NEW SOFTWARE METHODS IN NMR SPECTROSCOPY

C. L. Dumoulin, G. C. Levy, and
the Staff of the NMR and Data Processing Laboratory

Department of Chemistry
Syracuse University
Syracuse, New York 13210

I. INTRODUCTION

Over the past twenty years, NMR spectroscopists and instrumentation have not kept pace with rapidly advancing computer technology. Even today, as instrumentation incorporates significant new computing power, advances in digital hardware and software are accelerating beyond their application to NMR.

In what ways can these increased capabilities affect the way a spectroscopist analyzes the results from an experiment? This paper will address that question with illustrations of several data analysis techniques for the optimal extraction of information from NMR experiments.

Of all the growing areas of computer technology, none is progressing with as much speed and momentum as the field of hardware. Von Neumann computer architecture has been successful for the past 40 years due to (at least in part) the nature of hardware technologies and the continuous growth of these technologies. In fact today, computer architectures radically different from the von Neumann design are being implemented. Rapid development of electronic hardware is also responsible for making computers accessible to the general population and to the laboratory. Computing capability that cost millions of dollars twenty years ago, is today available to the individual researcher.

Current computer systems are substantially different from those used on NMR spectrometers only 5 years ago. CPUs, disks, memory and terminals are much faster. The drop in memory and disk prices has been so precipitious that memory and disk storage can no longer be considered limiting factors when designing a system. Far less expensive and new types of electronic components have resulted in new system designs. Modern NMR spectrometers routinely distribute computing functions among multiple processors; each performing a specific task. Array processors, for example, have been integrated into spectrometer systems for the explicit purpose of accelerating vector operations such as Fourier Transformations. Processors for data acquisition, data reduction, user I/O and spectrometer control can be found in most modern instrumentation. Performing data reduction on one or more computers independent of the data acquisition computer has advantages which will be discussed later.

Software developments are just as important as hardware developments, but the general consensus is that hardware advances have out-paced improvements in software. Nevertheless, substantial growth in software methods has been realized in the past few years. Until recently, operating systems were by necessity unique to each type of computer. UNIX is an operating system that has been designed to be machine independent and flexible enough to address the needs of a wide range of users. It is likely that some commercial NMR spectrometer systems will include UNIX in the near future. Standard high level computer languages were utilized long before standard operating systems; most NMR spectrometer systems are now programmed at least in part in PASCAL. This enables manufacturers to upgrade spectrometer performance with newer and more powerful computers without losing their investment in previously developed software. From the user's point of view, however, not much is gained with higher level languages unless the software can be modified. Since the manufacturers are justifiably reluctant to release source code, they generally accomodate users by supporting user-written modules.

More powerful computer hardware is resulting in more powerful and sophisticated software. Faster computers have made practical new algorithms for spectroscopic data processing. These new algorithms tend to heavily utilize statistics and thus be more rigorous and complicated. Fast floating point performance and large memory address spaces are design assumptions for these algorithms. A natural extension of this programming methodology is the application of artificial intelligence techniques to the processing of spectroscopic data.

This paper discusses only data reduction software; it does not address problems in optimizing data acquisition. Examples of sophisticated software developed at our laboratory will be presented in the next few sections. This software has been developed to process NMR data, but the techniques in use are applicable to a wide range of spectroscopies and chromatography. The software is designed to run on 32-bit mini or micro-computers (DEC VAX series, Data General MV series, IBM instruments CS-9000) in single or multiuser configurations. This software is part of a new generation of laboratory data processing methods. While the individual techniques employed are more evolutionary than revolutionary, the software as a whole represents a significant advance in laboratory data processing.

II. AUTOMATION

One of the most rapidly advancing fields of analytical chemistry is laboratory automation. It is generally recognized that efficiency, quality and cost effectiveness of many laboratory tasks can be greatly increased through the successful application of automation. FT-NMR spectroscopy was one of the first experimental techniques to benefit from automation due to the requirement of an instrument computer. Modern spectrometers are capable of running a series of very complicated experiments without operator attention. Even automatic sampling changing has been demonstrated at several laboratories. Fully automated data analysis is more difficult to accomplish, however, especially if that analysis is to include complicated tasks which the user would normally control based on the results of previous analysis steps.

The ease of automating a program is inversly proportional to the complexity and power of that program. Thus, in principle, it is relatively easy to automate simple software systems designed for the unsophisticated user. These automated and simplified systems are becoming increasingly popular. Of course, automating such a program does not afford the experienced spectroscopist much of the power or flexibility found in more sophisticated data analysis programs. Simple automated data processing, including data collecting, weighting or apodization, Fourier Transformation, and plotting) is much easier to realize than automated comprehensive data analysis (i.e. data processing, identification and characterization of all spectral peaks, quantification, evaluation, etc.).

Many of the steps taken in a typical processing sequence are linear and require only simple input. For example, an exponential weighting is generally performed in such a way that the user needs only to select a line broadening parameter. The success of processing is not extremely sensitive to the proper choice of this parameter. In fact, line broadening can be determined from one spectrum and then applied indiscriminately to all similar spectra. Phasing, on the other hand, is a linear process which is extremely sensitive to input parameters. Unless manually phased calibration spectra or phasing alternatives such as magnitude calculations are used, complicated algorithms are needed to reliably automate spectral phasing.

Automatic phase correction of Fourier transform NMR spectra was first demonstrated by Ernst (1). His iterative method arrives at optimum values for zero and first order corrections by taking linear combinations of the original spectrum and its Hilbert transform in various proportions until the ratio of the maximum positive signal excursion to maximum negative signal excursion is optimized. The applicability of the procedure is limited by spectral signal-to-noise and the accuracy with which the baseline can be characterized. Another iterative technique (2) optimizes phase corrections by a modified simplex method, using maximization of the smallest spectral absorption mode intensity as the optimization criterion. Limited automatic phase correction has also been achieved by precalibration of the detector transfer function (3). An automatic phasing technique developed in our laboratory (4) is based on Dispersion vs. Absorption (DISPA) lineshape analysis and is summarized below.

A plot of "absorption" versus "dispersion" (normalized to the maximum apparent "absorption" peak height) for a misphased Lorentzian line gives a unit circle passing through the origin, but which has been rotated about the origin by a number of degrees equal to the applied phase misadjustment. Fortunately, phase misadjustment appears to be the only mechanism (of those examined to date (5-9)) that displays this type of DISPA behavior. To automate phasing using this method, a robust and non-interactive peak indentification scheme must be used to locate each peak. We have found that locating peaks in the power spectrum gives the best results. The peak finding procedure, described elsewhere (10), tabulates peak location and peak phase information. A weighted linear least squares calculation is then performed to obtain the zero and first order phase corrections.

The DISPA phasing method has the advantage that it is non-interactive. Furthermore, a large number of peaks actually improves the quality of the spectral phase correction. Thus, best results are obtained on high resolution - narrow line spectra. A spectrum phased with this method is shown in figure 1.

Automatic phasing is the most difficult primary processing step to automate. Automated data analysis operations, however, are even more difficult to implement because of their non-linear nature. Furthermore, data analysis operations are characterized by extensive user interaction (and consequently user decision-making). To automate baseline calibration, spectral peak quantification (location, linewidth, intensity and integral) and final data presentation requires software capable of making decisions normally made by the user.

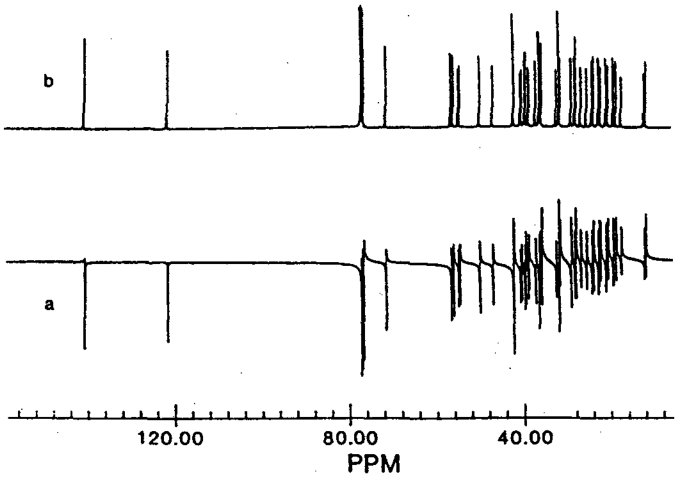

Figure 1. ^{13}C spectrum of a mixture of two steroids (a) before phasing and (b) after automatic phasing.

The peak quantification algorithms which we are using have multiple levels of operation - varing from total user control to fully automatic execution. The controlling software for automated data analysis has two manifestations. The first mechanism allows the user to string together processing commands which have only simple input (such as system parameters). The user can modify processing parameters during a "Dry Run" of the command execution if desired. Provisions exist to increment filenames or to internally loop within the command string.

A second mechanism is currently under development. It will allow a user familiar with the software system to train (or program) an automation sequence by processing, analyzing and generating output for a data file or a series of data files. During this training session, the trainer user can specify the required output, input and options that will be used for automated analysis. The end result of the training session will be a control file which a less sophisticated user can invoke to control processing and analysis for his own data sets. This automated sequence can include command strings (via the first mechanism described) or other automated sequences. Customized messages within the sequence will make it easy for users with all levels of spectroscopy experience to perform repetitive analyses to the full extent of their ability. This level of automation is not restricted to simple operations and, in fact, can make full use of the sophisticated subroutines such as the peak quantitation and curve fitting routines described below.

III. QUANTITATIVE ANALYSIS

One of the most important features of NMR spectroscopy is its quantitative nature. Unfortunately, the inherent low sensitivity of the technique usually forces the spectroscopist to obtain data under non-linear conditions, particularly for nuclei other than proton. Recent advances in instrument sensitivity, however, are making quantitative experiments increasingly practical.

The requirements for quantitative data acquisition are well known (11-15). Instrument and experiment parameters such as filter settings, pulse width, pulse repetition rate and decoupling all play important roles in establishing a quantitative experiment. In addition, the chemical system itself can influence the linearity of the experiment if magnetization is transfered between nuclei after excitation. Quantitative experiments may not always be possible and usually they require conditions which reduce spectral signal-to-noise ratios.

Regardless of whether or not an experiment has been performed under analytical conditions, the highest accuracy possible for peak integrations is desirable. A wide variety of integration techniques exist and have been applied to NMR data. Continuous wave instruments use a capacitor to electrically integrate spectral resonances. FT NMR data systems, on the other hand, allow the user to plot an integral trace superimposed on the spectrum. Since most NMR spectral baselines are not perfect, the user must adjust the trace (generally with interactive knobs) to insure that the integral is constant in regions without signals. The integral trace is usually generated by determining the running sum of all previous points. It is up to the user to measure the trace by hand to acquire the desired integrals. The user can introduce his bias at two points in the integral (baseline) determination - during the integral alignment phase and during the actual measurement. In particular, if proper procedures are not followed for high dynamic range spectra, very incorrect results can be obtained.

Clearly it is advantageous to remove the user from the integration mechanism altogether. However, it is not trivial to program a computer to automate this procedure. Numerical integration methods such as summing, trapezoidal rule and Simpson's rule are straightforward. Unfortunately, proper identification and correction of the baseline prior to integration is critical to the correct operation of all these techniques. In addition, the choice of integration limits can greatly affect the correctness of the results.

A number of mechanisms exist for the removal of distorted baselines. For example, the convolution difference method (16) removes broad features from a spectrum by subtraction after applying an appropriate weighting function in the data's Fourier co-domain. Successful baseline correction, however, is dependent on the proper choice of the ratio of narrow-to-broad component intensities. In practice, the convolution difference procedure is an iterative one in which the user performs the optimization.

A technique which we have found to be superior to all others can be performed automatically or interactively. The first step of the algorithm identifies baseline points by calculating the standard deviation for all points less than the previous estimate of the standard deviation. An n^{th} order polynomial (n = 0 to 9) is then fit to all the points identified as baseline. This polynomial function is subtracted from the data and the whole process is repeated until the coefficients of the polynomial approach zero.

The algorithm has provisions for user intervention if necessary. For example, the user can choose the order of the polynomial. The user can also delimit up to 20 regions within the data set which will be independently corrected. In another mode, regions which contain many peaks and thus should not contribute to the calculation of the polynomial can be identified. In this case, the correction applied to these regions is determined only by baseline characteristics outside the selected regions.

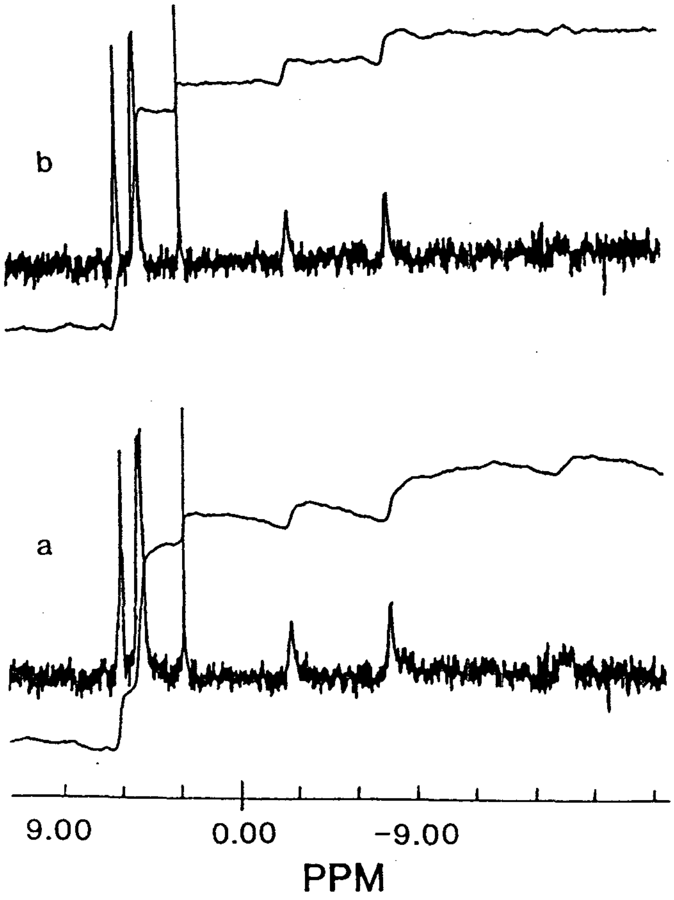

Figure 2. ^{31}P spectrum of human erythrocytes in Tris-ringer's buffer (a) before baseline flattening and (b) after a 4th order, 4 block baseline flattening.

This baseline flattening algorithm works well under most conditions. Figure 2 shows a low signal-to-noise ratio spectrum.

Simpson's rule integration is a powerful method for integrating well separated peaks. Peaks or interest often overlap, however, and alternative integration methods then become necessary.

IV. CURVE FITTING

The form of NMR spectral resonances can often be accurately described by one of several mathematical equations. The mathematical description for an entire spectrum is simply the sum of all the appropriate equations. For example, high resolution spectra of liquid samples (assuming no field homogenity distortion) can be described as a sum of several Lorentzians, one for each resonance. Since each Lorentzian peak is defined by three parameters (linewidth, intensity and position), an entire spectrum of N peaks can be defined with 3N parameters. Other NMR spectral lineshapes exist which may require more than 3 parameters. Thus, the total number of parameters needed to completely describe a spectrum depends not only on the number of peaks but also on the assumed functional form.

In a general sense, the whole purpose of spectral data analysis is to learn the number of spectral peaks and calculate all the parameters for all the peaks. In the case of well resolved spectra, the numerical techniques previously described are perfectly adequate for spectral parameterization. These techniques make no assumption of spectral form and thus work well under most conditions. When these techniques are applied to partially resolved spectra, however, accurate determination of individual peak locations, linewidths, intensities and integrals becomes difficult, if not impossible. By assuming a spectral lineshape, non-linear regression analysis techniques can be used to parameterize a data set. We have developed and are using in our laboratory a non-linear least squares technique based on Brown's modification of the Levenberg-Marquardt algorithm to analyze spectral data.

Non-linear least square techniques can be complex and much work has been done in the development of these algorithms (17-19). If we assume that the ith data point of a spectrum is $\mathbf{X_i} = \mathbf{R_i} + \mathbf{e_i}$ where $\mathbf{R_i}$ is the true value for point **i** and $\mathbf{e_i}$ is a random error with a Gaussian distribution, then a function $\mathbf{f(A,B,C.....)}$ exists such that the error function

$$E = \sum_{i=1}^{N} \left[f(A,B,C....) - X_i \right]^2$$

is minimized. Function **f** is simply the mathematical form of the spectral lines (Lorentzian, Gaussian, Voigt, their combinations, etc.). The independent variables **A,B,C...** represent the spectral quantities associated with the lineshapes of interest. At the global minimum of the error function, these variables provide the best approximation to the experimental data. The goal of a non-linear optimization algorithm is to efficiently find the global minimum of the error function while avoiding all other local minima.

The global minimum of the error function is a point in multi-dimensional space (where A,B,C... are the dimensions). Because the parameters are continuous, an exhaustive search is fundamentally impossible. Even if an exhaustive search could be performed it would be unacceptable in that calculations could take years on the fastest mainframe computers. Useful algorithms require that the error function be evaluated at only a relatively small number of points within the multidimensional space. From an examination of the local region surrounding a current point, an estimate is made for the optional direction and distance to travel to the next parameter point to be examined. The error function is reevaluated and the entire process is repeated until terminating criteria are satisfied. It is prudent to constrain the fit whenever possible so that the calculations are performed within a bounded region. thus, unreasonable values for some variables (for example, negative linewidths) can be prevented.

Determining which points to evaluate in multidimensional error space is the most difficult aspect of non-linear optimization. Good initial estimates of parameters are vital if the convergence is to be on the global minimum. Fortunately, the quantification algorithms described earlier usually provide reasonable initial estimates without user intervention. Many strategies for the location of maxima and minima of multidimensional functions have been shown to be successful (17-20) although no single algorithm is always optimal. The Taylor series method extrapolates to the nearest minimum using a Taylor series expansion on points within a small region of the error function. The Newtonian method is a 1 dimensional Taylor expansion applied one axis at a time. A more efficient Taylor expansion can be performed in all dimensions using a matrix of second partial derivatives. The gradient method, on the other hand, calculates the multi-dimensional gradient and extrapolates to the nearest minimum. Both methods reevaluate the error function at the extrapolated point and iterate until convergence at the minimum is obtained.

When convergence occurs, the Taylor series and gradient methods will usually converge on the same region, but follow very different paths. The Levenberg-Marquart method combines the Taylor series and gradiant methods to minimize the disadvantages of each. During optimization, this method selects new points through a suitable combination of the Taylor series and gradient method. The extent of combination is determined by the "Marquardt parameter" which is recalculated for each iteration.

Several terminating criteria can be used to stop the calculation. These include:

1. A preset number of iterations have been executed.
2. The proportional difference of successive error function evaluations falls below a preset valve.
3. The proportional difference in successive approximations of all the parameters becomes less than some preset valve.
4. The gradient defined by successive points in the multi-dimensional space becomes less than some preset valve. (Requires normalized parameters).

We have found that criteria 3 has the best performance when the approximations are close to the global minimum. If criteria 1 is met (we use 100 iterations as the limit), the user is given the opportunity to continue the calculation.

Figure 3 shows a spectrum with overlapping resonances which cannot be quantified by ordinary numerical methods. Curve fitting was employed to resolve the data.

Integration of the Lorentzian equation is straightforward using integral calculus techniques. Analytical integration of other lineshapes, however, can be non-trival. We use standard integration algorithms to numerically integrate these functions.

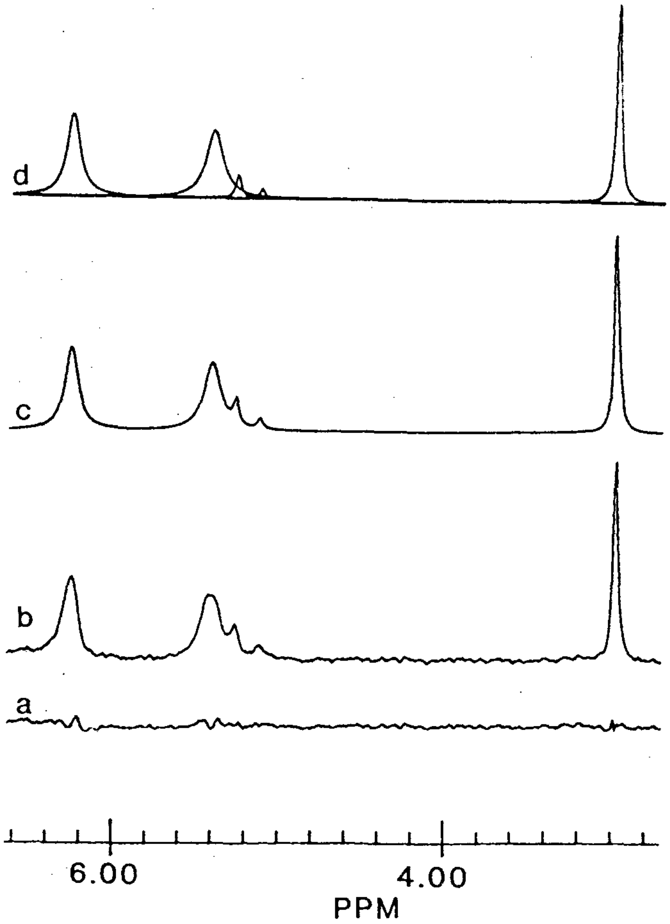

Figure 3. Downfield region of a ^{31}P spectrum of human erythrocytes after baseline flattening: (a) difference between the calculated and experimental spectra; (b) experimental spectrum; (c) calculated spectrum; (d) the individual components.

Even when the fit is very good, care should be taken when interpreting the results. Fitting more functions to the data than reasonably exist can lead to incorrect results despite good optimization and convergence. For example, a broad Gaussian peak can be approximated by two smaller overlapping Lorentzian peaks. This sort of over parameterization is one of the most serious problems in automating a curve fitting processing sequence. The number of peaks and lineshape of each peak must be correctly determined to insure a meaningful fit.

V. CONCLUSION

Large software systems such as the one developed in our laboratory are the product of many man-years of work. Furthermore, any large software system requires maintenance, particularly if new capabilities and features are being continually added. Clearly, in-house development of large software systems is not in the best interest of most chemical laboratories. As the size and sophistication of large systems grow, fewer laboratories will be able to afford the necessary development effort. If software designed in a few laboratories is to be useful to the scientific community as a whole, then that software must be: (1) machine independent; (2) well documented; (3) rigorously correct; and (4) continually supported.

Careful attention has been paid to the constraints of large software system design in the development of our NMR Spectroscopy data analysis software system. Over 20 academic institutions are using the initial release NMR1 (previously named ORACLE - which is a registered trademark of the ORACLE Corporation). It has been implemented on three families of 32 bit computers and supports 10 different graphics devices. Other software systems under development in our laboratory include a 2 dimensional NMR data processing system and systems for processing data from other spectroscopics and chromatography.

Space does not permit us to list all of the people who have contributed to NMR1. Special acknowledgements, however, belong to the following people: Salar Abosaiedi, Dan Aquilino, John Begemann, Joel Diamond, Judy Harrington, Sener Ilgen, William Kahn, Anil Kumar, Alex Macur, S. Manogaran, Mark Newsham, Chris Paul, Chris Sotak, and Pierre Wellner. We would also like to acknowledge support from the NIH Division of Research Resources (Grant RR 01317) and the National Science Foundation (Grant CHE 81-05109).

References

[1]R. R.. Ernst, J. Mag. Res. 1, 7 (1969).

[2]M. M. Siegel, Anal. Chem. Acta 133, 103 (1981).

[3]B. L. Neff, J. L. Ackerman, and J. S. Waugh, J. Mag. Res. 25, 335 (1977).

[4]C. H. Sotak, C. L. Dumoulin and M. D. Newsham, J. Mag. Res. Summitted.

[5]A. G. Marshall and D. C. Roe, Anal. Chem. 50, 756 (1978).

[6]D. C. Roe, A. G. Marshall, and S. H. Smallcombe, Anal. Chem. 50, 764 (1978).

[7]A. G. Marshall and D. C. Roe, SJ. Mag.Res. 33 551 (1979).

[8]A. G. Marshall, J. Phys. Chem. 83, 521 (1979).

[9]F. G. Herring, A. G. Marshall, P. S. Phillips and D. C. Roe, J. Mag. Res 37, 293 (1980).

[10]C. L. Dumoulin, and G. C. Levy in NMR of the New Nuclei: Chemical and Biochemical Applications, P. Laszlo ed., in press.

[11]C. H. Sotak, C. L. Dumoulin and G. C. Levy, Anal. Chem. 55, 782 (1983).

[12]J. N. Shoolery, Prog. NMR Spectros. 11, 79 (1977).

[13]T. H. Mareci, and K. N. Scott, Anal. Chem. 49, 2130 (1977).

[14]B. Thiault and M. Mersseman, Org. Mag. Res. 8, 28 (1976).

[15]J. W. Blunt, and M. H. Munro, Aust. J. Chem. 29, 975 (1976).

[16]J. C. Lindon and A. G. Ferrige, Prog. NMR Spec. 14, 27 (1980).

[17]K. Levenberg, Quart. J. Appl. Math., 2, 164 (1944).

[18]D. W. Marquardt, J. Soc. Ind. Appl. Math. 11, 431 (1963).

[19]K. M. Brown and J. E. Dennis, Numeriche Mathematic, 18, 289 (1972).

[20]P. B. Ryan, R. L. Barr and H. D. Todd, Anal. Chem., 52