JOURNAL OF MAGNETIC RESONANCE **129,** 165–172 (1997)
ARTICLE NO. MN971241

Use of Global Symmetries in Automated Signal Class Recognition by a Bayesian Method

Anja-Carina Schulte,* Adrian Görler,* Christof Antz,* Klaus-Peter Neidig,† and Hans Robert Kalbitzer*,‡,[1]

*Department of Biophysics, Max-Planck-Institute for Medical Research, Jahnstrasse 29, D-69028 Heidelberg, Germany; ‡Department of Biophysics, University of Regensburg, Universitätsstrasse 31, D-93040 Regensburg, Germany; and †Bruker Analytische Messtechnik, D-76287 Rheinstetten, Germany

Received March 7, 1997

Automated or semiautomated pattern recognition in multidimensional NMR spectroscopy is strongly hampered by the large number of noise and artifact peaks occurring under practical conditions. A general Bayesian method which is able to assign probabilities that observed peaks are members of given signal classes (e.g., the class of true resonance peaks or the class of noise and artifact peaks) was proposed previously. The discriminative power of this approach is dependent on the choice of the properties characterizing the peaks. The automated class recognition is improved by the addition of a nonlocal feature, the similarities of peak shapes in symmetry-related positions. It turns out that this additional property strongly decreases the overlap of the multivariate probability distributions for true signals and noise and hence largely increases the discrimination of true resonance peaks from noise and artifacts. © 1997 Academic Press

INTRODUCTION

A practical problem often encountered during the evaluation of multidimensional data is the occurrence of noise and artifact peaks which may result in ambiguous and erroneous assignments. This problem is particularly severe in automated assignment procedures where the occurrence of a large number of noise or artifact peaks can lead to an instability of the algorithms giving wrong results or alternatively a too large number of possible true solutions. As a practical solution for this problem, most more advanced program packages have interactive routines which allow one to remove wrong solutions at any stage of the evaluation. Since this interactive work is tedious, it is desirable to transfer at least part of the work to the computer.

To discriminate true resonance peaks from noise or artifact peaks, it is necessary to know features characteristic for each class of those peaks. Typical features which can be taken into account are local properties such as peak intensities and peak shapes or global spectral features such as t_1 ridges or the presence of symmetry-related partners (*1–7*).

Bayesian reasoning is a powerful statistical method of great flexibility. Its operational conditional formalism allows an easy expansion to more general and multivariate cases. Recently, we reported a Bayesian method coupled to a multivariate linear discriminant analysis (*8*). The actual implementation was limited to the use of local properties of resonance and artifact peaks. An important global property is the spectral symmetry, which occurs in many types of homonuclear two-dimensional spectra such as NOESY and TOCSY spectra. In the past this spectral feature was mainly used for an improvement of the spectral quality by the symmetry enhancement (*6, 9–14*). In the present paper we will show how it can be used for an efficient signal class recognition in two-dimensional NMR spectra.

MATERIALS AND METHODS

Software

Peak picking and integration was performed using the standard routines of the program package AURELIA (*15*). The method used by AURELIA for peak picking has been outlined earlier (*6*). The software developed and described here is implemented in the program package AURELIA but a stand-alone version can also be obtained from the authors.

NMR Spectroscopy

HPr protein from *Staphylococcus carnosus* was isolated as described (*16*). The sample contained 3.1 mM HPr protein and 0.05 mM EDTA in 500 μl of 95% H_2O/5% D_2O. The NOESY spectrum (*17*) was recorded on a Bruker AMX-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 (*18*). The 1K × 4K time domain data were recorded and transformed to different sizes as indicated. Frequency data were referenced to the internal standard 4,4-dimethyl-4-silapentane sulfonic acid (DSS).

THEORETICAL CONSIDERATIONS

According to Bayes's theorem (*19, 20*) the probability $P_j(C_\lambda|\mathbf{E}^j)$ that cross peak j with values E_k^j of the properties

[1] To whom correspondence should be addressed.

1090-7807/97 $25.00

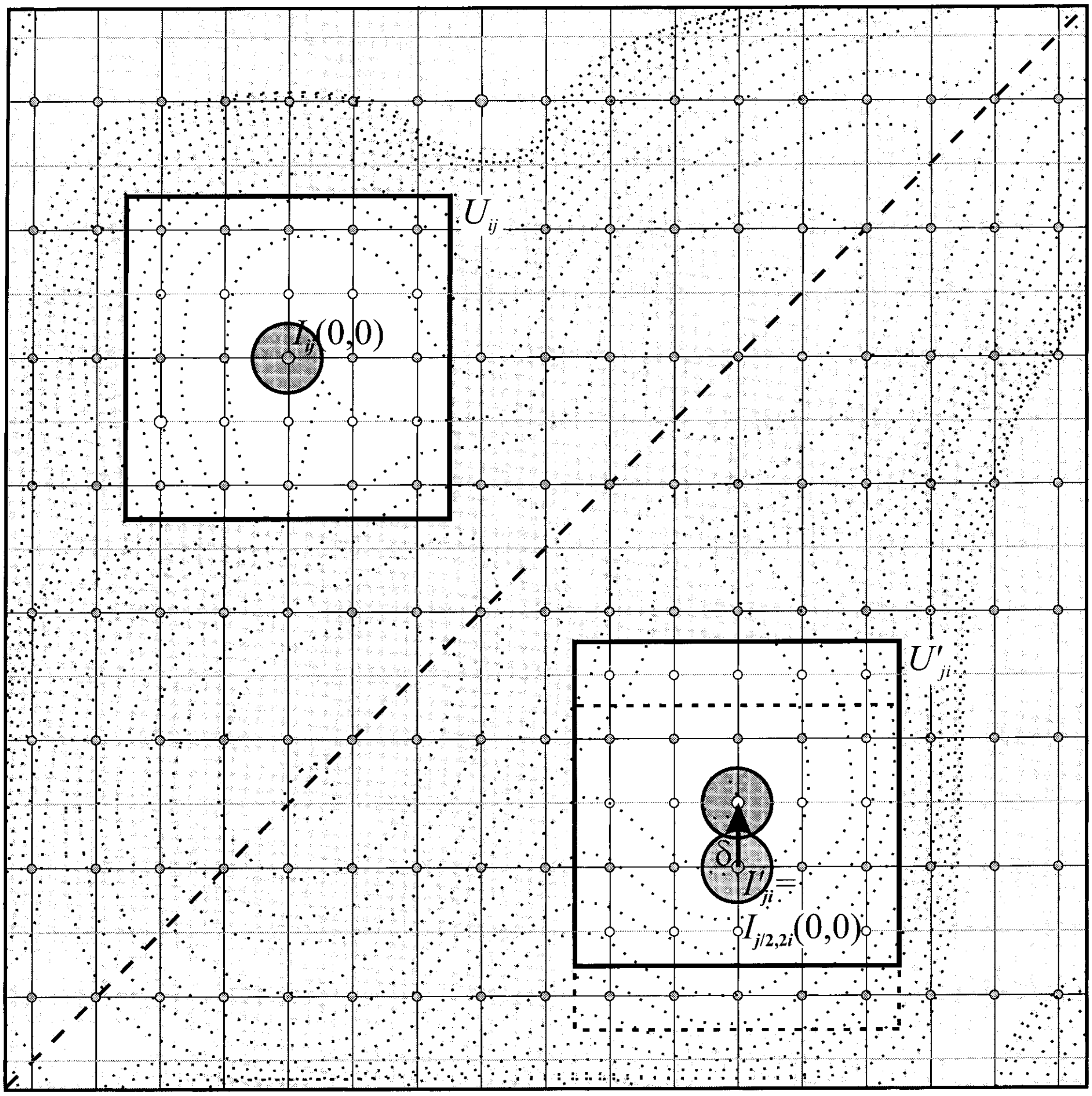

FIG. 1. Interpolation of asymmetric spectral areas. If the dimensions of the frequency domain data SI1 and SI2 are not equal, the discrete grid (gray dots) of the spectral areas U_{ij} and U'_{ji} used to calculate the match factor of a peak I_{ij} contains a different number of data points in both dimensions. To calculate the match factor M intermediate data points (white dots) are calculated by linear interpolation. If the peak I'_{ji} which is symmetric to I_{ij} does not lie on the discrete (gray) grid, the area U'_{ji} must be shifted by an offset δ before calculation of the match factor.

E_k ($k = 1, \ldots, K'$) belongs to class C_λ ($\lambda = 1, \ldots, \Lambda$) can be calculated as

$$P_j(C_\lambda|\mathbf{E}^j) = \frac{P(C_\lambda)P(\mathbf{E}^j|C_\lambda)}{\Sigma_{\mu=1}^{\Lambda} P(C_\mu)P(\mathbf{E}^j|C_\mu)} \qquad [1]$$

with the a priori probability $P(C_\lambda)$ of finding a peak of class C_λ and $P(\mathbf{E}^j|C_\lambda)$ the conditional probability of finding the properties $\mathbf{E}^j$ for a peak of class C_λ. If the multivariate probability distribution $p(\mathbf{E}|C_\lambda)$ is not known a priori, it must be obtained from a sample which, in general, must be sufficiently large in order to describe the multidimensional distribution function p 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 individual univariate distributions $p(E_k|C_i)$. If Q variables are independent and the remaining $K = K' - Q$ variables are correlated, the correlated variables can be replaced by a set of L new orthogonal variables Y_k by a linear discriminant analysis. The probability P_j for a given peak j to be member of class C_λ is then given by (*8*)

$$P_j(C_\lambda|Y_1, \ldots, Y_L, E_1, \ldots, E_Q)$$
$$= \frac{P(C_\lambda)\prod_{k=1}^{L} P(Y_k^j|C_\lambda)\prod_{i=1}^{Q} P(E_i^j|C_\lambda)}{\Sigma_{\mu=1}^{\Lambda} P(C_\mu)\prod_{k=1}^{L} P(Y_k^j|C_\mu)\prod_{i=1}^{Q} P(E_i^j|C_\mu)}. \qquad [2]$$

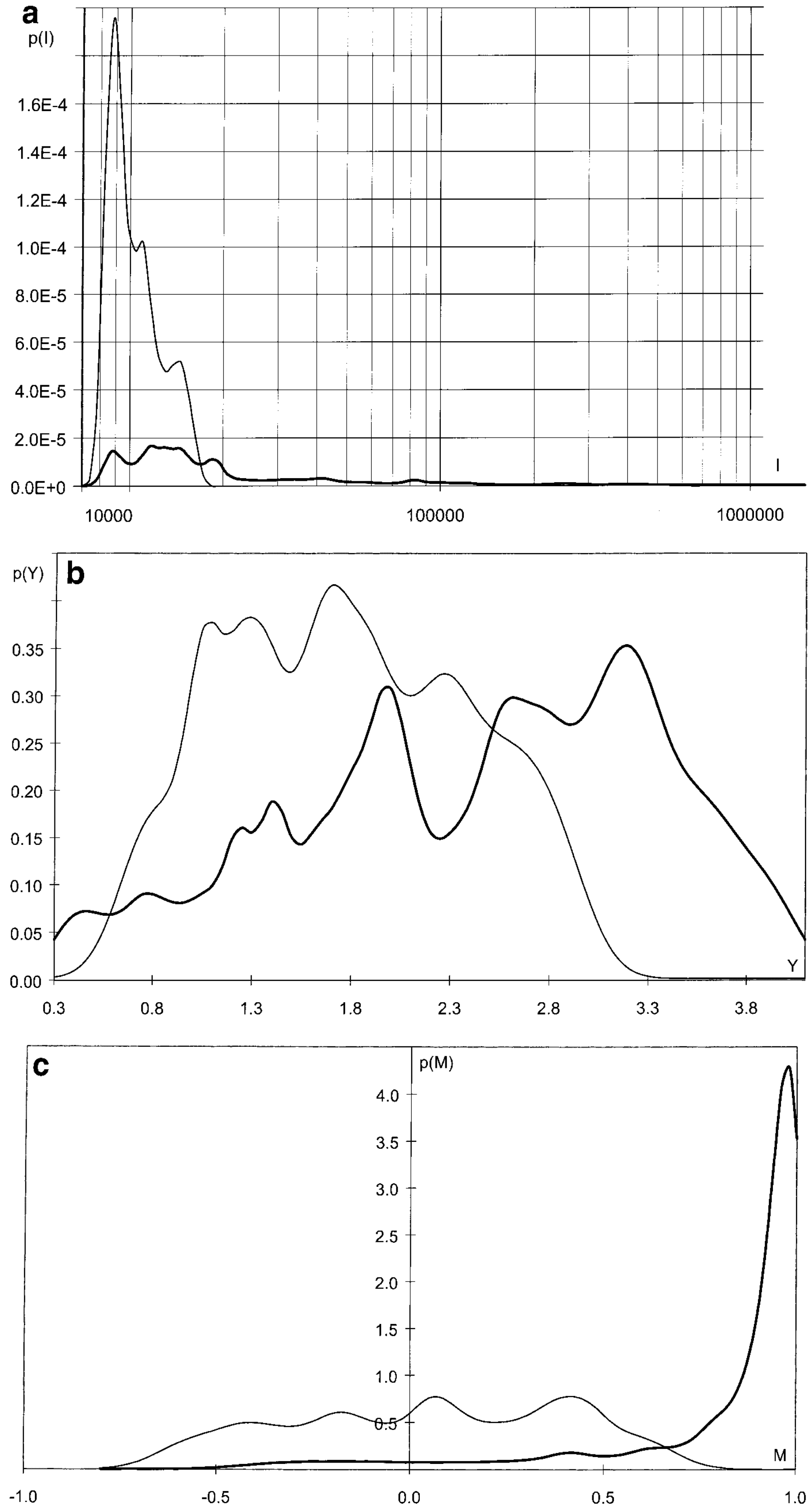

FIG. 2. Probability distributions for noise and true resonance signals. Probability distributions obtained in a Gaussian-filtered NOESY spectrum of HPr from *S. carnosus* (1024 × 1024 real frequency domain data points after Fourier transformation of 1024 × 1024 time domain data points). Thick lines, true signals; thin lines, noise and artifact signals. The distributions were smoothed as described in the text; (a) $p(I)$ is the intensity distribution, (b) $p(Y)$ is the probability distribution for the Y factor as described by Antz *et al.* (*8*), and (c) is the $p(M)$ is the probability distribution for the match factor M.

The already published signal recognition procedure made use of only local properties of the cross peaks (*8*). If one assumes that the global symmetry of the spectrum is statistically independent of the local properties describing a peak, this property can be easily added to the already existing properties using the basic equation [2]. What is essentially

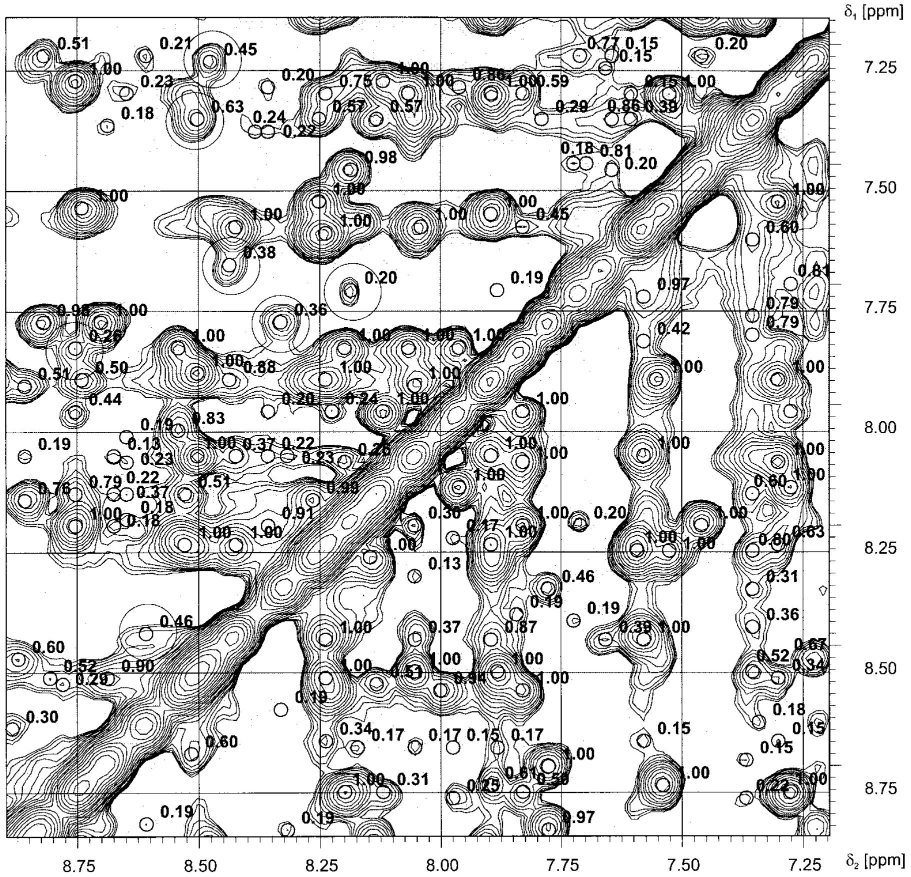

FIG. 3. Improvement of the signal class recognition by the use of spectral symmetry. Part of the spectrum used in Fig. 2; the probabilities of the peaks to be true resonance signals are indicated: (left) without use of the symmetry relation and (right) including global symmetry. A few peaks where the use of the symmetry relation is especially important are highlighted.

needed consists in a measure for the occurrence of a cross peak with similar shape at a symmetry-related position.

A very simple and normalized similarity measure for quantifying the property ''global symmetry'' for a given cross peak is the match factor *M* defined earlier for symmetry enhancement in NMR spectra (*13*). In its elementary form it is defined as

$$M'_{ij} = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}'_{ji}}{|\mathbf{r}_{ij}|\ |\mathbf{r}'_{ji}|}\,, \qquad [3]$$

with the vector $\mathbf{r}_{ij}$ containing the data points $\mathbf{I}_{ij}(\mu, \nu)$ of a rectangular area around the cross peak which have an offset of (μ, ν) grid units with respect to the cross peaks maximum at (i, j), and the vector $\mathbf{r}'_{ji}$ containing the data at the environment of the symmetry-related position (j, i). The match factor M'_{ij} equals 1 for peaks with identical shape and is independent of the absolute intensity. Since base plane variations may influence the effective peak shape, generally it is better to use the match factor in its offset corrected form (*14*)

$$M_{ij} = \frac{\mathbf{r}^*_{ij} \cdot \mathbf{r}'^*_{ji}}{|\mathbf{r}^*_{ij}|\ |\mathbf{r}'^*_{ji}|} \qquad [4]$$

with $\mathbf{r}^*_{ij} = \mathbf{r}_{ij} - \langle \mathbf{r}_{ij} \rangle\, 1$ and

$$\langle \mathbf{r}_{ij} \rangle = \frac{1}{(2n+1)(2m+1)} \sum_{\nu=-n}^{n} \sum_{\mu=-m}^{m} I_{ij}(\mu, \nu), \qquad [5]$$

with *n* and *m* the dimensions of the rectangular area.

PRACTICAL IMPLEMENTATION

As already described for the signal class discrimination from local properties (*8*) the probability distributions for noise and signal peaks are obtained by analyzing their distribution in specific spectral learning areas defined by the user. The definition of an area that contains only noise and artifact peaks but no true signals can usually be performed without larger difficulties. This is not possible for the learning area used for the recognition of true signal peaks since any area in the spectrum contains also noise peaks. However, since in a first approximation the occurrence of a symmetry-related true signal does not depend on the absolute intensity, the probability distribution can

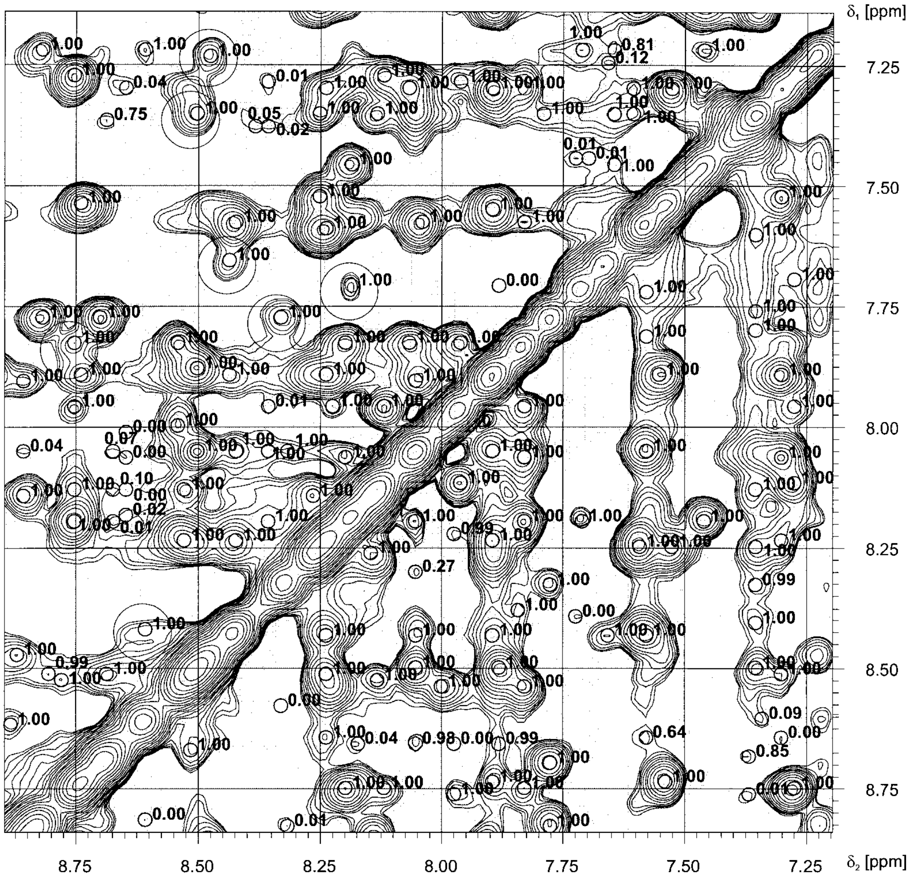

FIG. 3—*Continued*

be constructed from the most intense peaks, which are likely not to be noise peaks.

If one of the symmetry-related cross peaks is located on a t_1-noise ridge, it does not make sense to infer the probability of being a true signal from its calculated match factor. In this case an extreme difference in peak intensities of the compared cross peak areas would be observed. Only peaks in the learning area of true signals for which

$$0.1 < \frac{\langle \mathbf{r}_{ij} \rangle}{\langle \mathbf{r}'_{ji} \rangle} < 10 \qquad [6]$$

holds are included in the construction of the corresponding probability distribution. Analogously, if the intensities of a peak in the test area and its symmetric partner do not fullfill the criterion [6], the global symmetry is neglected when calculating the probability that this peak belongs to the signal class.

Under practical conditions the inherent global symmetry of two-dimensional NMR spectra is often perturbed by an asymmetry resulting from the data acquisition or processing. Although the match factor defined by the relations [3] to [5] is rather tolerant concerning these artifacts, ideally one would correct for these factors before the calculation of the symmetry match. A very common spectral asymmetry arises from differences in the number of data points recorded in the t_1 and t_2 directions. This asymmetry is not removed by a suitable zero-filling of the time domain data even if the same number of data points in the two dimensions has been obtained after Fourier transformation.

If the number of data points in the two dimensions of the frequency domain is not equal, the match factor can no longer be calculated according to Eq. [4]. The problems arise because a part of the symmetry-related data points to be compared does not exist in the digital grid of the experimental data. The situation is complicated by the fact that, even if such a symmetry-related grid point exists the ''true'' peak maxima (to be observed at infinite resolution) usually are not located exactly on the grid points but are now projected to an inherently asymmetric digital grid. The influence of these effects on the match factor can be reduced by transiently decreasing the digital resolution of the two areas compared to the lower limit (usually determined by the ω_1 dimension) or by interpolation of the missing data points. From these two possible methods the second method turned out to give superior results: the two spectral regions compared are enlarged by a linear interpolation prior to the calculation of the match factor (Fig. 1). The other alternative, decreasing the digital resolution, results in areas which are too small for the calculation of a match factor with high significance.

Since the probability distributions are deduced from a limited, discrete sample, the curves obtained must be smoothed. This has been done previously by a floating binomial average filter (*8*). In addition, values of the obtained probability distribution equal to 0 must be corrected and set to a small but finite value to allow well-defined probability calculations according to Eq. [2] because they are usually artifacts caused by the finite size of the sample.

To overcome these problems that can arise from small numbers of peaks in the training areas, in the new implementation of the program a different method of constructing a smooth distribution function has been chosen. The properties of the sample peaks are no longer ordered in classes of equal widths; instead, a distribution is constructed from classes with variable widths. First, the occurring values of properties E_k^j are sorted by magnitude, leading to $\cdots < E_k^{j-1} < E_k^j < E_k^{j+1} < \cdots$. From these sorted values a smoothed distribution is constructed which assumes that the 11 values $E_k^{j-5}, \ldots, E_k^{j+5}$ are normally distributed about their mean

$$\langle E_{k,j}\rangle = \frac{1}{11}\sum_{\alpha=-5}^{5} E_k^{j+\alpha} \qquad [7]$$

with the variance

$$\sigma_{k,j}^2 = \frac{1}{11}\sum_{\alpha=-5}^{5} (E_k^{j+\alpha} - \langle E_{k,j}\rangle)^2 + \sigma_{0,k}^2. \qquad [8]$$

The probability distribution of the property *k* can then be approximated as the sum of normalized Gaussians,

$$p(C_k)(x) = \frac{1-p_{0,k}}{J}\sum_{j=1}^{J}\frac{1}{\sigma_{k,j}\sqrt{2\pi}} \times \exp\left(-\frac{1}{2}\left(\frac{x-\langle E_{k,j}\rangle}{\sigma_{k,j}}\right)^2\right) + p_{0,k}, \qquad [9]$$

where $\sigma_{0,k}$ and $p_{0,k}$ are small values correcting for the limited accuracy of the values of the properties and for too small values of the distributions arising from the finite sample size, respectively. In the current implementation $p_{0,k}$ is chosen to fulfill

$$\int_{\text{range of values}} p_{0,k}\,dx = 0.005. \qquad [10]$$

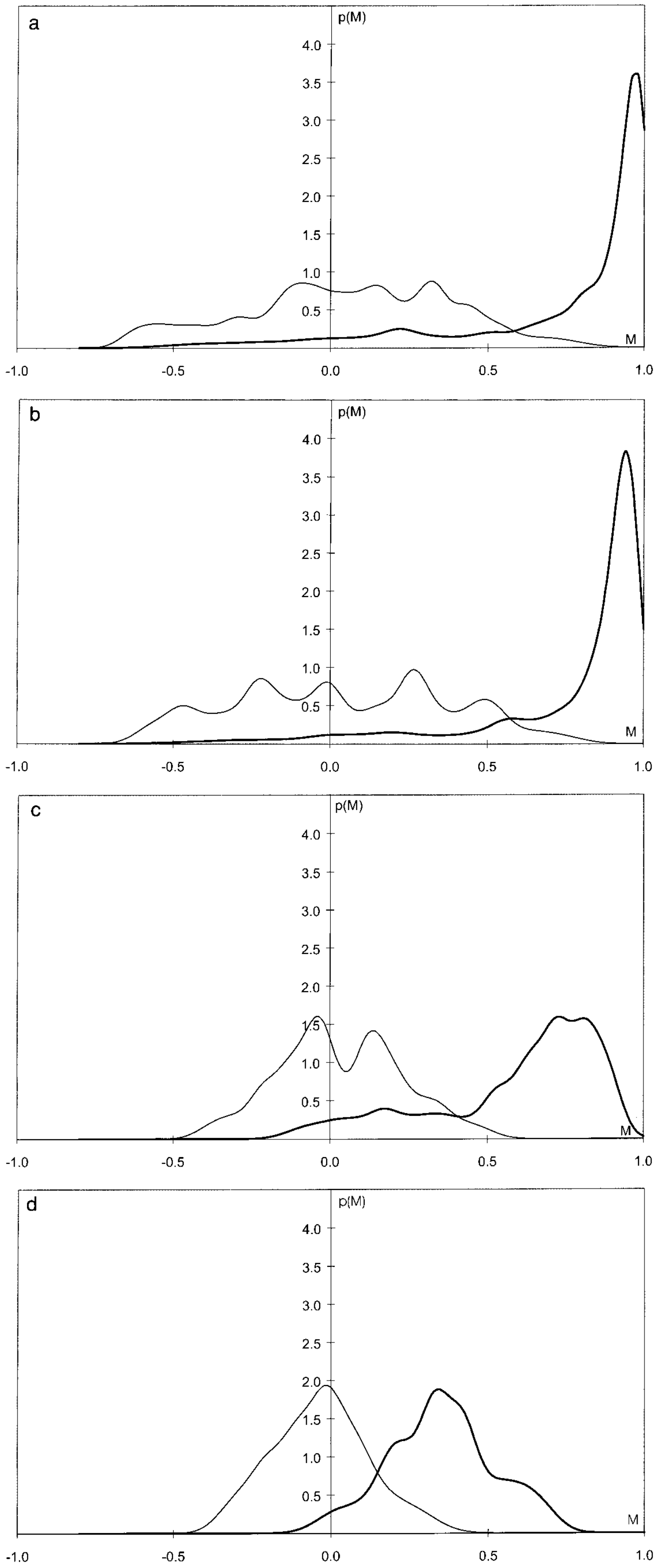

FIG. 4. Dependence of match factor distributions on data processing. The same NOESY data set as in Fig. 2 was used but processed in different ways. (a) 512 × 1024 time domain data points TD1 and TD2, Gaussian filtering of the data before Fourier transformation using the same filter function in both dimensions, and 1024 × 1024 real frequency domain data points SI1 and SI2. (b) As (a) but SI1 = 512; (c) TD1 = 256, TD2 = 1024, SI1 = 256, SI2 = 1024; (d) as (c) but SI1 = 128.

This method of constructing the probability distributions has the favorable property that the distribution functions obtained are well defined everywhere and are larger than zero in the entire range of valid values (in the case of the match factor M, this is the closed interval $[-1, 1]$).

RESULTS AND DISCUSSION

Although the general Bayes theorem is applicable for any probability distribution (objective or subjective), the discrimination power in signal class recognition depends on the separation of the multivariate distribution functions for noise and true signals. Figure 2 shows an example for the probability distributions for different properties obtained for a NOESY spectrum of HPr from *S. carnosus.* The intensity distribution (Fig. 2a) discriminates well between noise signals and true signals for high intensities. For peaks with low intensities the information obtained from the reduced variable Y which characterizes the peak shape is necessary (*8*). However, the Y distributions for noise and true peaks have a large range of overlap (Fig. 2b). In contrast, the distributions of the match factor for the two classes are very well separated (Fig. 2c).

With these distributions the probabilities of cross peaks being true resonance signals were calculated. Figure 3 depicts a typical part of the NOESY spectrum with the obtained probabilities indicated. Without using the symmetry information (Fig. 3, left), in general a satisfactory result is obtained for the majority of the cross peaks. However, a number of very weak cross peaks cannot be recognized safely on the basis of the intensity and peak shape information only. With addition of the symmetry information (Fig. 3, right), these peaks have high signal probabilities when a symmetrical partner with similar shape was found, and decreased probabilities if no partner in the symmetry-related position with similar shape was found. A few examples where the improvement of the recognition procedure is evident are highlighted in Fig. 3.

The underlying symmetry in the two-dimensional time domain data can be destroyed to some degree in the frequency domain by inappropriate processing of the data. Consequently, the discrimination power of the match factor for noise and true signals is reduced. The main source for asymmetries is the use of different digital resolutions in the frequency domain or different degrees of zero-filling before Fourier transformation. Another possibility for reducing the symmetry in the spectra is the application of different filter functions in the two dimensions.

Figure 4 shows some examples of the effects of data processing on the obtained probability distributions of the match factor M. The same data set was used for the calculations as in Fig. 2, but the processing parameters were varied. In Fig. 2c the same degree of zero-filling (onefold) in the t_1 and t_2 directions and the same digital resolutions in the δ_1 and δ_2 dimensions after Fourier transformation were used. Figure 4a shows that a twofold zero-filling in the t_1 direction and a onefold zero-filling in the t_2 direction does not change the probability distributions of M much. The noise distribution is still centered about its theoretical value of zero and well separated from the match factor distribution of true signals with its maximum near $M = 1$. In the second example, the number of data points in the δ_1 dimension was reduced by a factor of 2 (Fig. 4b). The distributions obtained are still rather well separated but the number of peaks with M close to 1 decreases; that is, even after interpolation the peaks have increasingly different shapes. Even in extreme cases where the size of the two dimensions differs by a factor of 4 (Fig. 4c) or 8 (Fig. 4d) the construction of $p(M)$ can be handled satisfactorily by the expansion procedure described above. The probability distributions obtained are still well separated although the ideal symmetry ($M = 1$) is clearly disturbed.

In summary, the inclusion of symmetry information in the signal recognition procedure is a valuable method for improving the discrimination power of the method. However, the use of this information requires that the data processing conserve this symmetry. If there are good reasons to destroy this symmetry in the spectrum under consideration (e.g., if the digital resolution of the time domain data in the t_1 and t_2 directions is very different), the full symmetry information can still be used by applying the recognition procedure to a second spectrum which is symmetrically processed. The probabilities obtained can then be transferred to the asymmetrically processed spectrum.

ACKNOWLEDGMENTS

This work was supported by the European Union. We thank W. Hengstenberg for providing the sample of HPr from *S. carnosus* used in this work.

REFERENCES

1. K.-P. Neidig, H. Bodenmueller, and H. R. Kalbitzer, *Biochem. Biophys. Res. Commun.* **125,** 1143 (1984).
2. S. Glaser and H.-R. Kalbitzer, *J. Magn. Reson.* **74,** 450 (1987).
3. G.-J. Kleywegt, R. M. J. N. Lamerich, R. Boelens, and R. Kaptein, *J. Magn. Reson.* **85,** 186 (1989).
4. G.-J. Kleywegt, R. Boelens, and R. Kaptein, *J. Magn. Reson.* **88,** 601 (1990).
5. V. Stoven, A. Mikou, D. Piveteau, E. Guittet, and J.-Y. Lallemand, *J. Magn. Reson.* **82,** 163 (1989).
6. K.-P. Neidig, R. Saffrich, M. Lorenz, and H. R. Kalbitzer, *J. Magn. Reson.* **89,** 543 (1990).
7. A. Rouh, A. Louis-Joseph, and J.-Y. Lallemand, *J. Biomol. NMR* **4,** 505 (1994).
8. C. Antz, K.-P. Neidig, and H. R. Kalbitzer, *J. Biomol. NMR* **5,** 287 (1995).
9. P. H. Bolton, *J. Magn. Reson.* **68,** 180 (1986).

10. R. Baumann, A. Kumar, R. R. Ernst, and K. Wüthrich, *J. Magn. Reson.* **44,** 76 (1981).

11. R. Baumann, G. Wider, R. R. Ernst, and K. Wüthrich, *J. Magn. Reson.* **44,** 402 (1981).

12. P. H. Bolton, *J. Magn. Reson.* **67,** 391 (1986).

13. K.-P. Neidig and H. R. Kalbitzer, *Magn. Reson. Chem.* **26,** 848 (1988).

14. K.-P. Neidig and H. R. Kalbitzer, *J. Magn. Reson.* **91,** 155 (1991).

15. K.-P. Neidig, M. Geyer, A. Görler, C. Antz, R. Saffrich, W. Beneicke, and H. R. Kalbitzer, *J. Biomol. NMR* **6,** 255 (1995).

16. R. Kruse, W. Hengstenberg, W. Beneicke, and H. R. Kalbitzer, *Protein Eng.* **6,** 417 (1993).

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

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

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

20. J. Cornfield, *Biometrics* **25,** 643 (1969).