Journal of Biomolecular NMR, 4 (1994) 645–652
ESCOM

J-Bio NMR 199

Quantification of two-dimensional NOE spectra via a combined linear and nonlinear least-squares fit

James W. Brown and Wray H. Huestis*

Department of Chemistry, Stanford University, Stanford, CA 94305-5080, U.S.A.

Received 13 December 1993
Accepted 29 March 1994

Keywords: NMR data processing; Two-dimensional NMR peak quantification; Data fitting; Parameter estimation

SUMMARY

Determining the volumes of peaks in 2D NMR spectra can be prohibitively difficult in cases of overlapping, broad lines. Deconvolution and parameter estimation can be attempted on either the time-domain or the frequency-domain data. We present a method of estimating spectral parameters from frequency-domain data, using a combination of Lorentzian and Gaussian lineshapes for reference lines. This approach combines a previously published method of projecting the data on a linear space spanned by reference lines with a nonlinear least-squares fitting algorithm. Comparison of this method with other published methods of frequency-domain deconvolution shows that it is both more precise and more accurate when estimating 2D volumes.

INTRODUCTION

Accurate and precise estimation of the peak volumes of 2D NOESY NMR data is essential to the task of determining the 3D structure of molecules in solution (Wüthrich, 1986). Methods used to determine the relevant parameters for signal quantification can be divided into two general categories: those that estimate the relevant parameters using time-domain data, such as linear prediction (Barkhuusen et al., 1985; Stephenson, 1988; Gesmar et al., 1990), state space modeling (DeBeer et al., 1992) and Bayesian analysis (Stephenson, 1988; Gesmar et al., 1990); and those that estimate the parameters using frequency-domain data, such as projection of the 2D data onto a set of reference lines (Denk et al., 1986; Olejniczak et al., 1989) and the combination of 1D row and column integrals (Holak et al., 1987). The latter group is attractive because of their intuitive relation to Fourier-transformed 2D NMR data and because they can more accurately quantify small volumes among much larger ones (Olejniczak et al., 1989).

*To whom correspondence should be addressed.
The source codes for MATLAB routines are available via Email to jbrown@nmrlab. stanford. edu.

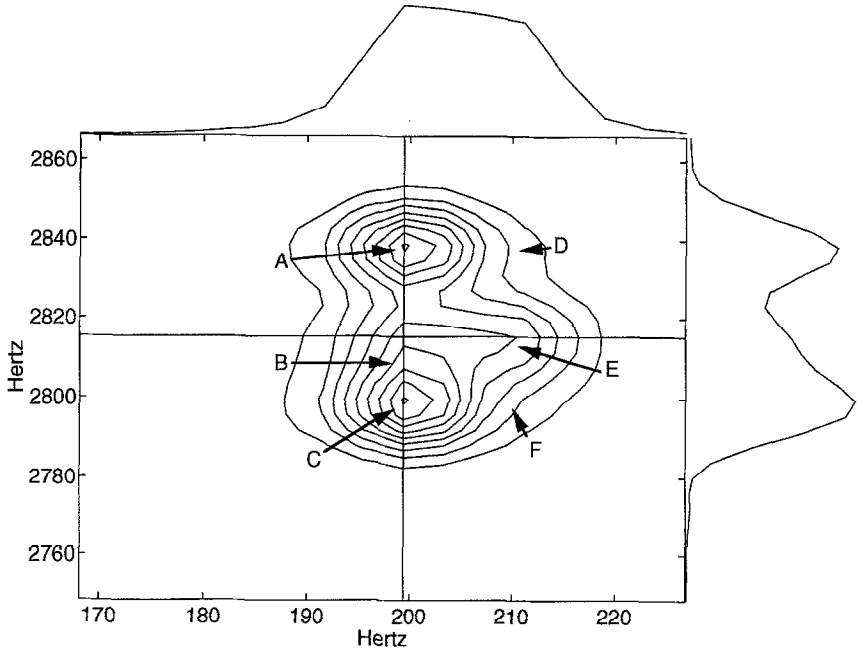

Fig. 1. Contour plot of synthetic NMR data, with representative 1D projections, used to evaluate different methods of fitting. Peaks D and F are nonexistent in the data, but the fitting methods will generate volumes for them, which should be near zero.

We present here a nonlinear least-squares fit of an entire 2D peak region as a method of quantifying peak volumes. An earlier published method (Denk et al., 1986) addressed this problem by expanding the 2D data set onto a set of reference lines via a linear least-squares fit (Denk method). Our contribution extends this approach by subjecting the parameters used to generate the reference lines in the Denk method, i.e., the chemical shifts and linewidths, to a nonlinear least-squares fit using the entire 2D data set. This method (modified Denk) results in more accurate and precise estimation of overlapping cross peaks, because the entire 2D data set is used to determine the linewidth as well as the intensities. The method also demands less user involvement, since it does not require user-determined reference lines, only initial estimates of the parameters that describe those reference lines. In addition, baseline correction is possible and arbitrary lineshapes can be used in cases where apodization has removed the Lorentzian character of the lines.

METHODS

The usefulness of our fitting procedure was tested by construction and analysis of a data set with known spectral parameters and comparison of this method with other methods of frequency-domain 2D fitting. The synthesis of the known data set and the fitting for all methods was done using the MATLAB 4.1 software, including the Optimization Toolbox (The Mathworks, Natick, MA), on a Sun SPARCstation 2. The MATLAB software and its associated Toolboxes are well

suited for NMR data analysis, as reported by other researchers (Van Tongeren et al., 1992). The synthetic data were generated using Eq. 1:

$$\mathbf{D} = \mathbf{F}_2\mathbf{X}\mathbf{F}_1 \tag{1}$$

where **D** is the synthetic data matrix, $\mathbf{F}_1$ and $\mathbf{F}_2$ are 2048×5 and 5×512 matrices of synthetic FIDS, respectively, and **X** is a 5×5 matrix of peak intensities, given by Eq. 2. The synthetic FIDs consisted of exponentially damped cosines. Both $\mathbf{F}_2$ and $\mathbf{F}_1$ spanned an 8000 Hz spectral width. The resulting matrix **D** was a 2048×512 real matrix. The first point of each row of **D** was multiplied by 0.5 to remove the baseline offset.

$$\mathbf{X} = \begin{pmatrix} 10 & 0 & 6 & 4 & 6 \\ 0 & 5 & 0 & 3 & 0 \\ 6 & 0 & 20 & 0 & 0 \\ 4 & 3 & 0 & 30 & 0 \\ 6 & 0 & 0 & 0 & 20 \end{pmatrix} \tag{2}$$

Each row of **D** was then apodized by a 60° phase-shifted squared sine-bell function and Fourier transformed. The last 1024 points of each row were discarded. The first point of each column of **D** was then multiplied by 0.5 to remove the baseline offset. Each column of **D** was zero-filled to 2048 points, apodized by a 60° phase-shifted squared sine-bell function, and Fourier transformed. The last 1024 points of each column were discarded. The resulting synthetic data set was a 1024×1024 matrix, spanning 4000 Hz in each dimension. A contour plot of the region to be analyzed and representative slices from each dimension are shown in Fig. 1.

The algorithm used to determine peak parameters is illustrated in Fig. 2. Equation 3 was used to generate the set of reference lines, $R_{n,m}$, where

$$R_{n,m} = f_{n,m}L + (1 - f_{n,m})G \tag{3}$$

n is the index of the reference line in dimension m, $f_{n,m}$ is the fraction of Lorentzian character of line n in dimension m, and G and L are given by Eq. 4,

$$L = \frac{1}{1 + z^2}$$

$$G = \exp(-(\ln 2)z^2) \tag{4}$$

$$z = \frac{s_{n,m} - x}{\frac{1}{2}w_{n,m}}$$

where $w_{n,m}$ and $s_{n,m}$ are the full width at half height and the chemical shift, respectively, of line n in dimension m (Gans, 1992). The matrix of intensities, **H**, was generated in MATLAB by the command shown in Eq. 5.

$$\mathbf{H} = \mathbf{R}_1\backslash((\mathbf{R}_2\backslash\mathbf{D}')') \tag{5}$$

The MATLAB operator '\', when used in a statement such as X = B\A, computes the solution, in

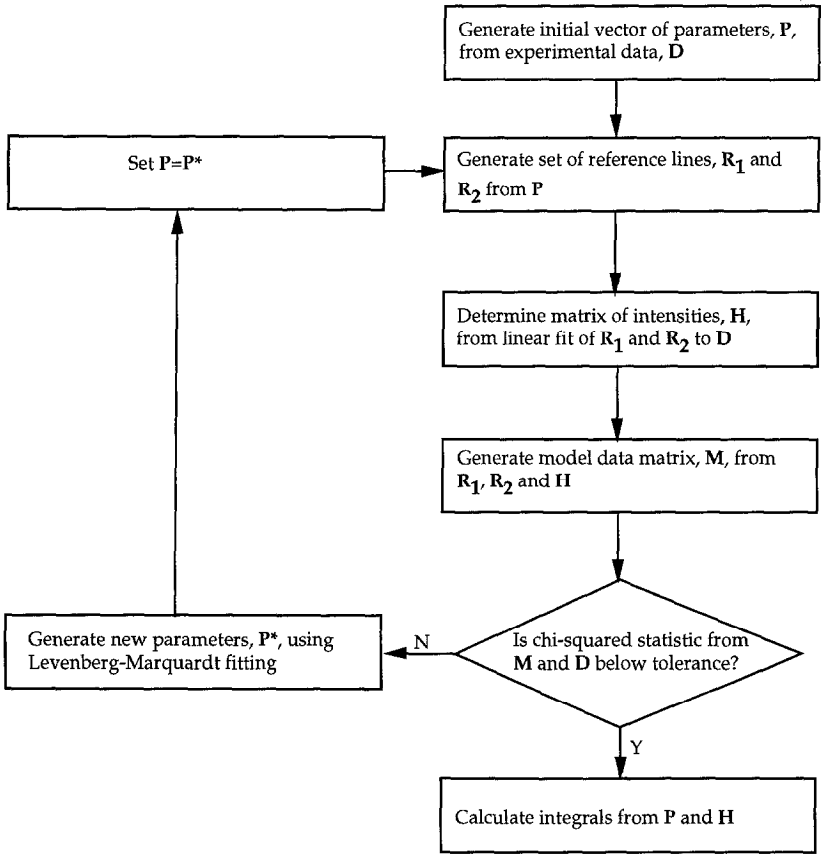

Fig. 2. Flow chart showing the fitting algorithm used to determine spectral parameters of 2D data.

the least-squares sense, to the overdetermined system of equations AX = B (The Mathworks, 1993). The model dataset, **M**,

$$\mathbf{M} = \mathbf{R}_1\mathbf{H}\mathbf{R}_2 \tag{6}$$

was then calculated by Eq. 6. The quantity **D**−**M** was then passed to the MATLAB function *leastsq* for subsequent nonlinear least-squares minimization with respect to the vector of parameters, **P**, which is the complete set of $f_{n,m}$, $s_{n,m}$ and $w_{n,m}$. The Levenberg–Marquardt minimization algorithm was used in this exercise (Grace, 1992). In cases where the fit generated negative values for f, the MATLAB function *constr*, an algorithm that uses Sequential Quadratic Programming (SQP) for constrained optimization (Grace, 1992) was used. In the constrained case, all elements of the vector **P** were constrained to be greater than zero and the set of $f_{n,m}$ were also constrained to be less than one. Finally, the 2D integral was calculated using Eq. 7,

$$I = hw_1w_2\left[f_1\pi + (1-f_1)\sqrt{\pi}\right]\left[f_2\pi + (1-f_2)\sqrt{\pi}\right] \tag{7}$$

where h is the intensity, w_1 and w_2 are the full widths at half height in each dimension, and f_1 and f_2 are the fractions of Lorentzian character in each dimension. The factors π and $\sqrt{\pi}$ in Eq. 7 represent the integral of a Lorentzian and a Gaussian, respectively, each of unit height and width.

We compared four different methods of calculating the volume integrals of the model data in Fig. 1. Two different variations of the method of Denk et al. (1986) were used, one in which the lineshape was assumed to be Lorentzian and another in which the lineshape was taken to be a linear combination of a Gaussian and a Lorentzian. In each case the parameters for the reference lines were generated from a least-squares fit of a selected 1D slice. The slices were selected such that the error in the width was a minimum. Volume integrals for this method were calculated using Eq. 7. We also used the combination of 1D integrals (Holak et al., 1987) to quantify peak volumes. Appropriate 1D slices were again chosen to minimize the error in the fitted parameters. Volume integrals were calculated as in Holak et al. (1987). For the modified Denk method presented here, the initial estimates of the nonlinear parameters were generated after inspection of the sum of the rows and columns.

RESULTS AND DISCUSSION

Analysis of the constructed data set

The results for the four methods of volume integration are compared in Table 1. The errors in the estimation of integrals were calculated using standard error propagation formulae (Gans, 1992). True integrals of individual peaks were estimated using the Holak, Denk, and modified Denk methods to fit a 2D data set generated by using Eqs. 1 and 2, with **X** having all nondiagonal elements set to zero except for the peak of interest. For each peak determined individually, all three methods yielded the same results.

Combination of 1D integrals, i.e., the method of Holak et al., accurately determines the integral of line E of Fig. 1, where there are no overlapping lines, but is poor in estimating the integrals of

TABLE 1
COMPARISON OF RESULTS OBTAINED WITH VARIOUS METHODS OF VOLUME INTEGRATION[a]

Cross peak	Holak et al.[b]		Denk et al.[c]		Denk et al.[d]		Modified Denk method[e]		True Integral
	Integral	Error	Integral	Error	Integral	Error	Integral	Error	
A	542.42	125.34	1159.66	507.71	661.46	165.83	717.42	59.53	731.49
B	807.71	233.95	799.84	619.16	618.03	168.81	555.81	69.83	520.18
C	675.05	96.35	1285.81	388.58	714.74	127.11	750.10	54.66	748.50
D	40.45	129.74	−235.28	118.55	−15.45	−5.70	−9.83	−2.53	0.00
E	396.03	129.14	831.50	661.65	393.38	90.58	402.80	42.33	377.16
F	9.94	173.35	−109.98	53.40	−4.95	−4.84	−5.50	−2.48	0.00

[a] Values in arbitrary units. Fitted chemical shifts: A: 200, 2838 Hz; B: 200, 2814 Hz; C: 200, 2797 Hz; D: 210, 2838 Hz; E: 210, 2814 Hz; F: 210, 2797 Hz. Errors are reported as two standard deviations (95% confidence level).

[b] Holak et al. (1987): combination of 1D integrals.

[c] Denk et al. (1986): pure Lorentzian lineshape.

[d] Denk et al. (1986) linear combination of Gaussian and Lorentzian lineshapes.

[e] Method presented in this paper: nonlinear fit.

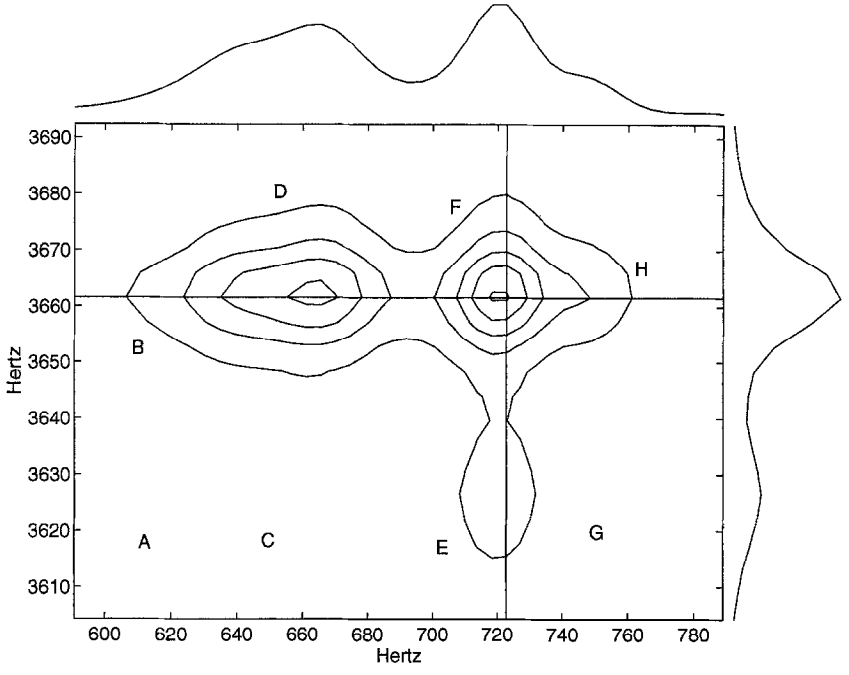

Fig. 3. 2D NOESY spectrum, with representative 1D projections, of 100 mM DMPC/25 mM AFA*t*Bu in D_2O. The region shown contains connectivities between the phenylalanine ring protons and the DMPC acyl chain methylene, peptide *tert*-butyl, Ala-[1] methyl and Ala-[3] methyl protons.

overlapping lines A, B and C. The main failure of this method is its overestimate of the linewidth of the 2814 Hz reference line (data not shown), resulting in an overestimate of the volume integral, especially in comparison to lines A, C and E. The method of Denk overestimates the value of the integrals, and is relatively imprecise when a pure Lorentzian lineshape is applied. On using a linear combination of a Lorentzian and Gaussian shape, the results improve significantly. Overall, the modified Denk method presented here performs better than the Denk and Holak methods, both in precision and accuracy. The modified Denk method is more precise because the entire 2D region is used to fit the widths of each peak, as opposed to using a single 1D slice. This increased precision is obtained at the price of increased computational complexity of performing a nonlinear fit to the entire 2D data set.

TABLE 2
2D INTEGRALS FROM ISOLATED CROSS PEAK[a]

Method	Integral	Error
Holak et al.	54.66	39.04
Denk et al.	55.56	25.34
This work	35.93	7.81

[a] Values in arbitrary units. A linear combination of Lorentzian and Gaussian lineshapes was used in each method. Errors are reported as two standard deviations (95% confidence level).

Analysis of an isolated cross peak

Since the analysis of model data is notorious for its ability to hide deficiencies in numerical methods in the presence of nonideal noise, we analyzed an isolated cross peak with the three methods described above. The peak was taken from one of our 2D NOESY spectra presented elsewhere (Brown and Huestis, 1993). The results from this analysis are presented in Table 2. As in the case of the constructed data set, our method proved to be more precise than the two previously presented methods.

Analysis of the 2D NOESY spectrum of a lipid–peptide complex

To demonstrate this method, we attempted to quantify a portion of the 2D NOESY spectrum, shown in Fig. 3, of a dimyristoylphosphatidylcholine (DMPC) alanine-phenylalanine-alanine-*O*-*tert*-butyl ester (AFA*t*Bu) complex in D_2O (Brown and Huestis, 1993). This region of the spectrum shows the cross peaks between the aromatic protons and the lipid acyl chain methylene, *tert*-butyl, Ala^1 methyl and Ala^3 methyl protons. The solution was sonicated such that the phospholipid molecules formed small, unilamellar vesicles of approximately 2000 kDa. An aggregate of this size exhibits very broad proton NMR lines, and quantification of the peaks cannot be achieved in the usual way, i.e. by numerical integration above a baseplane. Table 3 shows the results of volume integration of the peaks in Fig. 3, using the modified Denk method. Although the errors in the integrals of the non-zero cross peaks can be rather high, i.e., over 30%, our method provides a way to calculate relatively accurate and precise volume integrals of this severely overlapping region of the spectrum.

CONCLUSIONS

In summary, we have presented a modification of the method of Denk et al. for estimation of 2D volume integrals, which uses the entire data set for the estimation of the nonlinear parameters that describe the peak. This modification improves the accuracy and precision with which 2D volumes can be measured. The algorithm may not be robust enough to simultaneously simulate many hundreds of peaks present in more complex protein spectra, but as more computer power

TABLE 3
2D INTEGRALS FROM THE DMPC–AFA*t*Bu COMPLEX[a]

Cross peak	Integral	Error
A	226.14	79.71
B	979.60	137.35
C	73.76	34.54
D	586.09	194.60
E	271.54	98.89
F	865.90	147.83
G	−6.19	−5.21
H	181.33	66.53

[a] Values in arbitrary units. Fitted chemical shifts: A: 641, 3624 Hz; B: 641, 3662 Hz; C: 667, 3624 Hz; D: 667, 3624 Hz; E: 721, 3624 Hz; F: 721, 3662 Hz; G: 749, 3624 Hz; H: 749, 3662 Hz. Errors are reported as two standard deviations (95% confidence level).

becomes available and minimization routines improve, this may be possible. We feel that the method will be useful in obtaining quantitative information from overlapping 2D NOESY data and hence will allow better structure determination for large molecules and molecular aggregates.

ACKNOWLEDGEMENTS

This work was supported by Grant HL 23787 from the USPHS.

REFERENCES

Barkhuusen, H., DeBeer, R., Bovee, W.M.M.J. and Van Ormandt, E. (1985) *J. Magn. Reson.*, **69**, 180–186.
Brown, J.W. and Huestis, W.H. (1993) *J. Phys. Chem.*, **97**, 2967–2973.
DeBeer, R., Van Ormandt, D. and Pijnappel, W.W.F. (1992) *Pure Appl. Chem.*, **64**, 815–823.
Denk, W., Baumann, R. and Wagner, G. (1986) *J. Magn. Reson.*, **67**, 386–398.
Gans, P. (1992) *Data Fitting in the Chemical Sciences*, Wiley, Chichester.
Gesmar, H., Led, J.J. and Abildgaard, F. (1990) *Prog. NMR Spectrosc.*, **22**, 255–288.
Grace, A. (1992) *Optimization Toolbox for Use with MATLAB*, The Mathworks, Inc., Natick, MA.
Holak, T.A., Scarsdale, J.B. and Prestegard, J.H. (1987) *J. Magn. Reson.*, **74**, 546–549.
The Mathworks, Inc. (1993) *MATLAB Reference Guide*, Natick, MA.
Olejniczak, E.T., Gampe Jr., R.T. and Fesik, S.W. (1989) *J. Magn. Reson.*, **81**, 178–185.
Stephenson, D.S. (1988) *Prog. NMR Spectrosc.*, **20**, 515–595.
Van Tongeren, B.P.O., Mehlkopf, A.F., Van Ormandt, D., Schramp, M. and DeBeer, R. (1992) *J. Magn. Reson.*, **100**, 619–624.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.