Monte Carlo Estimation of Errors in ^{13}C-NMR Relaxation Studies of a DNA Oligomer Duplex

Ji-Qun Yu, Daniel S. Blumenthal, and Philip N. Borer*

Department of Chemistry, Syracuse University, Syracuse, New York 13244-4100

Received May 20, 1994®

An analysis of errors has been done with the Monte Carlo method for natural abundance ^{13}C-NMR relaxation studies of a DNA duplex. Repeated measurements of the longitudinal relaxation time, T_1, and the heteronuclear NOE were made at 90.6 MHz on the duplexed DNA pentanucleotide, [d(TCGCG)]$_2$.[1] The deviations averaged over all carbons were 13% for T_1 and 9% for NOE. These relative deviations were applied to generate 100 values of T_1 and NOE with normal distributions about the measured mean values for each carbon. A new version of MOLDYN,[2] called McMOLDYN, has been written, which was used to generate 100 values of T_1 and NOE with normal distributions corresponding to the measured errors; the same error distributions were also applied to measurements at 125.8 MHz. The order parameter, S^2, and the effective internal correlation time, τ_e, in the Model-Free Approach[3] have been optimized from the distributions simulated by McMOLDYN. McMOLDYN also permits the automated entry of multiple sets of initial guesses for the output parameters S^2, τ_e, and τ_m. In addition, McMOLDYN adds cross-relaxation terms from chemical shift anisotropy, increasingly important as spectrometer magnetic fields get higher. Between the two parameters optimized, S^2 has the smallest relative error, estimated at 15% on average, which means that S^2 is a well-defined parameter. However, τ_e is very poorly defined with the average relative error estimated 85%; it is typically found in the range of 30−300 ps.

INTRODUCTION

The model-free approach of Lipari and Szabo[3] describes isotropic overall reorientation by a single correlation time, τ_m, which is not coupled to internal motion. In its simplest form, internal motion is described by a generalized order parameter, S^2, which characterizes spatial restriction on the motion of an internuclear vector and τ_e, a quantity related to the rate of internal motion for the vector. Although τ_e has the appearance and units of a correlation time, it is also somewhat dependent on the amplitude of internal motion. For ^{13}C−^{1}H vectors, with lengths virtually fixed at 1.09 Å, the relaxation times, T_1 and T_2, and the nuclear Overhauser enhancement (NOE) can be expressed in terms of the spectral density $J(\omega)$ which is the Fourier transform of the time-dependent correlation function $C(t)$. By considering the properties of $C(t)$, which are independent of any particular model of motion, Lipari and Szabo introduced S^2, τ_e, and τ_m into the correlation function and gave an expression for $J(\omega)$. For a short DNA duplex, e.g., [d(TCGCG)]$_2$, the overall motion can be considered isotropic and $J(\omega)$ has a simple form[3]

$$J(\omega) = (2/5)(S^2\tau_m/(1 + (\tau_m)^2) + (1 - S^2)\tau/(1 + (\tau)^2) \quad (1)$$

where $1/\tau = 1/\tau_m + 1/\tau_e$. An optimal fit of $J(\omega)$ to the ^{13}C relaxation data provides values for the model-free parameters: S^2, τ_e, and τ_m.

We have published[1] relaxation data for [d(TCGCG)]$_2$ measured at 90.6, 62.9, and 125.7 MHz. The computer program MOLDYN[2] was used to optimize values of S^2, τ_e, and τ_m. In order to find out which model-free parameters are well-defined by experimental relaxation data, an error analysis is required. Since the MOLDYN optimization routine uses nonlinear least squares fitting, it is difficult to estimate the errors in these values directly from the errors in the measurements. In addition, using a large amount of experimental relaxation data to do error analysis is not practical because of limited experimental data.

This paper describes error estimates for S^2 and τ_e based on a Monte Carlo method. A large number of T_1 and NOE values were generated for each carbon at each of the field strengths, using the mean experimental values and a normal distribution of relative errors (the same as the average overall errors). A new version of MOLDYN, called McMOLDYN, has been developed, which includes the Monte Carlo routine and a correction for chemical shift anisotropy (CSA) and was used in this work. McMOLDYN can simulate normal distributions for several kinds of input relaxation measurements, automatically start the search procedure with eight combinations of initial guesses for S^2, τ_e, and τ_m to escape false minima,[1] and derive optimal values of S^2 and τ_e along with statistical error estimates. Our analysis shows that S^2 is a well-defined parameter that can provide reliable information about internal motion in DNA. A similar approach has been used to estimate errors in S^2 and τ_e for peptides from NMR relaxation data.[4]

METHODS AND COMPUTING PROGRAM

According to the central limit theorem, the distribution of experimental data will tend toward a normal distribution as long as the number of measurements is large. In general, a sample larger than 30 measurements is considered sufficient. However, NMR relaxation experiments are lengthy, and it is not practical to obtain such a sample (the experiments reported in ref 1 each required about a week of spectrometer

* Author to whom correspondence should be addressed.
® Abstract published in *Advance ACS Abstracts,* July 15, 1995.

0095-2338/95/1635-0803$09.00/0

Table 1. Experimental Relaxation Data for $[d(TCGCG)]_2$[a]

	90.6 MHz				125.8 MHz	
atom	MT_1	SDT_1	MNOE	SDNOE	T_1	NOE
T1,6	0.145	0.023	1.517	0.249	0.23	1.33
C2,6	0.159	0.020	1.507	0.050	0.19	1.46
G3,8	0.130	0.020	1.353	0.100	0.18	1.24
C4,6	0.131	0.014	1.583	0.133	0.17	1.26
G5,8	0.204	0.029	1.517	0.249	0.28	1.48
T1,5′	0.396	0.068	2.167	0.115	0.46	1.54
T1,4′	0.310	0.033	1.870	0.181	0.41	1.70
T1,3′	0.255	0.054	1.630	0.327	0.39	1.56
T1,2′	0.217	0.037	1.863	0.170	0.39	1.43
T1,1′	0.210	0.007	1.870	0.181	0.41	1.67
C2,3′	0.224	0.055	1.597	0.176	0.32	1.58
C2,1′	0.223	0.013	1.520	0.121	0.28	1.60
G3,5′	0.180	0.006	1.717	0.189	0.25	1.41
G3,4′	0.234	0.035	1.520	0.121	0.27	1.55
G3,3′	0.213	0.029	1.687	0.266	0.25	1.30
C4,5′	0.232	0.020	1.250	0.026	0.28	1.65
C4,4′	0.176	0.030	1.580	0.155	0.32	1.63
C4,3′	0.220	0.048	1.777	0.076	0.30	1.62
G5,5′	0.190	0.026	1.583	0.057	0.25	1.49
G5,4′	0.223	0.020	1.687	0.266	0.32	1.54
G5,3′	0.322	0.049	1.587	0.006	0.43	1.58
G5,2′	0.328	0.034	1.613	0.200	0.43	1.50

[a] The nomenclature for carbon atoms designates the nucleoside type first by letter, the chain position second, and the carbon group last. For example, C4,3′ represents the deoxyribose 3′ carbon in the cytidine located at residue 4, numbered from the 5′-end. The prefixes M and SD denote mean and standard deviation. The average relative error over all carbons for T_1 were 13% and 9% for NOE at 90.6 MHz.

time). In order to provide a reasonable estimate of the errors, we made four determinations of T_1 and three of NOE at 90.6 MHz and used the averaged relative deviations to generate normal distributions in T_1 and NOE at this field strength as well as at 125.8 MHz. The input data are given in Table 1. The error in T_1 was taken as 13% for all carbons and 9% for NOE. For each carbon at a given field strength, the mean measured T_1 value, and the average error was used to generate a normal distribution of 100 T_1 values; likewise, 100 NOE values were generated for each carbon at each field strength.

Since the overall correlation time, τ_m, should be the same for all $^{13}C-^{1}H$ vectors, τ_m was taken as a constant during optimization of S^2 and τ_e. The τ_m value was estimated from dynamic light scattering measurements on related molecules[5] to be 1.1 ns. An indistinguishable value of 1.02 ns was estimated from the experimental relaxation data (excluding the base carbons) using McMOLDYN. The csa contribution is significant for the base carbons and introduces an additional uncertainty. We have used $\Delta\sigma = 185$ ppm[1,6] as an estimate of the csa correction for the base carbons. We have made no csa corrections for the sp^3 carbons in the sugars since their distribution of electron density should be more nearly spherically symmetrical than for the sp^2 base carbons.

The automatic file reading and data processing are the basic technical differences that distinguish McMOLDYN from MOLDYN (Figure 1). The old MOLDYN reads one value per field per atom for each of T_1, T_2, and NOE and then does a series of calculations to produce an appropriate output file as requested by the user. After that, the user has to type in another set of relaxation data for the next run of calculations and reinitialize the program because the program does not save values for variables that change in the calculation process and does not restore the values held by

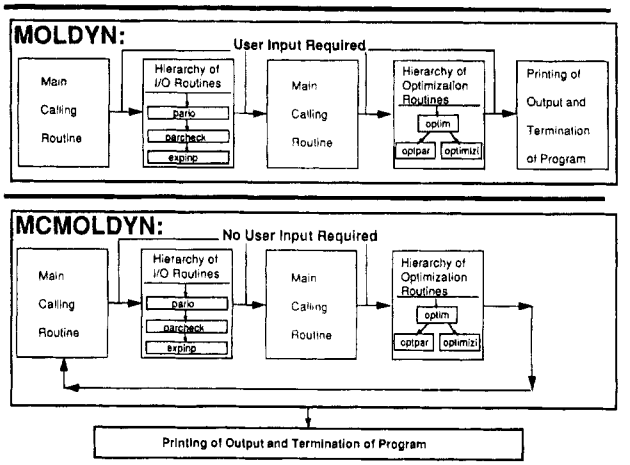

Figure 1. The normal operating sequence for MOLDYN and the new operating sequence facilitated by McMOLDYN.

variables prior to reading the input file. The new McMOLDYN can read the input relaxation data from the input data file and then does calculations just as MOLDYN would. McMOLDYN will then repeat the cycle of restoring or saving the necessary variables for the next run, reading the next set of input relaxation data and calculating new results until there are no further values to be read from the input data file. McMOLDYN can therefore read many values of relaxation data from the input data file without requiring input from the user.

McMOLDYN uses all the mechanisms of the MOLDYN program and adds two new options for dynamic studies using the Model-Free Approach, which are the Monte Carlo error estimation and correction for chemical shift anisotropy.

Monte Carlo Method. McMOLDYN generates simulated data randomly with the same mean value and standard deviation as those are obtained from experimental data or as input by the user. The program optimizes S^2 and τ_e in fitting simulated relaxation data by its nonlinear least squares subroutine. After finishing the optimization process, the program calculates the mean values and standard deviations for S^2 and τ_e. The program output of these statistical results can be used to study the errors transferred from experimental relaxation data to the parameters of the Model-Free Approach and to estimate the reliability of each parameter. In order to escape false minima in the optimization, the program can employ the search strategy described in ref 1. The optimization then automatically starts with eight combinations of initial estimates of S^2, τ_e, and τ_m.

Chemical Shift Anisotropy. McMOLDYN also contains new code for calculating chemical shift anisotropy (CSA) and cross-relaxation rates (σ_{ij}) in the model-free approach of Lipari and Szabo. The calculation of CSA is discussed in the article by Kay et al.,[7] and the equations for the (σ_{ij}) are described in the article by Eimer et al.[5] listed in references.

The source code for McMOLDYN is being prepared for submission to QCPE. It is also available on request from the authors. Our work has made some changes to the MOLDYN FORTRAN subroutines which were made specifically to simplify input and calculations for the Model-Free Approach. However, McMOLDYN should also allow a user to make custom modifications for automating input and calculations for other dynamical models.

Table 2. Parameters Describing Internal Motion in $[d(TCGCG)]_2$ with Fixed $\tau_m = 1.1$ ns

atom	S^2	ERR%	τ_e(ns)	ERR%
T1,6	0.71	12	0.077	140
C2,6	0.78	9	0.0016[a]	490[a]
G3,8	0.88	24	3.3[a]	130[a]
C4,6	0.61	62	2.2[a]	120[a]
G5,8	0.50	12	0.045	55
T1,5′	0.32	14	0.038	46
T1,4′	0.37	15	0.050	44
T1,3′	0.47	9	0.034	50
T1,2′	0.55	13	0.056	23
T1,1′	0.49	14	0.091	65
C2,3′	0.53	11	0.061	56
C2,1′	0.62	11	0.059	69
G3,5′	0.66	35	0.27	180
G3,4′	0.61	12	0.053	89
G3,3′	0.70	11	0.068	300
C4,5′	0.62	9	0.034	83
C4,4′	0.66	13	0.096	98
C4,3′	0.53	11	0.10	55
G5,5′	0.73	9	0.059	88
G5,4′	0.56	14	0.068	63
G5,3′	0.39	9	0.029	57
G5,2′	0.40	10	0.030	58

[a] Excluded in the statistical results.

RESULTS AND DISCUSSION

Table 2 summarizes the statistical results for S^2 (using $\tau_m = 1.1$ ns), reporting the means and standard deviations for the Monte Carlo simulated data set. The relative errors (ERR) for each carbon are defined as the standard deviation divided by the mean, expressed as a percentage. The mean values of S^2 are in close agreement with those published previously.[1] The average ERR in S^2 is 15%, with some especially large errors for the base carbons; the median of the distribution is 12%, and most of the values are in the range of 8−15%. If $\tau_m = 1.02$ ns is used, the S^2 values typically increase by 0.01−0.02 with ERR values nearly identical to those in Table 2. Thus S^2 should be a reliable indicator of the extent of internal motion in DNA oligomers.

To associate some physical meaning to the results, it is noted that in the model-free approach S^2 ranges from one (completely restricted internal motion) to zero (totally free internal reorientation). The average S^2 values for the sugar carbons in $[d(TCGCG)]_2$ are 0.54 ± 3% in contrast to 0.70 ± 24% for the base carbons. It is clear that the sugar internal motions are less restricted and are also more reliably determined. The data in Table 2 also quantify the relatively higher mobility of terminal residues (T1 and G5; average $S^2 = 0.50 \pm 12\%$) *vs* nonterminal residues (C2, G3, C4; av $S^2 = 0.65 \pm 19\%$).

Figure 2 presents a graphical view of the results for S^2, with the error bars indicating one standard deviation above and below the mean value for the 100 Monte Carlo simulations. A general trend that was noted previously[1] occurs in the graph: there is a general increase in S^2 for T1 from the free HO-C5′ along the chain 5′-4′-3′; the reversed order is seen for 5′-4′-3′ at the other end of the chain at G5 ending with the free C3′-OH. The error bars provided by the present analysis indicate that this trend occurs at about the level of the estimated error and must therefore be treated with caution. The figure also makes it clear that S^2 is not well-determined for some of the carbons, especially G3,5′, G3,8, and C4,6.

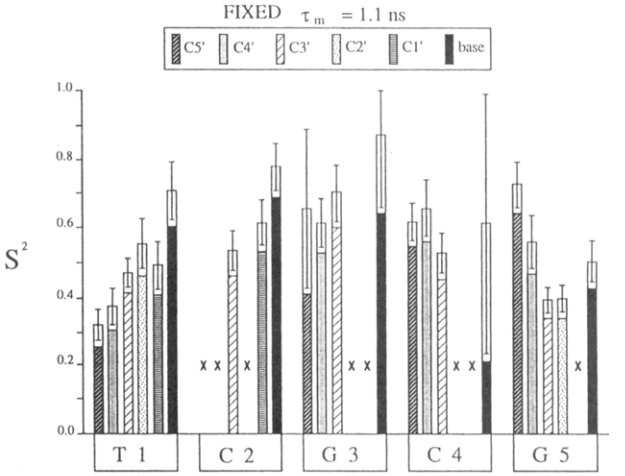

Figure 2. The distribution of S^2 in $[d(TCGCG)]_2$. The height of each bar is the mean value of S^2 corresponding to each carbon atom. The error bars show one standard deviation above and below the means. The 'x' symbols denote carbons for which overlap of signals prevented measurement of the relaxation properties.

Table 2 also summarizes the statistical analysis for τ_e (using $\tau_m = 1.1$ ns). The average value for $\tau_e = 0.070$ ns with a large average ERR = 85%. ERR > 100% for 6 of the 22 carbons. We conclude that τ_e is a poorly defined parameter, at least for this molecule, as was suggested by our earlier work. It is safest to say that τ_e occurs in the range from 0.03 to 0.3 ns, with relative errors from 40% to 150%. Although the estimated errors in τ_e are large for individual carbons, the central limit theorem suggests that random errors will tend to cancel for averages over several carbons. Thus, it may be significant that $\langle\tau_e\rangle$ 0.09 ns for the nonterminal residues, about twice as long as for the terminal residues $\langle\tau_e\rangle$ 0.05 ns. This indicates that the terminal residues move relatively faster (and have larger amplitudes, see the previous analysis of S^2) than the nonterminal residues.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant GM35069.

REFERENCES AND NOTES

(1) Borer, P. N.; LaPlante, S. R.; Kumar, A.; Zanatta, N.; Martin, A.; Hakkinen, A.; Levy, G. C. ^{13}C-NMR Relaxation in Three DNA Oligonucleotide Duplexes: Model-Free Analysis of Internal and Overall Motion. *Biochemistry* **1994**, *33*, 2441−2450.
(2) Craik, D. J.; Kumar, A.; Levy, G. C. MOLDYN: A Generalized Program for the Evaluation of Molecular Dynamics Models Using Nuclear Magnetic Resonance Spin-relaxation Data. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 30−38.
(3) Lipari, G.; 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.* **1982**, *104*, 4546−4559.
(4) Palmer III, A. G.; Rance, M.; 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.* **1991**, *113*, 4371−4380.
(5) Eimer, W.; Williamson, J. R.; Boxer, S. G.; Pecora, R. Characterization of the Overall and Internal Dynamics of Short Oligonucleotides by Depolarized Dynamic Light Scattering and NMR relaxation Measurements. *Biochemistry* **1990**, *29*, 799−811.
(6) Williamson, J. R.; Boxer, S. G. Synthesis of a thymidine phosphoramidite labeled with ^{13}C at C6: relaxation studies of the loop region in a ^{13}C labeled DNA hairpin. *Nucleic Acids Res.* **1988**, *16*, 1529−1540.
(7) Kay, L. E.; Torchia, D. A.; Bax, A. Backbone Dynamics of Proteins As Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease. *Biochemistry* **1990**, *28*, 8972−8979.

CI940059N