

Modeling Errors in NOE Data with a Log-normal Distribution Improves the Quality of NMR Structures

Wolfgang Rieping, Michael Habeck, and Michael Nilges*

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

Received August 22, 2005; E-mail: nilges@pasteur.fr

Nuclear Overhauser effect (NOE) data are routinely used to determine the structure of biomolecules.¹ In a first-order approximation, the isolated spin pair approximation (ISPA) relates the intensity of an NOE to the inverse sixth power of the inter-proton distance.² Molecular motion,³ spin diffusion,⁴ and experimental noise lead to additional contributions to cross-relaxation rates not captured by the ISPA. Calculated intensities will therefore differ from the true values. To avoid distortions in the calculated structure, intensities are usually converted into distance bounds, and structure calculation boils down to the generation of structures that satisfy the bounds.⁵ This approach seems plausible but comes at a price. First, the bounds have to be set empirically and often need to be adjusted selectively to be mutually consistent. Second, flat-bottom restraint potentials used to incorporate the experimental distances into a calculation weigh all measurements equally, provided that their back-calculated distances lie within the bounds. Hence, structurally supported measurements and those close to being violated contribute equally to the total energy. Third, bounds that are too large reduce the information content of the data, which leads to less precise and less accurate structures.

A probabilistic approach to structure determination⁶ avoids these difficulties. NOE intensities are described by means of a likelihood function which, in the present context, consists of the ISPA to predict the intensities from the structure and an error distribution to account for deviations between measured and calculated values. Ideally, the error model reproduces the true, however unknown, “experimental” error distribution in the data. Here, we develop formal and pragmatic arguments to show that a log-normal distribution is a natural choice for describing these deviations. Our model permits the calculation of a structure directly from the measured intensities and improves structural quality compared to the usual bounds representation.

NOE intensities are inherently positive. A calibration factor γ needs to be introduced in order to relate the intensity scale to a distance scale. However, changing the units does not affect the information content of the data. Hence, the distribution, $g(I_{\text{obs}}, I_{\text{calc}})$, of the deviations between observed and calculated intensities must be invariant under scaling, that is, $g(I_{\text{obs}}, I_{\text{calc}}) = \gamma g(\gamma I_{\text{obs}}, \gamma I_{\text{calc}})$, which follows from the transformation rule of probability densities. This general equation must also hold for the special case $\gamma = 1/I_{\text{calc}}$, and we obtain

$$g(I_{\text{obs}}, I_{\text{calc}}) = g(I_{\text{obs}}/I_{\text{calc}}, 1)/I_{\text{calc}} = h(I_{\text{obs}}/I_{\text{calc}})/I_{\text{calc}} \quad (1)$$

with the univariate density $h(\cdot) = g(\cdot, 1)$ defined on the positive axis. Equation 1 states that the error distribution depends on the ratio of observed and calculated intensity, meaning the error is multiplicative. This is in contrast to the usual, in our view, inappropriate practice of assuming that errors in NOE intensities or in NOE-derived distances are additive. We can freely choose the density h . In the absence of systematic errors, the log-error

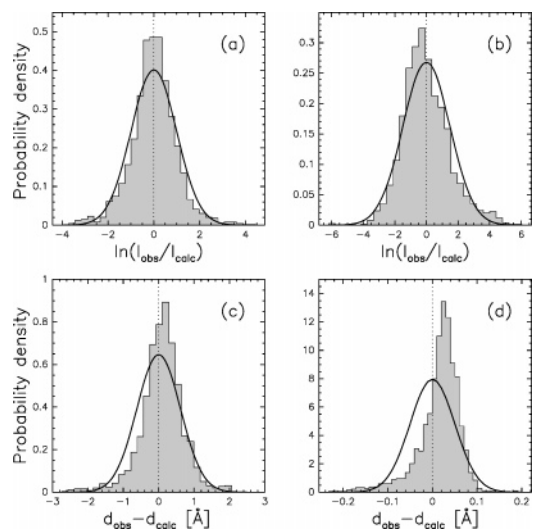


Figure 1. Experimental intensity error distributions for ubiquitin (a) and the Tudor domain (b). Solid lines indicate fitted log-normal distributions. (c,d) Corresponding distance error distributions and fitted Gaussians. Because intensities were unavailable for ubiquitin, we used the ISPA to convert the published distance data into intensities (1444 nonredundant distances taken from the restraint file, PDB code 1d3z). The Tudor data (1875 intensities from two ¹³C and ¹⁵N edited spectra) were calibrated such that observed and calculated values have the same geometric mean.

scatters around zero with a certain variance σ^2 . In this case, the Maximum entropy principle⁷ determines the least-biasing h , and we obtain

$$g(I_{\text{obs}}, I_{\text{calc}}) = \frac{1}{\sqrt{2\pi\sigma^2}I_{\text{obs}}} \exp\left\{-\frac{1}{2\sigma^2}\log^2\left(\frac{I_{\text{obs}}}{I_{\text{calc}}}\right)\right\} \quad (2)$$

The log-normal distribution in eq 2 is restricted to the positive axis and asymmetric around its median I_{calc} . Measurements are incorporated without bias in the sense that the probability of over- or underestimating the true intensity is both 1/2. This is not the case for error distributions defined on the entire axis, such as a Gaussian, which assign a nonvanishing probability to unobservable negative intensities. The parameter σ quantifies the relative deviation of the observed from the calculated intensity, provided that their difference is sufficiently small.

Figure 1a,b shows the “experimental” distributions of the intensity error for two data sets measured on the proteins ubiquitin⁸ and the Tudor domain⁹ using the published X-ray structures^{10,11} as reference. The fitted log-normal distribution captures most of the features of the experimental distributions. For comparison, we performed the same analysis for the distance differences $d_{\text{obs}} - d_{\text{calc}}$ as one would do when modeling the experimental error with, for example, a Gaussian distribution. Figure 1c,d demonstrates that the

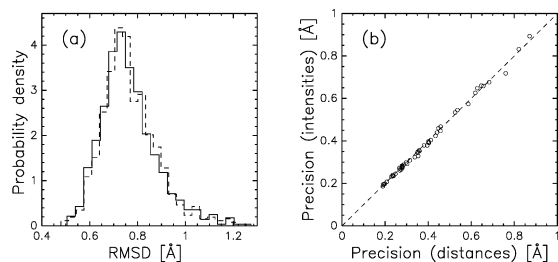


Figure 2. Power law invariance of the log-normal model. The conformational distributions of the Tudor domain obtained using intensities (solid line) and distances (dashed line) are identical in terms of heavy atom backbone RMSD to the X-ray structure (a) and C_{α} precision (b).

distance error distributions are asymmetric and long-tailed, and that a Gaussian can significantly underestimate the likelihood of observing outliers. Both properties are much better accounted for by the log-normal distribution due to the logarithmic transformation.

From a pragmatic point of view, the log-normal model has several favorable properties. Unlike a probability distribution corresponding to a flat-bottom potential, it has a unique maximum. Hence, measurements are not weighted equally but are penalized depending on the degree of disagreement with the structure. Furthermore, the log-normal distribution is invariant under power law transformations. If we consider $I_{\text{obs}}' = (I_{\text{obs}})^{\alpha}$ with exponent α , the transformed intensity still follows a log-normal law, however, with transformed median and error parameter $(I_{\text{calc}})^{\alpha}$ and $\sigma' = |\alpha|\sigma$, respectively. Because the ISPA is a special case of a power transformation ($\alpha = -1/6$), we obtain identical structures regardless whether we refine against intensities or distances, provided that σ has been transformed appropriately.

We have implemented the log-normal model in our software for probabilistic structure determination (ISD, in preparation). ISD is based on Monte Carlo sampling rather than energy minimization to explore the probability distribution over conformational space.¹² This allows us to determine the most likely conformation of a macromolecule, including its uncertainty. In an original structure determination, a reference structure that could be used to determine the optimal shape of the log-normal distribution by adjusting σ is usually unavailable. Since the error parameter is a priori unknown, it needs to be estimated during structure calculation, which is straightforward in a probabilistic framework.⁶ We employed the ISD software package to calculate the structure of ubiquitin and the Tudor domain from the published data sets.^{8,9} Estimation of σ yielded optimal values of $\sigma = 0.94$ for ubiquitin and $\sigma^{\text{C13}} = 1.24$ and $\sigma^{\text{N15}} = 0.98$, respectively, for the two data sets of the Tudor domain. The most probable structure of ubiquitin showed a heavy atom backbone RMS deviation of 0.61 Å to the X-ray structure, and for the Tudor domain, we found a value of 0.70 Å. A recalculation of the Tudor domain, this time using distances instead of intensities and a transformed error parameter $\sigma' = \sigma/6$, demonstrates the power law invariance of the log-normal model. Both simulations are identical in terms of accuracy and precision (Figure 2a,b).

To assess the impact of the log-normal model on structural quality, we repeated the calculations using distance bounds and a flat-bottom potential with harmonic walls and a force constant of 50 kcal mol⁻¹ Å⁻². Both calculations were performed with the ISD software; for the Tudor domain, we determined the distance bounds by using the rule implemented in the computer program ARIA.¹³ The structures generated with the log-normal model were found to be systematically closer to the X-ray structure than those obtained with the standard procedure (Table 1). We also observe an improvement in terms of precision. For ubiquitin, the heavy atom

Table 1. Comparison of Structural Quality Indicators^a

	Procheck	log-normal ^b	bounds ^b	log-normal ^c	bounds ^c
most favored	80.5 ± 3.1	76.8 ± 3.8	79.4 ± 4.0	68.9 ± 4.6	68.9 ± 4.6
allowed	19.3 ± 3.1	22.6 ± 3.9	19.9 ± 4.2	27.0 ± 4.7	27.0 ± 4.7
gen. allowed	0.2 ± 0.5	0.5 ± 0.8	0.7 ± 1.1	1.5 ± 2.1	1.5 ± 2.1
disallowed	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	2.6 ± 1.4	2.6 ± 1.4
WhatIf					
QUACHK ^d	-0.54 ± 0.25	-1.45 ± 0.31	-1.88 ± 0.28	-2.41 ± 0.25	-2.41 ± 0.25
NQACHK ^e	-1.34 ± 0.35	-2.02 ± 0.40	-0.38 ± 0.49	-1.29 ± 0.57	-1.29 ± 0.57
RAMCHK ^f	-3.14 ± 0.40	-3.30 ± 0.45	-1.64 ± 0.74	-4.15 ± 0.62	-4.15 ± 0.62
BBCCHK ^g	1.16 ± 0.57	0.76 ± 0.56	-0.10 ± 0.70	-2.36 ± 1.01	-2.36 ± 1.01
RMSD [Å] ^h	0.64 ± 0.06	0.72 ± 0.07	0.68 ± 0.06	0.99 ± 0.06	0.99 ± 0.06

^a Averages and standard deviations for the 100 most likely conformations generated with the log-normal model and a flat-bottom restraint potential. ^b Ubiquitin. ^c Structured part of the Tudor domain (residues 92–144). Procheck¹⁵ Ramachandran statistics are in percent, WhatIf¹⁴ values are Z-scores. ^{d,e} First and second generation packing quality. ^f Ramachandran plot appearance. ^g Backbone conformation normality score. ^h Heavy atom backbone RMS deviation to the X-ray structure.

backbone ensemble RMSD, calculated from the 100 most likely conformations, amounts to 0.30 Å, compared to 0.44 Å in case of the flat-bottom potential. For the structured part of the Tudor domain, the structural uncertainty is low for both models (0.35 and 0.37 Å, respectively). The gain in accuracy is further supported by a significant improvement of widely used indicators of structural quality (Table 1).

Application of the log-normal model is not limited to the description of errors in NOE data. The log-normal distribution has a clear statistical interpretation. It can be thought of as the “Gaussian for positive quantities” and is, therefore, expected to be equally well suited for modeling the error distribution of other experimental observables that are inherently positive. Due to its unique maximum and the resulting individual weighting of each measurement, the log-normal model exploits the experimental information to a greater extent than distance bounds. This improves both the quality and the accuracy of the calculated structures and is also expected to prove useful in defining a meaningful figure of merit for NMR structures solely based on the experimental data.

Acknowledgment. We thank Michael Sattler for kindly providing the experimental data on the Tudor domain. This work was supported by EU Grants QL2-CT-2000-01313 and QL2-CT-2002-00988.

References

- (1) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley: New York, 1986.
- (2) Neuhaus, D.; Williamson, M. P. *The Nuclear Overhauser Effect in Structural and Conformational Analysis*; VCH Publishers Inc.: New York, 1989.
- (3) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546–4558.
- (4) Kalk, A.; Berendsen, H. J. C. *J. Magn. Reson.* **1976**, *24*, 275–268.
- (5) Havel, T.; Kuntz, I. D.; Crippen, G. M. *Bull. Math. Biol.* **1983**, *45*, 665–720.
- (6) Rieping, W.; Habeck, M.; Nilges, M. *Science* **2005**, *309*, 303–306.
- (7) Jaynes, E. T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2003.
- (8) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.
- (9) Selenko, P.; Sprangers, R.; Stier, G.; Buehler, D.; Fischer, U.; Sattler, M. *Nat. Struct. Biol.* **2001**, *8*, 27–31.
- (10) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.
- (11) Sprangers, R.; Groves, M.; Sinning, I.; Sattler, M. *J. Mol. Biol.* **2003**, *327*, 507–520.
- (12) Habeck, M.; Nilges, M.; Rieping, W. *Phys. Rev. Lett.* **2005**, *94*, 018105.
- (13) Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* **2003**, *19*, 315–316.
- (14) Vriend, G. *J. Mol. Graph.* **1990**, *8*, 52–56.
- (15) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283–291.

JA055092C