Journal of Biomolecular NMR, 4 (1994) 505–518
ESCOM

J-Bio NMR 186

Bayesian signal extraction from noisy FT NMR spectra

Alain Rouh, Alain Louis-Joseph and Jean-Yves Lallemand*

Département de Chimie et de Synthèse Organique, Ecole Polytechnique, F-91128 Palaiseau, France

Received 28 September 1993
Accepted 13 December 1993

Keywords: Histogram; Bayesian detection; 3D NMR

SUMMARY

The statistical interpretation of the histogram representation of NMR spectra is described, leading to an estimation of the probability density function of the noise. The white-noise and Gaussian hypotheses are discussed, and a new estimator of the noise standard deviation is derived from the histogram strategy. The Bayesian approach to NMR signal detection is presented. This approach homogeneously combines prior knowledge, obtained from the histogram strategy, together with the posterior information resulting from the test of presence of a set of reference shapes in the neighbourhood of each data point. This scheme leads to a new strategy in the local detection of NMR signals in 2D and 3D spectra, which is illustrated by a complete peak-picking algorithm.

INTRODUCTION

In the last 20 years, NMR has become a major spectroscopic technique for the study of biological macromolecules in solution. Proteins are of special interest, and currently structures of proteins up to 174 residues (Ikura et al., 1992) can be resolved using multidimensional NMR. Yet NMR has limitations, such as its intrinsic low sensitivity, coupled with a water signal which is often much more intense than the signal of interest, giving rise to dynamic range problems. Furthermore, spectra are often affected by artifacts such as baseline distortions, intense solvent lines and t_1 noise. Signal processing is thus essential to extract information that can be used in structure determination.

In general, post-acquisition treatments need either a noise-level evaluation or a discrimination between useful NMR information and noise. The first class of post-acquisition treatments can be illustrated by the following three methods.

One of the most widely used techniques is the Maximum Entropy Method (Delsuc, 1989) for high-resolution spectral analysis. This method requires an estimation of the standard deviation of the data points to stop signal reconstruction. Recently, another method has been reported by

*To whom correspondence should be addressed.

Manoleras (Manoleras and Norton, 1992) to remove t_1 noise and artifacts. It is based on a weighted moving average, whose ponderations are calculated from a local estimation of the standard deviation of the noise. Iterative refinement of structures of proteins, based on the comparison between simulated NOE intensities and experimental ones, also requires some estimation of the local noise level (Borgias and James, 1988; Nibedita et al., 1992).

The second class of post-acquisition treatments is demonstrated by NMR data set compression (Zolnai et al., 1988) and automatic peak-picking programs. The latter rely generally on a criterion similar to the determination of a threshold superior to the local noise level (Stoven et al., 1989; Kleywegt et al., 1990).

In the first part of this paper we present a thorough discussion concerning the statistical properties of noise which occurs in NMR experiments. First, the statistical approach commonly used in NMR, i.e., the local maximum likelihood estimation, is presented. From this method, the two best estimators of the noise expectation and standard deviation are derived. Another statistical approach, allowing the representation of a spectrum as a histogram, will then be discussed. This representation not only permits the evaluation of the Gaussian distribution hypothesis of the noise, but also the calculation of an alternative set of estimators.

The second part of this paper introduces a new procedure for signal detection in 2D and 3D NMR spectra. A statistical distribution model of the noise and a deterministic local model of a line, made of a set of characteristic neighbourhood configurations, are discussed. The signal detection is reduced to testing of the presence of reference shapes in the neighbourhood of each data point. The Bayesian approach to this problem leads to a testing procedure that combines prior knowledge, obtained from histogram interpretation, with posterior knowledge, extracted from the neighbourhood of a data point. The difference between the experimental histogram and the stimulated one for a pure noise data set provides a prior estimation of the signal probability and a good representation of artifact and NMR signal effects. The comparison between the neighbourhood of a point and the reference shapes, based on the noise distribution model, results in a posterior signal probability estimation. The testing procedure then combines these two information sets, depending on the noise level.

THEORETICAL CONSIDERATIONS ABOUT NOISE

In NMR experiments, data points are affected by some noise N. In order to deal with noise, hypotheses are formulated in this section and evaluated in the next one. The noise is assumed to rely on a stochastic process, in other words, for each acquisition time t, N(t) is a random variable. By definition the noise under study is uncorrelated with NMR signals, and the so-called t_1 noise is not under investigation here.

Assuming that N is white noise, then N is a series of uncorrelated random variables (Arques, 1982):

$$E[N(t)N(t + \Delta t)] = a(t)\,\delta\,(\Delta t) \tag{1}$$

where E is the mathematical expectation, δ the Dirac function and a(t) the noise energy at time t. When N is a stationary process, the probability properties of N are independent of the time origin and Eq. 1 becomes:

$$E[N(t)N(t + \Delta t)] = a\delta(\Delta t) \tag{2}$$

It is then possible to define time averages, and if N is an ergodic process (Arques, 1982; Ernst et al., 1987) these time averages shall converge to the ensemble averages over different realizations of a given random variable N(t). In this case, the study of the probability properties of N (ensemble averages) is equivalent to the study of the statistical properties (time averages).

After these statements about the properties of noise, one can attempt to estimate some of its parameters. Usually the expectation m and the standard deviation σ of the noise are estimated by interactively selecting an area assumed to be free of NMR signal, and computing the average and the average of the square in this area. In the case of a Gaussian distribution, the expectation and the standard deviation completely define the distribution law, and the following expressions are their respective maximum likelihood estimators:

$$\begin{aligned} \widehat{m_{lh}} &= \tfrac{1}{n} \sum_{i=1}^{n} x_i \\ \widehat{\sigma^2_{lh}} &= \tfrac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{m_{lh}})^2 \end{aligned} \tag{3}$$

where the subscripts lh denote maximum likelihood estimators.

Rather than manually selecting an area A of appropriate size, such that $\widehat{m_{lh}}$ and $\widehat{\sigma_{lh}}$ have acceptable biases and the signal contribution is low, it is possible to find automatically the best area of a given size by minimizing the signal influence. The intensity D of each data point is the sum of the noise N and possibly the NMR signal s:

$$D = N + s \tag{4}$$

As stated before, the noise and the signal are uncorrelated, so:

$$\sigma^2_D = \sigma^2_N + \sigma^2_s \tag{5}$$

and the maximum likelihood estimator of the standard deviation of D in any area A is:

$$\widehat{\sigma^2_{lh}}(A) = \widehat{\sigma^2_N}(A) + \widehat{\sigma^2_s}(A) \tag{6}$$

This equation shows that there is a systematic bias to the estimation of σ^2_N with $\widehat{\sigma^2_{lh}}(A)$, due to the signal contribution $\widehat{\sigma^2_s}(A)$.

With the hypothesis of a uniform noise distribution we obtain:

$$\widehat{\sigma^2_N}(A) = \widehat{\sigma^2_N} \tag{7}$$

The best estimate of σ^2_N is $\widehat{\sigma^2_{lh}}(A)$, so that the bias $\widehat{\sigma^2_s}(A)$ is minimal, i.e., $\widehat{\sigma^2_{lh}}(A)$ is minimal. This leads, within the framework of the hypotheses, to an automatic, and also the best, method to estimate the noise standard deviation (see Fig. 7).

The histogram representation is a much more powerful tool for statistical analysis. Not only does it give access to the probability distribution of the population under study, but also a way to

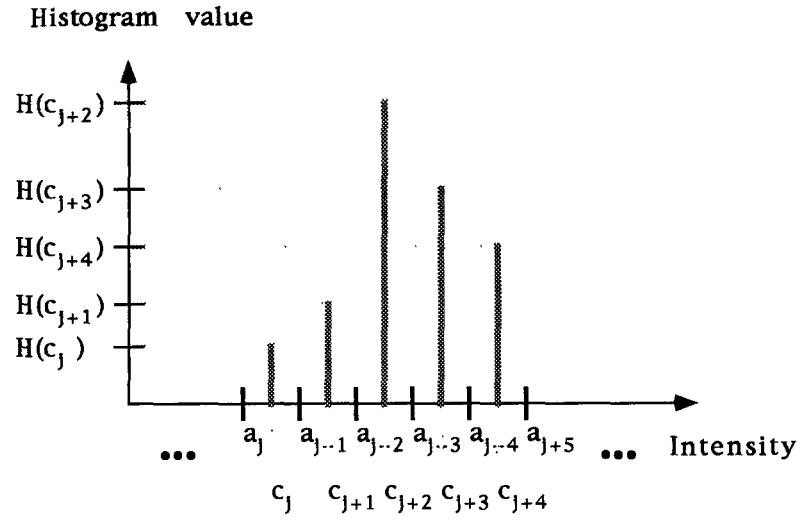

Fig. 1. Histogram representation of a spectrum. The horizontal axis represents the intensity of the points of the spectrum, and is divided in intervals $[a_j, a_{j+1}[$. The vertical axis is the histogram value $H(c_j)$ at c_j, that is, the number of data points whose intensity lies in the interval $[a_j, a_{j+1}[$.

estimate some statistical parameters. Let X be the random variable associated with the intensity of a population of points i, $i \in [1, N_p]$, for example the points of a spectrum, drawn from the same distribution probability. In the case of NMR data, this condition is fulfilled by the noise ergodicity assumption. The values of X are partitioned in an arbitrary number of intervals called bins B_j, $j \in [1, N_B]$, defined as follows:

$$\begin{aligned} &B_j = (c_j, [a_j, a_{j+1}[), \quad j \in [1, N_B] \\ &c_j \in [a_j, a_{j+1}[\,, j \in [1, N_B]; a_{j+1} > a_j, j \in [1, N_B] \end{aligned} \tag{8}$$

The histogram H(j) of the points {i} with bins B_j is then constructed by counting the number of points whose intensity lies in bin j, that is, in interval $[a_j,a_{j+1}[$ (see Fig. 1):

$$H(j) = \text{card}\ \{\, i, i \in [1, N_p] : x_i \in [a_j, a_{j+1}[\} \ , \quad j \in [1, N_B] \tag{9}$$

If f_X is the probability density function of X, then the probability P_j that x falls in bin j is:

$$P_j = \int_{a_j}^{a_{j+1}} f_X(x)dx, \quad j \in [1, N_B] \tag{10}$$

Let k be the histogram value in bin j, that is, H(j). The probability that k of N_p samples fall in bin j is given by the binomial law:

$$P_j(k) = C_{N_p}^k P_j^k (1 - P_j)^{N_p - k}, \quad j \in [1, N_B] \tag{11}$$

The expected value for k results in:

$$E(k) = N_p P_j \tag{12}$$

A good estimate for P_j is deduced from Eq. 12, by replacing k by its value H(j):

$$\widehat{P_j} = \frac{H(j)}{N_p}, \quad j \in [1, N_B] \tag{13}$$

When $[a_j, a_{j+1}]$ is so small that f_X is nearly constant in this interval, Eq. 10 becomes:

$$P_j \simeq f_X(c_j)(a_{j+1} - a_j), \quad j \in [1, N_B] \tag{14}$$

An estimate of $f_X(c_j)$ can then be deduced from Eqs. 13 and 14:

$$\widehat{f_X}(c_j) = \frac{H(j)}{N_p(a_{j+1} - a_j)}, \quad j \in [1, N_B] \tag{15}$$

This relation is the link between the statistical study of the data $\{i\}$, $i \in [1, N_p]$ with the interpretation of the histogram H(j), $j \in [1, N_B]$, and the probability density function f_X of the underlying random variable X.

The identification of $\widehat{f_X}$ with a model leads to an estimation of the model parameters. Considering a Gaussian noise, the histogram strategy gives estimators $\widehat{m_H}$ and $\widehat{\sigma_H}$ of the mean and the standard deviation of the noise (see Fig. 3):

$$\begin{cases} \widehat{\sigma_H} = \dfrac{N(a_{k+1} - a_k)}{\sqrt{2\pi H(k)}} \\ \widehat{m_H} = c_k \end{cases} \quad \text{with } H(k) > H(j)_{j \neq k}, \quad (k, j) \in [1, N_B]^2 \tag{16}$$

For the large amount of data used in 2D and 3D NMR, the ratio of the noise data points to the signal data points is high, and the identification of the mode of the histogram with the maximum of the Gaussian model gives an accurate estimator $\widehat{\sigma_H}$.

RESULTS AND DISCUSSION OF THE NOISE STUDY

The experimental study of the noise properties with the tools presented here is clearly exemplified in the case of 3D NMR spectra, where the large amount of data leads to the most significant statistical results. Any linear transform of a Gaussian process leads to another Gaussian process (Coulon, 1984), so the study in the frequency domain after Fourier transformation is equivalent to the study of the original data set in the time domain.

The ergodicity and Gaussian distribution of the noise are evaluated as shown in Fig. 2, representing the experimental histogram of a 3D homonuclear proton NOE-HOHAHA spectrum of $256 \times 256 \times 256$ points, and the corresponding simulated histogram for a Gaussian distribution. A first histogram is built for the full range of the spectrum values, and then a more accurate histogram is built around the interesting intensities. The Gaussian hypothesis appears to be relevant; the slight variation in the shape of the histogram results from baseline distortions and

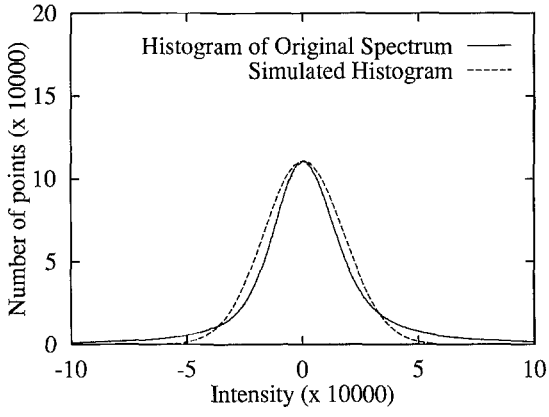

Fig. 2. Histogram of 3D NOE-HOHAHA spectrum of Capsicein in H_2O, pH = 6.35, concentration = 4 mM, T = 318 K. Spectrometer frequency: 600 MHz.

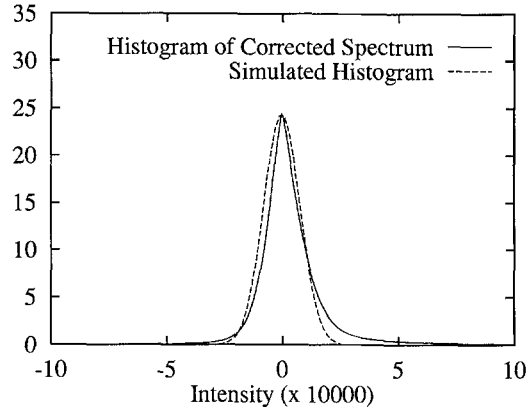

Fig. 3. Histogram of the same spectrum as in Fig. 2, after baseline distortion correction. The identification of the maximum of the experimental distribution with the Gaussian model gives two estimators of the expectation and standard deviation, used to simulate the corresponding Gaussian distribution.

NMR signal, which tend to move some points from low-intensity bins into higher intensity bins. Figure 3 illustrates the influence of baseline distortion correction (Rouh, 1993) when applied to the same data. The difference between the experimental distribution and the Gaussian distribution for positive values now mainly results from the positive lines corresponding to signal in the phase-corrected spectrum.

The results of the different noise estimators and the white-noise hypothesis are discussed together. A noise profile in a 3D NMR spectrum is constructed in the following manner: for each plane perpendicular to a given dimension, the best two-dimensional maximum likelihood estimator of a given size is computed. This profile for an estimator perpendicular to the acquisition dimension, together with the best three-dimensional maximum likelihood estimator for the whole spectrum, and the histogram estimator are represented in Fig. 4. The noise profile is not flat, due to dispersive solvent line components. The maximum likelihood estimator is $\widehat{\sigma_{lh}} = 9775$ and the histogram estimator is $\widehat{\sigma_H} = 17\,269$, i.e., a difference of 43%. This difference is due to the fact that the histogram estimator reflects the whole spectrum, whereas the best maximum likelihood estimator is a local estimator, less influenced by the solvent line. The noise profiles for the planes perpendicular to the other two dimensions are rather flat, confirming the influence of the solvent line.

Figure 5 displays the same estimators when baseline distortion correction is applied, resulting in a drop of the solvent effect. In this case the two estimators are $\widehat{\sigma_{lh}} = 6949$ and $\widehat{\sigma_H} = 7831$, respectively, a difference smaller than 12%. The deviation from a pure white-noise profile results from residual solvent lines and from the t_1 noise (Manoleras and Norton, 1992).

The computation times for a 16×10^6 point spectrum on a Silicon Graphics 4D30 workstation are 8853 and 69 s, respectively, which demonstrates the efficiency of histogram estimation. In addition, the histogram strategy has the ability to validate the probability distribution model, and may use different models apart from the Gaussian one, for example multimode distributions.

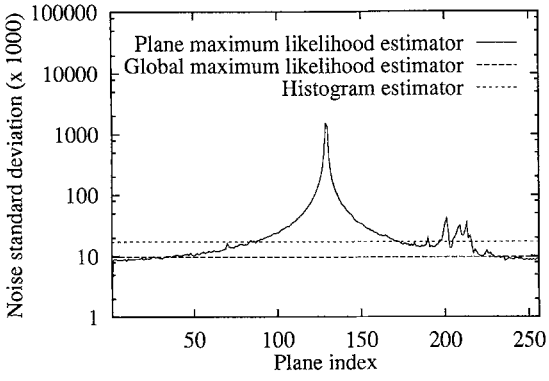

Fig. 4. Noise standard deviation estimators of the same spectrum as in Fig. 2. The lower dashed line is the histogram estimator and the upper one the best maximum likelihood estimator of size $20 \times 20 \times 20$, scanned over the whole spectrum. The continuous curve is a plot of the best maximum likelihood estimator of size 30×30 for each plane perpendicular to the acquisition dimension, versus the plane number. The sizes of the maximum likelihood estimators are chosen with reference to an area free of NMR signal, and large enough so that the biases of the estimators are acceptable.

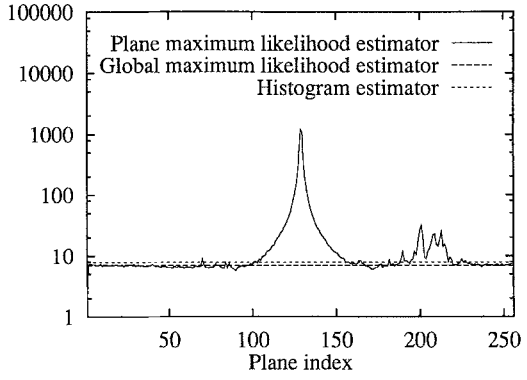

Fig. 5. Noise standard deviation estimators of the same spectrum as in Fig. 2, after baseline distortion correction. The lower dashed line is the histogram estimator and the upper one is the best maximum likelihood estimator of size $20 \times 20 \times 20$ on the whole spectrum. The continuous curve is a plot of the best maximum likelihood estimator of size 30×30 for each plane perpendicular to the acquisition dimension, versus the plane number.

THEORY OF BAYESIAN DETECTION

The second part of this paper is devoted to a statistical method to discriminate signal from noise. The statistical model of the noise, associated with a deterministic model of NMR peaks can be used in a Bayesian detection of the signal (Duda and Hart, 1973). Given a hypothesis H about a phenomenon, the Bayesian approach combines prior knowledge of this hypothesis with an observation $\vec{x}$ of the phenomenon, in order to evaluate the hypothesis (Howson and Urbach, 1991). Bayes' theorem says that if we know the prior probabilities $P(\vec{x})$ and P(H), and the conditional probability $p(\vec{x}|H)$ of $\vec{x}$ given the hypothesis H, we can evaluate the posterior probability $p(H|\vec{x})$ in the following manner:

$$p(H|\vec{x}) = \frac{p(\vec{x}|H)}{P(\vec{x})} P(H) \tag{17}$$

For the sake of clarity the method is presented for the 2D case, but it remains valid for the 3D case.

The procedure operates on a local elementary neighbourhood for each data point, represented by a vector $\vec{x}$. Each data point is regarded as being either a summit or a side (side A, side B, side C and side D) of a peak and is characterized by a special configuration of a 3×3 neighbourhood, $\vec{S}_S$, $\vec{S}_A$, $\vec{S}_B$, $\vec{S}_C$ and $\vec{S}_D$, where the subscripts denote the summit and the four types of side, respectively (see Fig. 6). The signal detection consists of testing the following hypothesis: is there one characteristic configuration in the neighbourhood of each data point?

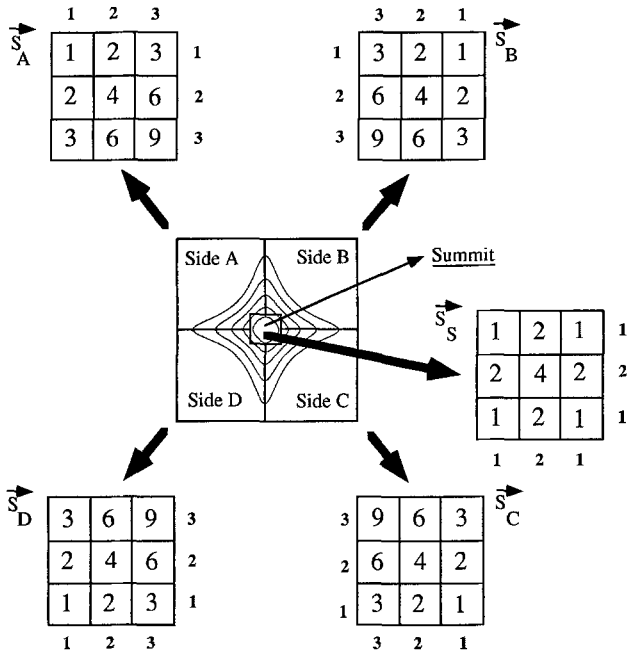

Fig. 6. The different types of points of a 2D NMR Lorentzian line are characterized by a local model, made of a set of special configurations of the intensities in the neighbourhood of the point. These configurations are compared to the spectrum in the statistical detection step, to discriminate the NMR signal from noise in the spectrum.

The alternative approach is to use a set of reference peaks, good and spurious, and classify a candidate peak according to the k-nearest-neighbour rule (Kleywegt et al., 1990). A match factor is calculated between the candidate peak and each reference peak. If the majority of the k-nearest reference peaks are true peaks, the peak will be classified as a true peak, otherwise it will be classified as a false one. Note that the corresponding hypothesis is the presence of a line at a given position, and not the determination of the extension of a peak.

The elementary reference shapes are scaled to fit $\vec{x}$ with respect to a chi-2 criterion:

$$\vec{S}'_y = \alpha \vec{S}_{y,y=S.A.B.C.D} + \beta \vec{1} / \min_{\alpha,\beta} \left[\chi^2_{\alpha,\beta} = (\vec{x} - \alpha \vec{S}_y - \beta \vec{1})^2 \right] \tag{18}$$

Under the hypothesis of a white Gaussian noise, this least-square fitting is the maximum likelihood estimation of α and β (Press et al., 1986). This fit is constrained, so that the resulting shape is a real candidate peak and not an artefact:

$$\begin{cases} \beta > \text{Threshold}_1 \\ \alpha > \text{Threshold}_2 \end{cases} \quad \text{with Threshold}_2 > 0 \tag{19}$$

The first condition corresponds to a minimum ground value of the reference shape, and the second one to an upward positive peak with a minimum energy level.

A hypothesis H_y(y = S, A, B, C, D) is defined as the presence of the shape $\vec{S}'_y$, y = S, A, B, C, D in the neighbourhood $\vec{x}$, with $\vec{S}'_y$, y = S, A, B, C, D different from noise. The null hypothesis H_0 is the lack of signal in $\vec{x}$. The decision process consists of two steps: first a choice of the best hypothesis among H_S, H_A, H_B, H_C and H_D and then a choice between the best hypothesis and the null hypothesis.

The criterion for the first step is the minimisation of the average error probability, which leads to the Bayes decision rule:

$$\text{decide } H_y \text{ if } P(H_y|\vec{x}) > P(H_z|\vec{x})_{z \neq y} \tag{20}$$

where $P(H_y|\vec{x})$ is the probability of the hypotheis H_y when the observation is $\vec{x}$. This rule is equivalent to:

$$\text{decide } H_y \text{ if } P(H_y)p(\vec{x}|H_y) > P(H_z)p(\vec{x}|H_z)_{z \neq y} \tag{21}$$

The second criterion is the minimisation of the average risk of the decision, and the associated rule is:

$$\text{decide } H_y \text{ if } \frac{p(\vec{x}|H_y)}{p(\vec{x}|H_0)} > \frac{P(H_0)}{P(H_y)} \quad \frac{C_{y0} - C_{00}}{C_{0y} - C_{yy}} \tag{22}$$

where the left part of the inequality is the likelihood ratio, and the right part is the threshold of the test. The elements of the threshold are the costs of each type of decision, i.e., C_{00} is the cost to decide H_0 when H_0 is true; C_{y0} is the cost to decide H_y when H_0 is true; C_{0y} is the cost to decide H_0 when H_y is true; and C_{yy} is the cost to decide H_y when H_y is true. The prior probabilities of H_y, y = 0, S, A, B, C, D are estimated from the experimental histogram H_{exp} and the theoretical one H_{theo} using Eq. 15 for a pure noise data set, that is, under the assumption that the density distribution f_X is Gaussian. The NMR signals move some points of the experimental histogram from low intensity bins into higher ones, so the difference between the two distributions provides an estimation of the largest number of signal points N_{Si}:

$$N_{Si} = \frac{1}{2} \sum_{j=1}^{N_B} |H_{theo}(j) - H_{exp}(j)| \tag{23}$$

Following the same reasoning as in Eq. 13, an estimation of the probability of signal P_{Si} is:

$$P_{Si} = \frac{N_{Si}}{N_p} \tag{24}$$

Given the average size A_S of a 2D peak, the estimated prior probabilities P_S of a summit, P_A, P_B, P_C, P_D of a side point, and P_0 of a noise point are:

$$\begin{aligned} P_S &= \frac{P_{Si}}{A_S} \\ P_A = P_B = P_C = P_D &= \frac{(A_S - 1)}{4} P_S \\ P_0 &= 1 - P_{Si} \end{aligned} \tag{25}$$

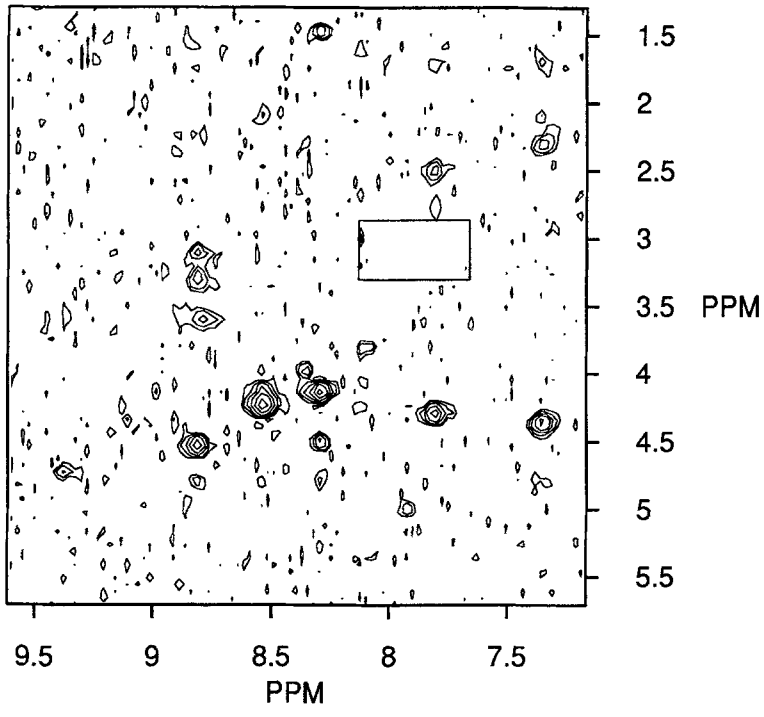

Fig. 7. Extract from the (f_2, f_3) plane 35, perpendicular to the ^{15}N dimension of the 3D HOHAHA-HMQC spectrum of ^{15}N-enriched Capsicein in H_2O, pH = 6.35, concentration = 4 mM, T = 318 K, spectrometer frequency: 400 MHz. The box is the area automatically determined for the calculation of the best maximum likelihood estimators of size 10×20.

The estimated prior probabilities for the hypotheses are directly deduced from these probabilities:

$$\begin{aligned} P(H_0) &= P_0 \\ P(H_S) &= P_S \\ P(H_A) &= P(H_B) = P(H_C) = P(H_D) = P_A \end{aligned} \tag{26}$$

The evaluation of the quantities $P(\vec{x}|H_y)_{y=0,S,A,B,C,D}$ is based on the Gaussian probability distribution of the noise N:

$$p_N(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2\right], \quad i \in [1, N_p] \tag{27}$$

We can simplify this equation by subtracting m from the whole spectrum, so that the new distribution is centered.

As the noise is assumed uncorrelated with the signal, the probability of having $\vec{x}$ under the null hypothesis is equal to the probability that $\vec{x}$ is equal to the noise vector $\vec{n}$:

$$p(\vec{x}|H_0) = p(\vec{x} = \vec{n}|H_0) \tag{28}$$

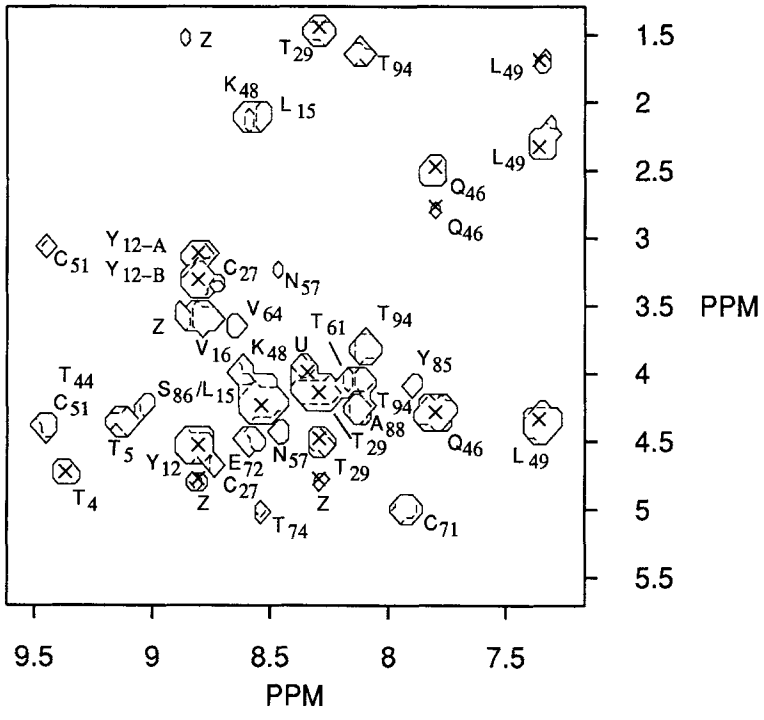

Fig. 8. Solid curve: statistical detection of signal in the same plane area as in Fig. 7; dashed curve: peak extensions determined by a 3D peak-picking algorithm using the statistical detection step; crosses: peak summits in plane 35, calculated from the peak extension. The numbered labels correspond to the manual assignment of the protein, the letter U to an unidentified peak, and the letter Z to artifacts.

The noise components are uncorrelated, based on the white-noise assumption. Consequently, the probability density that there is no signal in the neighbourhood $\vec{x}$ is:

$$p(\vec{x} = \vec{n}|H_0) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^9 \exp\left(\frac{-1}{2\sigma^2} \vec{x}^2 \right) \tag{29}$$

For the other hypotheses, we have:

$$p(\vec{x}|H_y)_{y = S.A.B,C,D} = p(\vec{x} = \vec{n} + \vec{S}'_y|H_y) \tag{30}$$

Under the hypothesis that the noise and the signal are uncorrelated, Eq. 30 becomes:

$$p(\vec{x}|H_y)_{y = S,A,B,C.D} = p(\vec{n} = \vec{x} - \vec{S}'_y|H_y) \tag{31}$$

so:

$$p(\vec{x}|H_y)_{y = S,A,B,C,D} = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^9 \exp\left[\frac{-1}{2\sigma^2} (\vec{x} - \vec{S}'_y)^2 \right] \tag{32}$$

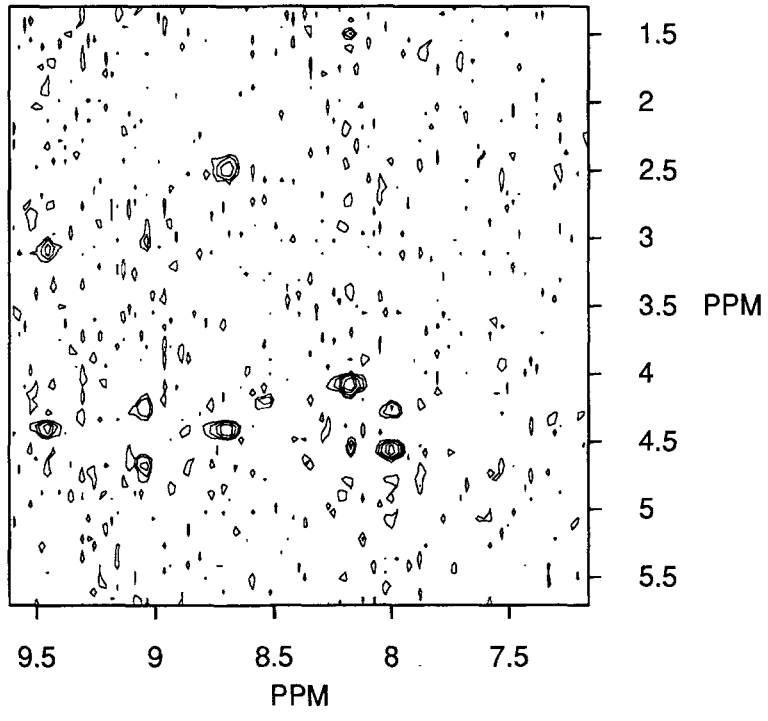

Fig. 9. Extract from the (f_2, f_3) plane 37 perpendicular to the ^{15}N dimension of the 3D HOHAHA-HMQC spectrum of ^{15}N-enriched Capsicein in H_2O, pH = 6.35, concentration = 4 mM. T = 318 K, spectrometer frequency: 400 MHz.

The logarithm of rules 21 and 22 and the use of Eqs. 29 and 32 give the final detection rule, i.e., decide $H_{y,y=0,S,A,B,C,D}$ that maximises the following quantity:

$$\begin{cases} \text{if } y = 0: & \sigma^2 \left[\log P(H_0) + \log \left(\dfrac{C_{y0} - C_{00}}{C_{0y} - C_{yy}} \right) \right] \\ \text{otherwise:} & \vec{x}.\vec{S}'_y - \dfrac{\vec{S}_y^{\,2}}{2} + \sigma^2 \log P(H_y) \end{cases} \tag{33}$$

In 3D NMR experiments, while the number of data points increases, the density of signal and artifact data points decreases. Consequently, the prior probability estimation is better than in the 2D case. An extended peak model is then made of a set of 3D shapes in a three-dimension $3 \times 3 \times 3$ elementary neighbourhood, and the vectorial signal representation enables a straightforward generalisation from the 2D case.

RESULTS AND DISCUSSION

The strategy of statistical detection of NMR signal is exemplified by a complete peak-picking algorithm that will be described elsewhere. The first step of this algorithm is the signal detection, performed by the Bayesian detection presented above. The support of each individual peak is then built, separating overlapping peaks when the summits are distinct. The last step consists in

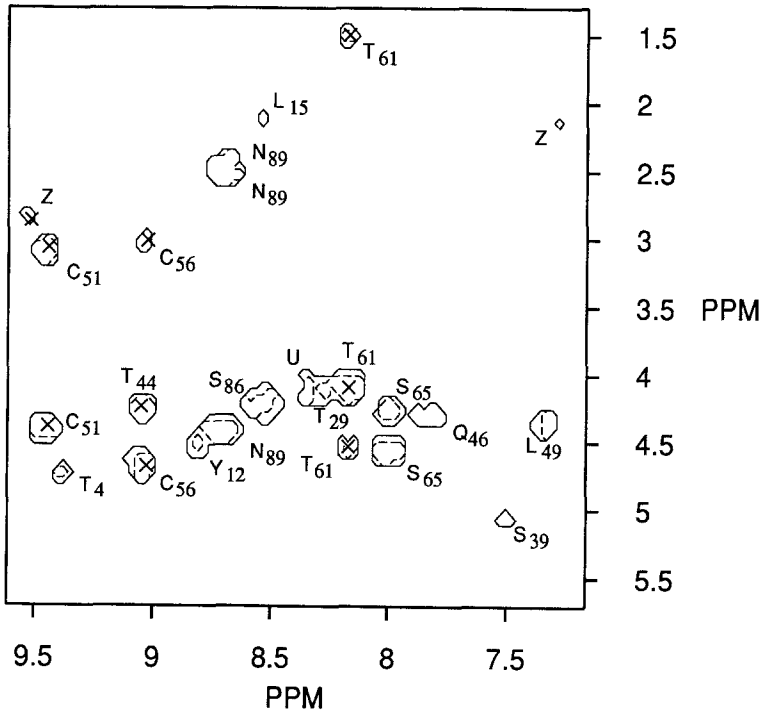

Fig. 10. Solid curve: statistical detection of signal in the same plane area as in Fig. 9; dashed curve: peak extensions determined by a 3D peak-picking algorithm using the statistical detection step; crosses: peak summits in plane 37, calculated from the peak extension. The numbered labels correspond to the manual assignment of the protein, the letter U to an unidentified peak, and the letter Z to some noise.

measuring the parameters of each peak, in particular the summit coordinates. All these procedures are implemented in a program named GIFC*, designed to process and display 2D and 3D spectra. The peak-picking using a $3 \times 3 \times 3$ lineshape model was applied to a 3D $64 \times 256 \times 256$ points heteronuclear HOHAHA-HMQC spectrum of Capsicein, a protein of 98 residues. The computation time on a Silicon Graphics 4D30 workstation was 269 min and 807 peaks were detected.

A contour plot representation of an area of the (f_2, f_3) planes 35 and 37 is shown in Figs. 7 and 9. In Figs. 8 and 10, the results of signal detection are represented by solid lines, the peak extensions after the complete peak-picking procedure are represented by dashed lines, and the summits found in each plane are marked by crosses. The numbered letters correspond to the manual assignment of the spectrum, representing the best way to evaluate the quality of the results.

The detection step appears to be reliable, since every peak found by manual examination is found by the algorithm, and there is little residual noise, denoted by the letter Z, apart from the two false peaks found in plane 35, which result from the residual solvent line. It is worth noting that the peaks Q_{46}, L_{49} and Y_{12}, found in plane 37, result from the influence of some peaks that lie

* A copy of the source code for the routines described in this paper is available from the authors.

in plane 35. This is due to the extension of the reference shapes, which gives a good discrimination between noise and signal, but with a loss of locality. Moreover, the rest of the process is able to find the correct extensions of the peaks, and the deduced summits are the same as those manually determined.

CONCLUSIONS

In conclusion, we have developed an original procedure for signal detection in noisy spectra. Based upon a noise estimation by means of the statistical interpretation of spectrum histograms, it allows line-shape preservation. Bayesian detection proves to give better results than a standard contour plot. Our new peak-picking procedure provides accurate results for automatic assignment of multidimensional NMR spectra.

ACKNOWLEDGEMENTS

The authors thank S. Bouaziz for providing the experimental spectra, and J. Timmerman, R. Klinck and E. Jolivet (Département de Biométrie, I.N.R.A., Jouy en Josas, France) for reading the manuscript.

REFERENCES

Arques, P.Y. (1982) *Décision et Traitement du Signal*, Masson, Paris.
Borgias, B.A. and James, T.L. (1988) *J. Magn. Reson.*, **79**, 493–512.
Coulon, F. (1984) *Théorie et Traitement des Signaux*, Presses Polytechniques Romandes, Lausanne.
Delsuc, M.A. (1989) In *Maximum Entropy and Bayesian Methods* (Ed., Skilling, J.) Kluwer, Dordrecht, pp. 285–290.
Duda, R.O. and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*, Wiley, New York, NY.
Ernst, R.R., Bodenhausen, G. and Wokaun, A. (1987) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford.
Howson, C. and Urbach, P. (1991) *Nature*, **350**, 371–387.
Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B. and Bax, A. (1992) *Nature*, **256**, 632–638.
Kleywegt, G.J., Boelens, R. and Kaptein, R. (1990) *J. Magn. Reson.*, **88**, 601–608.
Manoleras, N. and Norton, R.S. (1992) *J. Biomol. NMR*, **2**, 485–494.
Nibedita, R.N., Kumar, R.A., Majumdar, A. and Hosur, R.V. (1992) J. Biomol. NMR, **2**, 467–476.
Press, W.H., Flannery, B.P., Tevkolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge.
Rouh, A., Delsuc, M.A., Bertrand, G. and Lallemand, J.Y. (1993) *J. Magn. Reson.*, **102**, 357–359.
Stoven, V., Mikou, A., Piveteau, D., Guittet, E. and Lallemand, J.Y. (1989) *J. Magn. Reson.*, **82**, 163–168.
Zolnai, Z., Macura, S. and Markley, J.L. (1988) *J. Magn. Reson.*, **80**, 60–70.