Journal of Biomolecular NMR, 5 (1995) 287–296
ESCOM

J-Bio NMR 243

A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis

Christoph Antz[a], Klaus-Peter Neidig[b] and Hans Robert Kalbitzer[a,*]

[a]*Department of Biophysics, Max-Planck-Institute for Medical Research, P.O. Box 103820, D-69028 Heidelberg, Germany*
[b]*Bruker Analytische Messtechnik, D-76287 Rheinstetten, Germany*

Received 5 August 1994
Accepted 11 October 1994

Keywords: Multivariate discriminant analysis; 2D NMR spectroscopy; NOESY; Bayesian analysis; Peak recognition; NMR molecular structure

Summary

A generally applicable method for the automated classification of 2D NMR peaks has been developed, based on a Bayesian approach coupled to a multivariate linear discriminant analysis of the data. The method can separate true NMR signals from noise signals, solvent stripes and artefact signals. The analysis relies on the assumption that the different signal classes have different distributions of specific properties such as line shapes, line widths and intensities. As to be expected, the correlation network of the distributions of the selected properties affects the choice of the discriminant function and the final selection of signal properties. The classification rule for the signal classes was deduced from Bayes's theorem. The method was successfully tested on a NOESY spectrum of HPr protein from *Staphylococcus aureus*. The calculated probabilities for the different signal class memberships are realistic and reliable, with a high efficiency of discrimination between peaks that are true NOE signals and those that are not.

Introduction

NMR has become a powerful tool for the determination of macromolecular structures in solution (see, e.g., Ernst et al., 1986; Wüthrich, 1986; Griffey and Redfield, 1987; Fesik, 1991; Hausser and Kalbitzer, 1991). A practical problem often encountered during the evaluation of multidimensional NMR data is the occurrence of noise or artefact peaks which may result in erroneous conclusions (assignments) if they are not recognized. In a manual evaluation of the data, these peaks can usually be recognized if the spectra are not too crowded. The problem is more severe in automatic or semiautomatic procedures where the occurrence of a large number of noise and artefact peaks tends to lead to an instability of the algorithms (resulting in wrong results or a large number of possibly true solutions). As a practical solution for this problem, most of the more advanced program packages use a program architecture which allows the user to examine the results and remove wrong solutions at any stage of the evaluation. However, this interactive work is tedious and time-consuming and it is desirable to transfer at least parts of this task to the computer. After an optimal preprocessing (usually including time-domain filtering, base line correction and sometimes enhancement of symmetrical spectral features), at the lowest level true resonance peaks have to be distinguished from noise and artefact peaks. To be able to do this, it is necessary to know the features of true resonance peaks. For example, the probability that a signal is due to (thermal) noise decreases with increasing intensity. Since the first publication of a program for automated spin pattern search in 2D NMR spectra (Neidig et al., 1984), this feature has been used (in a rather primitive way) in all programs that use a peak recognition step, by defining a lower threshold for the intensity of true signals. Additional features which can be taken into account are the peak shape, occurrence at positions where artefacts are to be expected, or the presence of symmetry-related partners (see, e.g., Glaser and Kalbitzer, 1987; Kleywegt et al., 1989,1990; Stoven

*To whom correspondence should be addressed.

et al., 1989; Neidig and Kalbitzer, 1990; Garett et al., 1991; Rouh et al., 1994). In the past, these features have been used not only for peak assessment in automated pattern recognition but also in the context of automated base plane corrections which, likewise, have to distinguish true signals from (base plane) artefacts (Dietrich et al., 1991; Güntert and Wüthrich, 1992; Manorelas and Norton, 1992).

For structural determination of macromolecules it is important to have a set of NOESY cross peaks as large and as reliable as possible; two requirements which are in some respects mutually exclusive. Here, reliability means the certain knowledge that the signal $f(\omega_1,\omega_2)$ is due to the magnetization transfer between two protons with resonances at the chemical shift positions ω_1 and ω_2, but not due to artefact or noise effects. Even after a complete assignment of the spin systems and with exact knowledge of all relevant chemical shifts, with $\bigcup_i \omega_i = \Omega$, there may be a number of signals $S(\omega_i,\omega_j)$ with $\omega_i,\omega_j \in \Omega$ whose physical origin is of pure statistical or systematic (artefact) nature, such as thermal noise from the sample or the probehead, electronic noise from the hardware, digital noise, strong solvent stripes or t_1 noise (Mehlkopf et al., 1984). Such misinterpreted signals can cause severe convergence problems and/or can lead to less accurate structures when they serve as input information for the subsequent calculations. Very often such convergence problems, emerging during a constrained molecular dynamics (MD) run as NOE violations, serve the expert as a corrective for filtering of the original data input set. It often turns out that some of the signals originate from misinterpreted noise or artefact signals (apart from misinterpreted solvent or impurity signals) and have to be rejected, whereas other signals, originally thought to be irrelevant noise signals, must be added.

Bayesian reasoning is a widely used statistical method of great flexibility, one of its advantages being that it can be used even if only partial information about the system is available. Bayesian methods have been applied to a variety of problems in NMR, including signal reconstruction, line fitting and peak recognition (Jaynes, 1985; Skilling and Gull, 1985; Bretthorst et al., 1988; Rouh et al., 1994). In a recent paper, Rouh et al. (1994) used a Bayesian method to predict if a data point in the spectrum is part of a peak or noise, with the assumption of Gaussian functions. In this paper we do not deal with the separation of data but with a different problem: we present a generally applicable Bayesian method coupled to a multivariate linear discriminant analysis (Fisher, 1936; Tatsuoka, 1970) which calculates the probability that a given signal in an NMR spectrum is a member of a predefined class. It does not depend on special assumptions on noise distributions and can serve as a robust decision aid for the classification of signals into valid signals and noise and artefact signals. The method is based only on the general idea that signal classes having different physical origins, such as NOE transfer signals in NOESY spectra, noise, and solvent signals have distinct and measurable properties in their frequency domain.

Materials and Methods

Sample preparation

HPr protein from *Staphylococcus aureus* was isolated as described by Kalbitzer et al. (1982). The sample contained 4.9 mM HPr protein and 0.05 mM EDTA in 500 µl of 95% H_2O/5% D_2O. Prior to sample preparation, oxygen was removed by flushing the solvents with helium and by storing the lyophilized protein in a helium atmosphere for several hours. After dissolving the HPr protein in the oxygen-free solvent, the pH was adjusted to 7.8 by adding appropriate amounts of NaOD. The solution was transferred to the sample tube, which was then sealed off in helium.

NMR spectroscopy

NOESY spectra (Jeener et al., 1979) were recorded on a Bruker AM-500 NMR spectrometer operating at 500 MHz. The water signal was suppressed by selective presaturation. A mixing time of 100 ms was used. Phase-sensitive detection in the t_1 direction was obtained according to Marion and Wüthrich (1983). 4K × 1K time domain data were recorded and Fourier transformed to obtain 1K × 1K real data points in the frequency domain. The spectral widths in the two dimensions were 6849.31 Hz, resulting in a final digital resolution of 6.68 Hz/point. Prior to Fourier transformation the data were filtered exponentially with a line broadening of 8 Hz. The region between 5 and 11.5 ppm (relative to internal 4,4-dimethyl-4-silapentane sulfonic acid) was baseline corrected in the ω_2 direction according to Saffrich et al. (1992).

Software

Peak picking and integration was performed using the standard routines of the program package AURELIA. The method used by AURELIA for peak picking has been outlined earlier (Neidig et al., 1984). The software developed and described here is implemented in the latest release version of AURELIA. Relaxation matrix calculations were performed with the program X-PLOR 3.1 (Brünger, 1993). The calculation of Spearman's rank correlation coefficients C_s (Spearman, 1904,1908) and of Hoeffding's correlation coefficients C_H (Hoeffding, 1948; Hollander and Wolfe, 1978) was performed with the SAS package, Version 5 (1985).

Theoretical considerations

A satisfactory method for peak assessment must provide a measure of the reliability of the cross peak under

consideration. Ideally, it gives the probability that the peak is a member of a certain class C_i (e.g., a class of true resonance signals or a class of noise and artefact signals). In addition, it should fulfill the following conditions: (i) it has to be flexible; that is, it should be applicable to different kinds of multidimensional spectra, independent of special data processing; (ii) it must be easy to computerize and should require little information from the spectroscopist. For the estimation of the probabilities, the cross peaks must be characterized in some way. In the present implementation, local properties E_k (1,....,K') of peaks are used for the classification. They are derived from the peak shapes, the peak intensities and the peak widths. However, the method developed below could also be applied to nonlocal properties such as the existence of a symmetry-related cross peak or the location of t_1 noise. The properties characterizing peaks of a given class are usually not single-valued, but are defined by continuous distribution functions. Theoretically, the peak shape for a single transition is well defined in solution state NMR, and is a Lorentzian. Therefore, a two-dimensional resonance peak in the absorption phase could be defined by three parameters: the amplitude and the line widths in the two dimensions. However, in practice this does not work, since the observed lines are inhomogeneously broadened by unresolved J-couplings, and their shapes are distorted by the time-domain filtering of the data (except where an exponential filter has been applied). The exact shape of individual noise peaks, however, is in principle unpredictable, because it is produced by stochastic and noncoherent time processes.

Therefore, different signal classes C_i (i = 1,...,I) are characterized by multivariate probability distributions $p(\mathbf{E}|C_i)$ where $\mathbf{E}$ is a K'×1 matrix containing K' properties (measured continuous variables) E_k as components. In principle, we are free in the selection of the variables E_k that are used for the classification of the peaks. However, since we want to discriminate between different classes, the choice of E_k is critical; they should be based on characteristic features of the cross peaks, which are different for peaks of different classes.

Starting with Bayes's theorem (Cornfield, 1967,1969), the probability $P(C_i|\mathbf{E}^j)$ that the cross peak j with the values E_k^j of the properties E_k (k = 1,...,K') belongs to class C_i can be calculated as

$$P(C_i|\mathbf{E}^j) = \frac{P(C_i)P(\mathbf{E}^j|C_i)}{\sum_{i=1}^{I} P(C_i)P(\mathbf{E}^j|C_i)} \quad (1)$$

with $P(C_i)$ the a priori probability of finding a cross peak of class C_i and $P(\mathbf{E}^j|C_i)$ the probability of finding the property matrix $\mathbf{E}^j$ for peaks j of class C_i. In the following, we denote the probability distributions by p and their probabilities by P. If the multivariate probability distribution $p(\mathbf{E}|C_i)$ is not known a priori, it has to be obtained from a sample which, in general, must be rather large in order to describe the multidimensional distribution function completely. The distribution can be obtained from a much smaller sample if the properties E_k are statistically independent. In this case, the multivariate distribution can be obtained as a product of the univariate distributions $p(E_k|C_i)$. If Q variables are independent and the remaining K = K' − Q variables are correlated, the probability distribution $p(\mathbf{E}|C_i)$ becomes, according to the multiplication theorem for the probabilities of independent events:

$$p(\mathbf{E}|C_i) = p(\mathbf{R}|C_i) \prod_{q=1}^{Q} p(E_q|C_i) \quad (2)$$

with $p(\mathbf{R}|C_i)$ the reduced, K-dimensional probability distribution. The probability distribution $p(\mathbf{R}|C_i)$ can be analysed by searching for a new set of independent variables Y which are a linear combination of the components $R_k = E_k$ ($1 \le k \le K$). With $\mathbf{R}$ the reduced vector (K × 1 matrix) of properties and $\mathbf{a}' = (a_1,...,a_K)$ the coefficient matrix, Y can be written as $Y = \mathbf{a}'\mathbf{R}$. Y corresponds to the linear discriminant function of the reduced problem. The corresponding matrix of coefficients $\mathbf{a}'$ is calculated by maximizing λ with

$$\lambda = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} \quad (3)$$

The matrices $\mathbf{B}$ and $\mathbf{W}$ are calculated from the samples for the classes C_i as:

$$\mathbf{B} = \sum_{i=1}^{I} (\bar{\mathbf{R}}_i - \bar{\mathbf{R}})(\bar{\mathbf{R}}_i - \bar{\mathbf{R}})' \quad (4)$$

$$\mathbf{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{R}_i^j - \bar{\mathbf{R}}_i)(\mathbf{R}_i^j - \bar{\mathbf{R}}_i)' \quad (5)$$

with n_i the number of elements in the sample of class C_i, I the number of classes and $\mathbf{R}_i^j$ the reduced property vector for the peak j in class i. Furthermore:

$$\bar{\mathbf{R}}_i = 1/n_i \sum_{j=1}^{n_i} \mathbf{R}_i^j$$

and

$$\bar{\mathbf{R}} = 1/I \sum_{i=1}^{I} \bar{\mathbf{R}}_i$$

As usual, the optimum can be found by setting the first derivative $\partial\lambda/\partial\mathbf{a}$ to zero, that is:

$$\frac{\partial\lambda}{\partial\mathbf{a}} = \frac{2[(\mathbf{B}\mathbf{a}) - \lambda(\mathbf{W}\mathbf{a})]}{\mathbf{a}'\mathbf{W}\mathbf{a}} = 0 \Leftrightarrow (\mathbf{B} - \lambda\mathbf{W})\,\mathbf{a} = 0 \quad (6)$$

If **W** is not singular (which is usually true), Eq. 6 can be rewritten as:

$$(\mathbf{W}^{-1}\mathbf{B}-\lambda\mathbf{1})\,\mathbf{a} = 0 \tag{7}$$

with **1** the unity matrix.

In order to find nontrivial solutions of this homogeneous equation system, we have to solve the characteristic equation

$$\det|\mathbf{W}^{-1}\mathbf{B}-\lambda\mathbf{1}| = 0 \tag{8}$$

This solution yields a polynomial of degree $r = \mathrm{rank}(\mathbf{W}^{-1}\mathbf{B})$ with r eigenvalues λ_h (h = 1,...,r) and r corresponding eigenvectors $\mathbf{a}_h$, fulfilling Eq. 7. The resulting linear transformations $Y_h = \mathbf{a}_h'\mathbf{R}$ are orthogonal and thus statistically independent (Wiesböck, 1987). With $\mathbf{a}_h$, every vector $\beta\,\mathbf{a}_h$ ($\beta \in \Re$, $\beta \neq 0$) will fulfil the above condition. This means that scaling of the coefficient eigenvector $\mathbf{a}_h$ has no effect on the separation power of Y_h. Thus, the eigenvectors can be normalized by:

$$\mathbf{a}_{h,\mathrm{norm}} = \frac{\mathbf{a}_h}{\sqrt{\mathbf{a}_h'\,\mathbf{a}_h}} \tag{9}$$

The new probability distribution $p(\mathbf{Y}|C_i)$ can be written as:

$$p(\mathbf{Y}|C_i) = \prod_{k=1}^{r} p(Y_k|C_i) \tag{10}$$

The dimensionality of the new probability distribution $p(Y,C_i)$ can be reduced (with a possible loss of information) by the selection of the r' (r' < r) solutions with the largest values of λ. Then the total probability distribution becomes:

$$p(\mathbf{E}|C_i) = \prod_{k=1}^{r'} p(Y_k|C_i) \prod_{q=1}^{Q} p(E_q|C_i) \tag{11}$$

By substituting this into Bayes's formula and replacing p with P one obtains Eq. 12:

$$p(C_i|Y_1,\ldots,Y_{r'},E_1,\ldots,E_Q) = \frac{P(C_i)\prod_{k=1}^{r'} P(Y_k|C_i)\prod_{q=1}^{Q} P(E_q|C_i)}{\sum_{j=1}^{I}\left[P(C_j)\prod_{k=1}^{r'} P(Y_k|C_j)\prod_{q=1}^{Q} P(E_q|C_j)\right]} \tag{12}$$

Definition of variables and calculation of probability distributions

The probability distributions of the variables in Eq. 12 must be known to calculate the probabilities $P(C_i|\mathbf{Y}^j,\mathbf{E}^j)$ of a given peak j to be member of the class C_i. These are not known a priori and are expected to vary between different NMR spectra. Therefore, they must be extracted from the spectra. A simple way to obtain the distributions of the different signal classes consists in the definition of spectral regions in which predominantly signals of only one class C_i are present. Figure 1 shows these regions in the NOESY spectrum used below for characterization of the basic properties of our discrimination method.

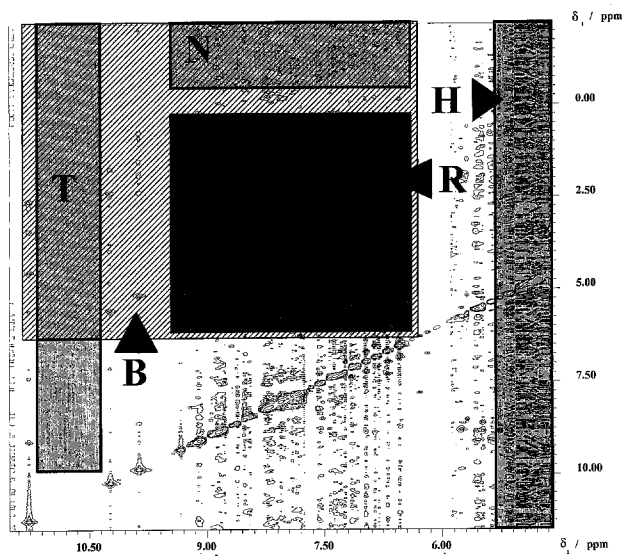

Fig. 1. Training areas for different peak classes, indicated in part of a NOESY spectrum of HPr protein from *S. aureus*, dissolved in H_2O. The lowest contour level depicted is 80 000. In the shaded areas, predominantly peaks from a single class are to be expected (**S**: true NOE signals; **T**: thermal noise; **N**: t_1 noise and artefacts; **H**: artefacts from the suppressed water signal). These areas were used to create the different probability distributions. Additional regions (**B** and **R**) are indicated which were used to test the performance of the discriminant analysis.

The selected regions are indicated by different shadowed squares, highlighted with capital letters. We denote **T** as a region of thermal noise (peak class T), **N** as a region containing the more intense spectral artefacts such as t_1 noise (peak class N), **H** as a region containing the typical artefacts from the suppressed H_2O signal (peak class H), **S** as a region containing true NMR signals (peak class S), **B** as the test region with a mixture of NOE signal peaks and noise signals, and **R** as a smaller part of region **B**, where the procedure has been tested by a recalculation of the NMR spectrum from the known three-dimensional structure.

The definition of our actual variables was guided by our experience in analysing NOESY spectra. What are the features and patterns that determine the decision of

an experienced NMR spectroscopist that a single and unknown peak is an NOE or a non-NOE signal, even without a time-consuming and often ambiguous lineshape analysis? One would be very sceptical about peaks with broad tails, tall peaks with high intensities looking like noise spikes, or extended rectangular peaks with flat tops, the typical texture of solvent stripes. Recognition of these patterns by an automaton instead of our visual system requires the evaluation of different peak curve integrals, normalized to the respective peak intensities. We constructed distributions of the following variables for the different signal classes, where $f(\omega_1,\omega_2)$ denotes the formal description of the 2D signal:

(i) The absolute intensity (amplitude) of the signal, denoted by E1;

(ii) The ratio of peak volume to peak intensity, denoted by E2. This is defined as:

$$E2 = \frac{\iint_{\substack{\omega_1,\omega_2 \text{ for} \\ f \geq 0.05E1}} f(\omega_1,\omega_2)\,d\omega_1 d\omega_2}{E1} \tag{13}$$

(iii) The relative volume of the tail of the peak, denoted by E3. This is defined as:

$$E3 = \frac{\iint_{\substack{\omega_1,\omega_2 \text{ for} \\ f \geq 0.05E1}} f(\omega_1,\omega_2)\,d\omega_1 d\omega_2 \; - \iint_{\substack{\omega_1,\omega_2 \text{ for} \\ f \geq 0.2E1}} f(\omega_1,\omega_2)\,d\omega_1 d\omega_2}{E1} \tag{14}$$

(iv) The relative volume of the top of the peak, denoted by E4. This is defined here as:

$$E4 = \frac{\iint_{\substack{\omega_1,\omega_2 \text{ for} \\ f \geq 0.5E1}} f(\omega_1,\omega_2)\,d\omega_1 d\omega_2}{E1} \tag{15}$$

The class membership of the peaks used for the construction of the probability distributions should be as definite as possible, which means that the regions in which the recognition procedure is learned should contain only peaks of one class. This condition can only be approximated in real 2D spectra. Most areas contain peaks from more than a single class. As can be seen in our example (Fig. 1), it is easy to find a region where thermal noise is present exclusively (region **T**). The area containing artefacts like t_1 noise (region **N**) is also easy to define, since it is always possible to find areas where no true NMR signals are to be expected. However, even here one has to be aware that this class comprises at least two types of peaks, with possibly different characteristics, i.e. thermal noise peaks and artefact peaks. The training area for true signals (region **S**) always contains noise peaks and artefact peaks. The region with the artefacts resulting from imperfect water suppression usually contains numerous peaks of the classes T, N and S.

For an effective discriminant analysis, it is essential that the samples used for the calculation of the probability distributions contain predominantly peaks of only one class. If this cannot be achieved easily by the definition of the training areas, additional selection rules should be applied. For properties which, at least in a first approximation, are independent of the intensity (properties E2 to E4), this can be done in a simple way by applying an intensity threshold. Peaks of the classes S, N and H can be separated from thermal noise in region **T** by using only peaks with intensities higher than the 95th percentile of peaks in area **T**. True NOE signals can be separated from artefacts and thermal noise by selecting peaks in area **S** with intensities higher than the 95th percentile found in area **N**. The only problem is the construction of the intensity distribution of signals at low intensities, for which no straightforward method exists. However, if the area is not too crowded, and with the approximation that peak overlaps or superpositions can be neglected, the distribution p(E1|S) obtained from the area **S** can be corrected in a simple way. The corrected probability distribution $p(E1|S^{corr})$ is given by:

$$p(E1|S^{corr}) = \left(1 + \frac{D^N}{D^S}\right) p(E1|S) - \frac{D^N}{D^S} p(E1|N) \tag{16}$$

with D^S and D^N the peak densities in areas **S** and **N**. In our spectra the signal area is rather crowded. Here, it turned out that better results could be obtained by replacing the peak densities D^S and D^N in Eq. 16 by the intensity densities A^S and A^N, which are defined by:

$$A^S = \frac{\sum_{i \in S} E1_i}{F(S)} \tag{17a}$$

$$A^N = \frac{\sum_{i \in N} E1_i}{F(N)} \tag{17b}$$

with F(S) and F(N) the areas of **S** and **N**, respectively. The sum in Eqs. 17a and b is carried out over all peaks in the **S** or **N** areas. $p(E1|S^{corr})$ is then defined by:

$$p(E1|S^{corr}) = \left(1 + \frac{A^N}{A^S}\right) p(E1|S) - \frac{A^N}{A^S} p(E1|N) \tag{18}$$

The only factors not yet determined in Eq. 12 are the probabilities $P(C_i)$, i.e., the general a priori probabilities of finding peaks of a given class C_i. There is no really obvious way to approximate these probabilities. One solution would consist in setting the (subjective) probabilities P(N) and P(S) to 0.5, which would be equivalent to the statement that no a priori knowledge about the occurrence of noise and signal peaks is available. Another way to approximate the a priori probabilities in NOESY-type spectra could be based on the intensities at the diag-

onal of the spectrum. In this case the a priori probability could be dependent on two frequency coordinates. Since not all types of two-dimensional spectra contain diagonal signals, this method is not generally applicable. A simple estimate for P(N) and P(S) can be obtained on the basis of the same arguments used for the derivation of Eq. 18. Because of the high peak density, the intensity density is again probably somewhat better suited than the pure peak density for the approximation of the a priori probabilities. Therefore, the present implementation is based on the normalized intensity densities, namely:

$$p(N) = \frac{A^N}{A^S + A^N} \tag{19}$$

and

$$p(S) = \frac{A^S}{A^S + A^N} \tag{20}$$

However, a general advantage of the Bayesian approach is that it is not critically dependent on the exact knowledge of the a priori probabilities, as long as they are not wrongly assumed to be almost zero or one.

Since the probability distributions are obtained from a limited, discrete sample, they must be smoothed. In a first step, the range of possible values is divided into discrete intervals containing a sufficiently high number of events. In a second step, a moving average filter, defined as:

$$P_i = \left(3P_i + \sum_{k=i-3;\, k\neq i}^{i+3} P_k\right) / 9 \tag{21}$$

(with $P_{i,k}$ the values in the intervals i,k) is applied to these intervals. If there are still intervals with zero probability, the zero is replaced by a value of 10^{-8} to avoid discontinuities in the distribution function due to too small a size of the sample. Finally, the distribution functions are renormalized to a total integral of 1.

Results and Discussion

Figure 2 shows the probability distributions of the properties E1, E2, E3 and E4 for the different classes C_i defined above. For all signal classes the abscissa has been divided into 100 equal intervals. Peak picking was performed above a threshold level of 80 000, which corresponds to the 75th percentile of the E1 distribution for all positive **N** area peaks. Below this level, noise and artefacts dominate the entire signal and the evalution of the data is not meaningful. Since the thermal noise is very weak in the analysed spectrum (maximum intensity 56 000), in the **T** area a lower intensity threshold of 24 000 has been applied. The E1 distribution of these peaks is not shown in Fig. 2a, since the lowest intensity depicted is 80 000. Inspection of these distributions shows that all classes differ significantly in their distribution functions for the properties E1–E4. As expected for small ampli-

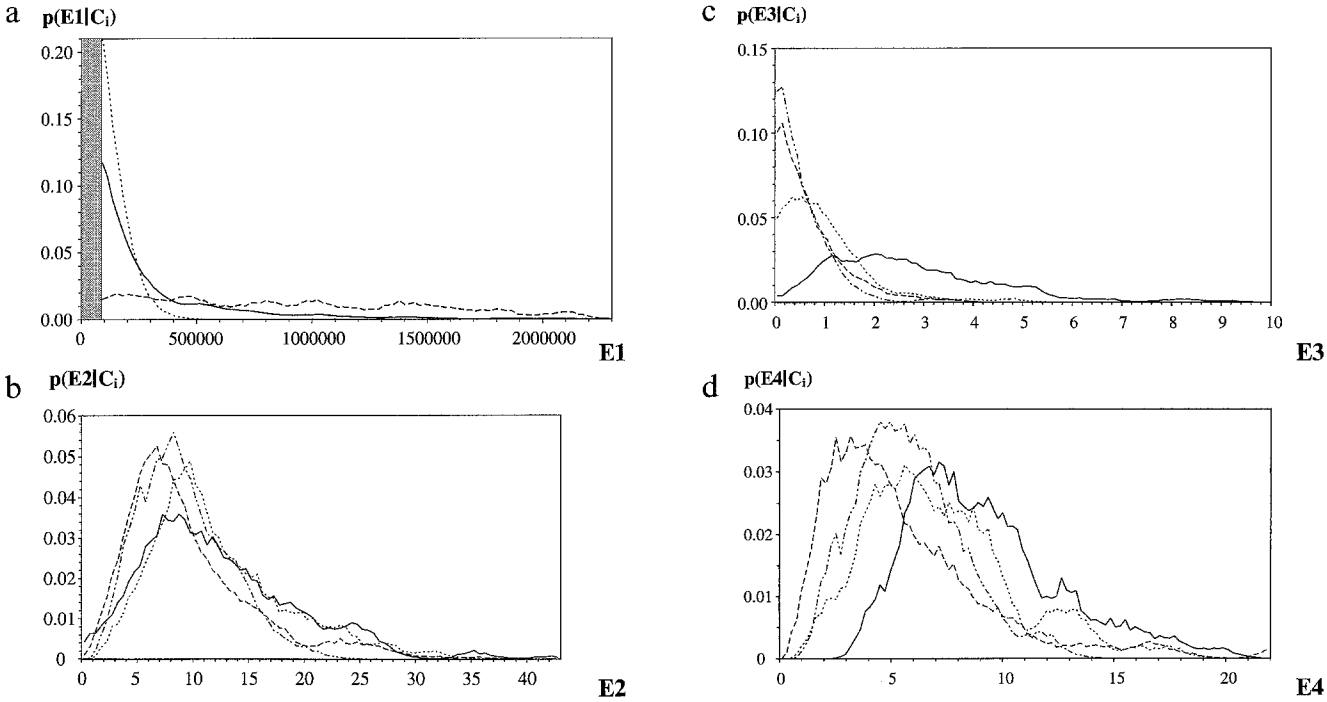

Fig. 2. Uncorrected probability distributions: (a) $p(E1|C_i)$; (b) $p(E2|C_i)$; (c) $p(E3|C_i)$; and (d) $p(E4|C_i)$ for the different signal classes C_i. (·····) = signals of the noise class N; (——) = true signals of class S; (- ·· -) = thermal noise peaks of class T; and (------) = signals from the suppressed water resonance (class H). Only signals with intensities above 80 000 were taken into account for signal classes S, N and H. Since the maximum intensity in the **T** area is 60 000, its area distributions were calculated from signals with intensities higher than 24 000. No signals of the classes S, N and H were sampled in the shaded intensity interval of Fig. 2a.

TABLE 1
CHARACTERISTIC STATISTICAL PROPERTIES OF THE DIFFERENT PROBABILITY DISTRIBUTIONS

Region (n[a])	Variable	Maximum	Minimum	s[b]	P25[c]	Median	P75[c]
S[d] (795)	E1	2 275 778	80 052	348 642	121 338	195 738	408 224
	E2	48.31	1.07	6.80	8.87	12.41	16.94
	E3	9.27	1.00	3.49	5.19	6.84	9.36
	E4	21.85	0.04	1.36	0.59	1.10	2.04
N[d] (192)	E1	384 900	80 066	49 869	99 392	123 522	159 878
	E2	32.29	2.69	5.98	8.24	10.73	15.55
	E3	17.89	1.58	3.31	4.73	6.60	8.95
	E4	4.88	0.02	0.44	0.44	0.87	1.34
H[d] (352)	E1	253 777 584	90 358	37 407 077	1 048 982	3 520 726	8 997 188
	E2	35.73	1.0	5.58	5.69	7.94	11.78
	E3	25.63	0.05	3.88	2.94	4.61	7.10
	E4	3.73	0.05	0.70	0.25	0.53	1.03
T (182)	E1	56 026	24 082	5311	25 510	27 696	31 072
	E2	21.19	1.75	3.77	6.08	8.56	11.39
	E3	12.44	1.62	2.27	4.16	5.59	7.12
	E4	3.79	0.05	0.58	0.19	0.42	0.75

[a] n = number of peaks.
[b] s = standard deviation.
[c] P25 and P75 are the 25th and 75th percentile, respectively.
[d] Data were taken from the probability distributions presented in Fig. 1, except for the signals in the **T** area where a lower threshold of 24 000 was used.

tudes, the probability distributions of peak amplitudes (property E1) show a pronounced probability for noise peaks (area **N**) compared to the distributions of peaks from the signal area **S**. In contrast, the artefacts from the suppressed water signal (area **H**) are characterized by a very flat distribution function with rather high probabilities for high intensities. (Fig. 2a). Correspondingly, the statistical analysis of the distributions (Table 1) results in the smallest value of the median and the standard deviation for thermal noise T, the noise and artefact peaks N having an almost fourfold higher median and an approximately 10-fold higher standard deviation than peaks of class T. Although there is only a factor of two between the medians of peaks of the noise class N and the signal class S, the distribution of the signal peaks is significantly broader, with a sevenfold larger standard deviation. The peaks in the **H** area are more intense by more than a factor of 10 compared to the signal peaks S; this is the reason why signals in this area are usually not evaluated. It is clear from Fig. 2 that the differences between the peak classes also hold for the other properties; the distributions overlap but are different.

As stated above, the E1 distribution of signals obtained from the **S** area must be corrected, since it contains a certain amount of noise peaks. Figure 3 shows the effect of such a correction using Eq. 18 for the E1 distribution of class S signals. Under the assumption that the published NMR structure of HPr from *S. aureus* is correct (Kalbitzer and Hengstenberg, 1992), it is possible to calculate the 'true' noiseless NOESY spectrum on the basis of the full relaxation matrix formalism (Borgias and James, 1984; Keeper and James, 1984). The distribution obtained in this way agrees fairly well with our corrected probability distribution, i.e., our method of correction represents a good approximation to the 'true' NOE signal distribution.

In most practical cases only the assignment to two classes, that of true NOE signals and that of noise and artefact peaks, is of interest. In the following we restrict our analysis to these two classes. In order to determine the mutual dependence of our variables, a correlation analysis for all proposed variables was performed. As before, only peaks of area **S** with intensities higher than 80 000 were used for this analysis. The calculated Spearman's rank correlation coefficients C_s (Spearman, 1904, 1908) and the Hoeffding's correlation coefficients C_H

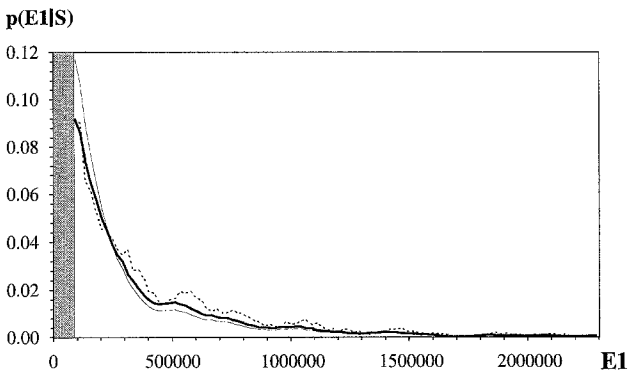

Fig. 3. Corrected probability distribution of intensity E1, together with the 'ideal' probability distribution. Uncorrected probability distributions of signals in area **S** (——), probability distribution noise-corrected according to Eq. 18 (——), and 'true' NOE signal distribution (------). No signals were sampled in the shaded intensity interval.

TABLE 2
CORRELATION COEFFICIENT MATRICES FOR S AREA SIGNALS

	E1	E2	E3	E4
Spearman's rank correlation coefficients				
E1	1	−0.4185	0.5033	0.2224
E2	−0.4185	1	−0.7622	−0.8975
E3	0.5033	−0.7622	1	0.6468
E4	0.2224	−0.8975	0.6468	1
Hoeffding's correlation coefficients				
E1	1/30	0.0017	0.0027	0.0005
E2	0.0017	1/30	0.0079	0.0139
E3	0.0027	0.0079	1/30	0.0050
E4	0.0005	0.0139	0.0050	1/30

(Hoeffding, 1948; Hollander and Wolfe, 1978) are given in Table 2. C_S can vary between −1 and 1; values near −1 or 1 indicate correlated variables. C_H can vary between −1/60 and 1/30; values near 1/30 reveal correlated properties.

Examination of the data shows that the intensity distribution E1 is rather independent of the other distributions, whereas the other variables are mutually correlated. Such behaviour is to be expected, since E2–E4 describe properties of the peak shape which should be largely independent of the intensity E1 for noise and signal peaks. Therefore, to a first approximation, E1 can be treated as an independent variable but the remaining three variables should be subjected to the discriminant analysis described above.

For the evaluated NOESY spectrum, the solution with the highest eigenvalue and thus the highest discrimination power gives the normalized eigenvector components a2 = 0.027, a3 = 0.999 and a4 = 0.021 (with the reduced variable Y = (a2 E2 + a3 E3 + a4 E4)). Obviously, in this data set the variable E3 yields the most important contribution for the construction of Y. However, this is probably not generally true for all kinds of data sets, since E2, E3 and E4 describe different aspects of the line shape. In different data sets different eigenvectors will be found; this feature becomes obvious if one realizes that a different digital filtering of the NOESY spectrum used as example here is sufficient to change the calculated eigenvector components. The probability distribution obtained after this transformation for N and S peaks is depicted in Fig. 4.

The distributions of Y obtained for the two classes appear rather different, although there is still a considerable overlap. Although one can expect three (formally) independent solutions from the discriminant analysis of our data, in discriminant analysis usually only a subset of the possible solutions is used (one reason for this is that the limited size of the samples suggests a number of formally independent solutions which do not actually exist). For the analysis of our data it turned out to be sufficient to take only the solution with the highest eigenvalue, i.e., the final probability distribution (Eq. 11) reduced to the case r' = 1, Q = 1, where it is only dependent on two variables: E1 and Y.

In order to compare our calculated probabilities P(S|Y,E1) with the 'ideal' probabilities RS, it is necessary to know which peaks belong to which class in a test spectrum. Such a decision can be made (at least to a first approximation) by the calculation of a theoretical NOESY spectrum from the three-dimensional structure of the protein (see above). RS indicates the relative number of true signal peaks (i.e., peaks confirmed by the back-calculation) in a given probability interval. Ideally, the two probabilities should be identical. Figure 5 shows that our procedure provides a satisfactory result, the calculated probabilities approximating rather well to the expected linear behaviour. Our resulting data can be used for an a posteriori analysis of the training sets. Such an analysis was performed for our training sets of noise and signal peaks, where the mean value of P(S|Y,E1) was 0.24 for peaks in the **N** area and 0.63 for peaks in the **S** area.

A typical result of the Bayesian procedure described is depicted in Fig. 6. The figure shows a 2D contour plot of the **R** region defined in Fig. 1. Note that the lowest contour level of 80 000 shows a large number of peaks that could not be used in the published structure calculation (Kalbitzer and Hengstenberg, 1993) since they have the same intensity as artefact peaks. The typical minimal intensity of peaks used for these calculations corresponded to four contour lines in the plot. Peaks which were not predicted from the back-calculation, and which are most probably artefacts, are shaded. The probability P(S|Y,E1) that a given peak is a true NOE peak is indicated. In general, the calculated probabilities match very well with expectation, true peaks having probabilities near 1 and artefact peaks displaying low probabilities. However, miracles cannot be expected: very weak signal peaks usually obtain probabilities around 0.5, thus not yielding sufficient information to decide with certainty whether they represent true signals. The arrow indicates an extremely weak peak which represents an NOE contact

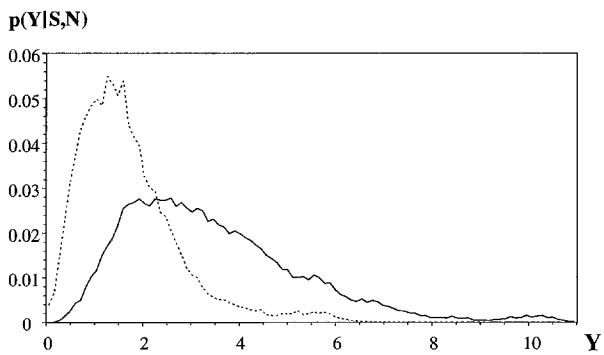

Fig. 4. Probability distributions of the reduced variable Y for peaks of the corrected S class (——) and of the N class (·····).

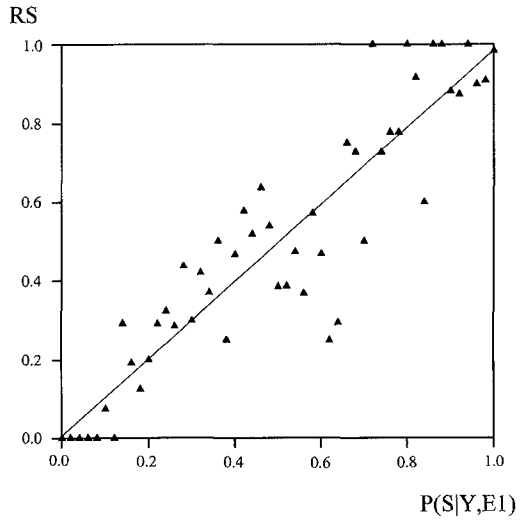

Fig. 5. Comparison between the expected probability RS and the calculated probability P(S|Y,E1). RS is plotted as a function of the probability P(S|Y,E1). RS was obtained by checking how many cross peaks in a given interval of P(S|Y,E1) were predicted from the back-calculation of the NOESY spectrum, i.e., $RS = (N_B^{NOE}(P(S|Y,E1)))/(N_B(P(S|Y,E1)))$, with $N_B^{NOE}(P(S|Y,E1))$ the number of predicted NOE peaks in **B** with the probability in the interval P(S|Y,E1), and $N_B(P(S,N|Y,E1))$ the total number of peaks in **B** in this interval (50 intervals with a width of 0.02 were used to divide the abcsissa in classes of similar probabilities). The straight line shows the best linear fit of $RS = A + B \times P(S|Y,E1)$ with $A = 0.006$, $B = 0.98$, a regression coefficient of 0.91 and a standard deviation of 0.13.

between residues 59 and 64. Despite its very low intensity, it has obtained an ambient probability.

For practical applications, the effective computing time is an important aspect. This time could be shortened if one tries to produce a standard library containing the probability distributions for true signal and artefact peaks together with their normalized eigenvectors. This would have the additional advantage that the training set could be created from a larger number of spectra and therefore include a larger number of samples for the creation of the distribution functions. However, a disadvantage is that the application of such a library would require a rigid standardization of acquisition and processing conditions, which most users would not accept. On a Bruker Aspect station, the definition of training sets and the calculation of distribution functions and eigenvectors takes only approximately 10 min, so computing time is not really an argument for the limitation of flexibility by standardization. However, it could be useful for the analysis of closely related data sets, such as a series of NOESY spectra recorded with different mixing times, where the improved definition of the distribution functions could lead to more accurate results. The calculation of the probabilities can also be performed in a reasonable time. For the spectrum shown in Fig. 1 the evaluation of 1000 peaks required approximately 10 s. These computing times allow application of the data analysis under experimental conditions which are optimized for the actual problem and may change from case to case. With regard to the data processing, it is not clear a priori which digital filter function is optimal for the discriminant analysis because the convolution with the filter function may reduce the discriminatory power. However, since optimal filtering enhances typical properties of signals and suppresses noise and artefacts, in our opinion it is better to use the discriminant analysis on a data set that has been optimized for such an enhancement. In the spectrum shown, we used an exponential filtering adapted to the expected line widths and line splittings of our protein spectrum. With such a line broadening, the multiplet structure of the peaks vanishes. Another possibility would be a Lorentzian-to-Gaussian transformation; such a transformation, however, would probably decrease the discrimination power of E3. The extreme alternative, the application to nonfiltered data, would probably have resulted in the correct assessment of a much larger number of noise peaks, simply because the number of noise peaks would have increased very much. However, such an improved discrimination of peaks that are not present in the correctly processed spectrum does not make much sense.

In our example, the intensity distribution was separated from the other variables and not included in the discriminant analysis. Such a procedure is not required by the algorithm itself. Essentially, the same results would have been obtained by performing the discriminant analysis in four-dimensional space, and retaining the two eigen-

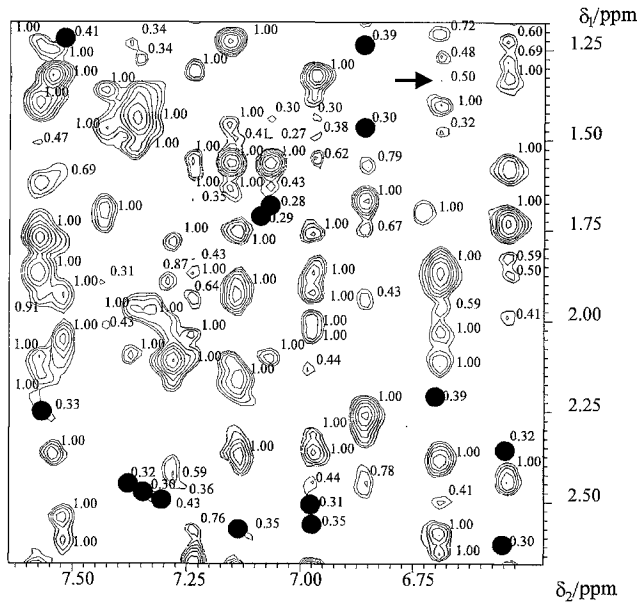

Fig. 6. Peak assessment in an experimental NOESY spectrum. The figure shows a contour plot of the R region of Fig. 1. Peaks not expected from the back-calculation of the NOESY spectrum are indicated by grey circles. The numbers represent the computed probabilities P(S|Y,E1) of a peak to be a true signal. The arrow indicates a very weak long-range NOE between residues 59 and 64. The lowest contour level starts at 80000, with a multiplication factor of 1.5 for subsequent levels.

vectors with the highest eigenvalues. However, initial separation of E1 has the practical advantage that it speeds up the calculation and this separation is justified, since E1 is only weakly correlated to the other properties.

Conclusions

The method presented in this paper is flexible and efficient, it is generally applicable to any kind of n-dimensional NMR spectrum (n > 1), and it gives the probability that a given peak in the spectrum is a member of a predefined class, such as the classes of noise or true signals. It does not depend on a priori assumptions about the expected peak shapes, the only condition for its use being the existence of some properties which have different probability distributions for the classes to be separated. It is very easy to use, since it only requires a definition of a training set by the user. In the present implementation we used the four characteristic properties E1–E4 for the discriminant analysis. We have shown that the choice of these properties is well suited for the analysis of our experimental data and we think that most types of two-dimensional spectra can be analysed sufficiently well with these basic properties.

However, one has to be aware that the algorithm described is very general; in particular, it can be used for any number and type of properties. Since the probabilities obtained reflect the information available, their discriminating power will increase with the amount of information used. Therefore, the choice of additional properties such as local peak symmetries and global symmetries may be useful, and can be fitted into the general algorithm as described above.

The probability information may be useful in many practical situations. We developed the algorithm mainly with two applications in mind, namely automatic spin system assignment and automatic calculation of a three-dimensional structure from a partly unassigned NOESY spectrum. In both applications the probability information could be used for an iterative procedure, where first the signals with high probability are used for the analysis and subsequently peaks with lower probabilities are taken into account as well.

References

Borgias, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **59**, 493–512.

Bretthorst, G.L., Hung, C.-C., D'Avignon, D.A. and Ackerman, J.J.H. (1988) *J. Magn. Reson.*, **79**, 369–376.

Brünger, A.T. (1993) X-PLOR Manual, Version 3.01, Yale University, New Haven, CT.

Cornfield, J. (1967) *Rev. Int. Statist. Inst.*, **35**, 34–49.

Cornfield, J. (1969) *Biometrics*, **25**, 643–657.

Dietrich, W., Rüdel, C.H. and Neumann, M. (1991) *J. Magn. Reson.*, **91**, 1–11.

Ernst, R.R., Bodenhausen, G. and Wokaun, A. (1986) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Oxford Science Publication, Oxford.

Fesik, S.W. (1991) *J. Med. Chem.*, **34**, 2937–2945.

Fisher, R.A. (1936) *Ann. Eugenics*, **7**, 179–188.

Garett, D.S., Powers, R., Gronenborn, A.M. and Clore, G.M. (1991) *J. Magn. Reson.*, **95**, 214–220.

Glaser, S. and Kalbitzer, H.R. (1987) *J. Magn. Reson.*, **74**, 450–463.

Griffey, R.H. and Redfield, A.G. (1987) *Q. Rev. Biophys.*, **19**, 51–82.

Güntert, P. and Wüthrich, K. (1992) *J. Magn. Reson.*, **96**, 403–407.

Hausser, K.H. and Kalbitzer, H.R. (1991) In *NMR in Medicine and Biology: Structure Determination, Tomography, In Vivo Spectroscopy*, Springer, Heidelberg, pp. 39–77.

Hoeffding, W. (1948) *Ann. Math. Statist.*, **19**, 546–557.

Hollander, M. and Wolfe, D.A. (1973) In *Nonparametric Statistical Methods* (Eds, Bradley, R., Hunter, H.S., Kendall, D.G. and Watson, G.S.) Wiley, New York, NY, pp. 228–236.

Jaynes, E.T. (1985) In *Maximum-Entropy and Bayesian Methods in Inverse Problems* (Eds, Smith, C.R. and Grandy, W.T.) Reidel, Dordrecht, pp. 21–58.

Jeener, J., Meier, B.H., Bachmann, P. and Ernst, R.R. (1979) *J. Chem. Phys.*, **71**, 4546–4553.

Kalbitzer, H.R., Hengstenberg, W., Rösch, P., Muss, P., Bernsmann, P., Dörschug, M. and Deutscher, J. (1982) *Biochemistry*, **21**, 2879–2885.

Kalbitzer, H.R. and Hengstenberg, W. (1992) *Eur. J. Biochem.*, **216**, 205–214.

Keeper, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.

Kleywegt, G.J., Lamerichs, R.M.J.N., Boelens, R. and Kaptein, R. (1989) *J. Magn. Reson.*, **85**, 186–197.

Kleywegt, G.J., Boelens, R. and Kaptein, R. (1990) *J. Magn. Reson.*, **88**, 601–608.

Manorelas, N. and Norton, R.S. (1992) *J. Biomol. NMR*, **2**, 485–494.

Marion, D. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.*, **113**, 967–974.

Mehlkopf, A.F., Korbee, D. and Tiggelman, T.A. (1984) *J. Magn. Reson.*, **58**, 315–323.

Neidig, K.-P., Bodenmueller, H. and Kalbitzer, H.R. (1984) *Biochem. Biophys. Res. Commun.*, **125**, 1143–1150.

Neidig, K.-P. and Kalbitzer, H.R. (1990) *J. Magn. Reson.*, **88**, 155–160.

Neidig, K.-P. (1993) AURELIA User's Guide, Version 931101, Bruker Analytische Messtechnik, Rheinstetten.

Rouh, A., Louis-Joseph, A. and Lallemand, J.-Y. (1994) *J. Biomol. NMR*, **4**, 505–518.

Saffrich, R., Beneicke, W., Neidig, K.-P. and Kalbitzer, H.R. (1992) *J. Magn. Reson. Ser. B*, **101**, 304–308.

SAS (1985) *SAS User's Guide: Basics*, Version 5, SAS Institute Inc., Cary, NC.

Skilling, J. and Gull, S.F. (1985) In *Maximum-Entropy and Bayesian Methods in Inverse Problems* (Eds, Smith, C.R. and Grandy, W.T.) Reidel, Dordrecht, pp. 83–132.

Spearman, C. (1904) *Am. J. Psychol.*, **15**, 72–101.

Spearman, C. (1908) *Br. J. Psychol.*, **2**, 227–242.

Stoven, V., Mikou, A., Piveteau, D., Guittet, E. and Lallemand, J.-Y. (1989) *J. Magn. Reson.*, **82**, 163–168.

Tatsuoka, M.M. (1970) *Multivariate Analysis Techniques for Educational and Psychological Research*, Wiley, New York, NY, pp. 94–190.

Wiesböck, K. (1987) Ph.D. Thesis, University of Passau, Passau.

Wüthrich, K. (1986) *NMR of Proteins and Nucleid Acids*, Wiley, New York, NY.