 2004, pp. 157–166.

[14] W. Rieping, M. Habeck, M. Nilges, Inferential Structure Determination, Science 309 (2005) 303–306.

[15] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. Roy. Soc. A 186 (1946) 453–461.

[16] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes: The Art of Scientific Computing, Cambridge University Press, Cambridge, UK, 1989.

[17] A. Wang, A. Bax, Determination of the backbone dihedral angles ϕ in human ubiquitin from reparametrized empirical Karplus equations, J. Am. Chem. Soc. 118 (1996) 2483–2494.

[18] G. Cornilescu, J.L. Marquardt, M. Ottiger, A. Bax, Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase, J. Am. Chem. Soc. 120 (1998) 6836–6837.

[19] M.H. Chen, Q.M. Shao, J.G. Ibrahim, Monte Carlo methods in bayesian computation, Springer, New York, 2002.

[20] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images, IEEE Trans. PAMI 6 (6) (1984) 721–741.

[21] M. Habeck, M. Nilges, W. Rieping, Replica-exchange Monte Carlo scheme for bayesian data analysis, Phys. Rev. Lett. 94 (2005) 018105.

[22] G.W. Vuister, A. Bax, Quantitative J correlation: a new approach for measuring homonuclear three-bond $J(H^N–H^\alpha)$ coupling constants in ^{15}N-enriched proteins, J. Am. Chem. Soc. 115 (1993) 7772–7777.

[23] J.M. Schmidt, M. Blümel, F. Löhr, H. Rüterjans, Self-consistent ^3J coupling analysis for the joint calibration of Karplus coefficients and evaluation of torsion angles, J. Biomol. NMR 14 (1999) 1–12.

[24] J.C. Hoch, C.M. Dobson, M. Karplus, Vicinal coupling constants and protein dynamics, Biochemistry 24 (1985) 3831–3841.

[25] R. Brüschweiler, D.A. Case, Adding harmonic motion to the Karplus equation for spin–spin coupling, J. Am. Chem. Soc. 116 (1994) 11199–11200.