JOURNAL OF MAGNETIC RESONANCE **135,** 288–297 (1998)
ARTICLE NO. MN981570

Automated Peak Picking and Peak Integration in Macromolecular NMR Spectra Using AUTOPSY

Reto Koradi,*[,1] Martin Billeter,*[,2] Max Engeli,† Peter Güntert,* and Kurt Wüthrich*[,3]

**Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg, CH-8093 Zürich, Switzerland; and †Institut für Werkzeugmaschinen und Fertigung, Eidgenössische Technische Hochschule-Zentrum, CH-8092 Zürich, Switzerland*

Received May 1, 1998

A new approach for automated peak picking of multidimensional protein NMR spectra with strong overlap is introduced, which makes use of the program AUTOPSY (*aut*omated *p*eak picking for NMR *s*pectroscop*y*). The main elements of this program are a novel function for local noise level calculation, the use of symmetry considerations, and the use of lineshapes extracted from well-separated peaks for resolving groups of strongly overlapping peaks. The algorithm generates peak lists with precise chemical shift and integral intensities, and a reliability measure for the recognition of each peak. The results of automated peak picking of NOESY spectra with AUTOPSY were tested in combination with the combined automated NOESY cross peak assignment and structure calculation routine NOAH implemented in the program DYANA. The quality of the resulting structures was found to be comparable with those from corresponding data obtained with manual peak picking.

Key Words: **peak picking; peak integration; noise level calculation; lineshapes; AUTOPSY.**

1. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is by now a well-established method for biomacromolecular structure determination (*1, 2*). Further development of NMR structure determination is in part focused on increased efficiency of labor-intensive steps by computer-supported automation. One of these steps is the identification of the NMR signals in two- and higher-dimensional spectra, often referred to as "peak picking." In present practice this step in the evaluation of complex NMR spectra is usually done manually, with the aid of interactive computer programs (e.g., (*3, 4*)). The reason is that even advanced recognition methods, such as neural networks (*5, 6*), statistical approaches (*7, 8*), or numerical analysis of various properties of the data points (*9, 10*) often fail for complex spectra, mainly because of strong overlap of peaks and spectral distortions due to artifacts. The main weakness of most automated approaches is the fact that they analyze only the data points around a local maximum that is part of a potential peak. When interpreting spectra manually, an experienced spectroscopist will make use also of information outside of the data points near the local maximum. In this context it is important that multidimensional spectra typically contain multiple peaks that have the same lineshape and the same chemical shift in one frequency domain. This property has so far mainly been used for signal integration (e.g., (*11–13*)). The AUTOPSY method presented in this paper makes use of this observation for resolving strongly overlapping signals in a fully automated way.

Some methods for peak integration assume that the lineshapes can be expressed by an analytical function, typically a mixed Gauss/Lorentz function (*14, 15*). Peaks in real spectra often have lineshapes that are significantly different from Gauss/Lorentz functions, for example, when there is peak splitting due to scalar couplings. Assumptions that lineshapes follow analytical functions are therefore avoided for the most part of the method presented here. Instead, we use more generally valid criteria for evaluating potential peaks, in particular their symmetry and their regular shape. Symmetry considerations have previously been used for analyzing anti-phase peak patterns, e.g., in COSY spectra (*16–18*), but only rarely for the analysis of spectra with in-phase peaks, such as NOESY.

2. THE COMPUTATIONAL TOOLS OF AUTOPSY

2.1. General Strategy

The presently introduced approach for automated peak picking of complex multidimensional NMR spectra consists of the following steps, which are in the following sections described in more detail for the treatment of 2D data sets. The actual implementation of the algorithm in the program AUTOPSY is designed for spectra with an arbitrary number of dimensions.

- Determination of noise level. Exact determination of the noise level is important, so that weak peaks can also be recognized. The noise level is determined locally, so that the

[1] Present address: Tripos, Inc., 1699 S. Hanley Rd., St. Louis, MO 63144-2913.

[2] Present address: Biochemistry and Biophysics, Box 462, Göteborg University, SE 40530 Göteborg, Sweden.

[3] To whom correspondence should be addressed.

1090-7807/98 $25.00

algorithm can deal with noise bands, water lines, and similar artifacts.

- Segmentation. The spectrum is decomposed into connected regions made up of data points with signal intensities above the noise level. Data points outside of these regions are not considered for further analysis.
- Identification of separated peaks. Peaks that are well separated from others are identified first, using criteria such as symmetry and regular shape. Their lineshapes and chemical shift positions are extracted and used for the further steps.
- Comparison and grouping of lineshapes from all regions. If a combination of a given shift and lineshape was found in several peaks during the previous step, each occurrence will be included in a list of all lineshapes. In this way all lineshapes are compared for each frequency dimension and combined into groups of (approximately) equal lineshapes, which results in a reduced list of lineshapes that are characterized with higher precision.
- Resolving regions with strong overlap. Parts of regions with strong peak overlap were not treated in the previous steps. These are now resolved using the lineshapes collected from separated peaks. Shapes of potential peaks are constructed from different combinations of lineshapes in each dimension. These peak shapes are then used for explaining residual intensity that cannot be accounted for by the previously identified peaks, and for calculating the amplitudes.
- Integration. Calculating peak integrals is based on the lineshapes and amplitudes that were previously evaluated for all peaks.
- Symmetrization and filtering. For spectra that are expected to be symmetric with regard to their diagonal, an optional symmetrization step can be performed on the peak list. Before output, the peak list can also be filtered based on various other criteria, such as a peak quality factor or the linewidths.

2.2. Noise Level Calculation

A useful peak picking algorithm must be able to find peaks with intensities that are only slightly over the noise level without erroneously detecting a sizeable number of noise peaks. Noise in a NMR spectrum is generally not uniform, and may be larger close to the edges than in the central regions. Many spectra also have characteristic noise bands (t_1-noise, water line, diagonal). The following strategy for local noise level calculation was developed.

A noise level value is calculated for each 1D slice (rows and columns in 2D spectra) through the spectrum. For this purpose, a section of given length (typically 5% of the total length) is determined so that the standard deviation within this window is minimal. The noise level amplitude is then obtained by multiplying this standard deviation by an empirical factor, typically between 2 and 3, to make sure that only values significantly above the noise level are larger than this reference. Noise level values within the spectrum are then modeled as a base noise level present in the whole spectrum plus additional noise that can be present in each individual slice. With $\delta_{d,i}$ being the noise level of slice i in dimension d of an n-dimensional spectrum, the base level δ_b is obtained as the minimum of all these values,

$$\delta_b = \min_{d,i}(\delta_{d,i}). \qquad [1]$$

The additional noise levels for individual rows and columns, i, relative to the base level of the noise are calculated as

$$\delta'_{d,i} = \sqrt{\delta_{d,i}^2 - \delta_b^2}, \qquad d = 1, \ldots, n. \qquad [2]$$

The noise level at a given position P with coordinates (i_1, . . . , i_n) is then calculated from the base value and the additional values for the slices that pass through the data point,

$$\text{noise}(P) = \sqrt{\sum_{d=1}^{n} \delta'^2_{d,i_d} + \delta_b^2} = \sqrt{\sum_{d=1}^{n} \delta^2_{d,i_d} - (n-1)\cdot\delta_b^2}. \qquad [3]$$

More complex statistical methods have been proposed for noise level calculation (e.g., (*19*)). However, these cannot account for the often very characteristic, uneven noise distribution (bands) in NMR spectra and only calculate one global noise level value for the whole spectrum. They therefore do not seem suitable as robust methods for noise analysis of complete spectra.

2.3. Segmentation

Segmentation of spectra into connected regions of data points above the noise level is done with a "flood fill" algorithm (*20*), that was generalized to an arbitrary number of dimensions. Local maxima are used as seeds, where the filling algorithm is only started for maxima that are not within an already determined region. To make this test efficient even for large numbers of regions, all local maxima of previously determined regions are stored in a hash table. The implementation of the algorithm is such that it is never necessary to hold the whole spectrum in memory, only the currently processed parts of the rows are loaded. The data points within the bounding box of the region are kept in memory, in conjunction with an equally sized table of boolean values that indicate whether the corresponding data point lies within the region.

Many important spectra (e.g., TOCSY and NOESY) have a diagonal where the signals overlap so strongly that they cannot be evaluated. These diagonal signals can be excluded from the segmentation. For this the extent of the diagonal is determined by first using the aforementioned flood fill algorithm with points on the diagonal as seeds. If the noise level is used as the threshold for this filling step, strong cross peaks close to the diagonal are also excluded, even though they could subse-

quently be handled by the algorithm used. To avoid such loss of informative peaks, the threshold value for determining the diagonal can be gradually increased, and additional segmentation steps can be made.

Rows and columns close to the outer confines of a spectrum may suffer from poor processing (base line distortions). They can be excluded in the same way as described above for the diagonal, except that points on the outer boundaries of the spectrum are then used as seeds for the fill algorithm.

2.4. Identification of Separated Peaks

Peak Symmetry

Even though the lineshapes of the NMR signals are often close to Gauss or Lorentz functions, the error with which potential peaks can be fitted with such functions turned out to be an insensitive criterion for discriminating well-separated peaks from peaks that strongly overlap with other peaks. Instead, a more generally valid criterion of symmetry is used here. Assuming proper processing (phase correction) of the spectrum, the errors in symmetry in each frequency domain should be smaller than the noise level for peaks without overlap. To make use of this fact a measure for symmetry violation with regard to a given symmetry center is defined. We minimize this function with the position of the symmetry center as a parameter, and use the symmetry center as the position of the peak. The symmetry center is a good estimate for the position of the strongest peak even in the case of strong overlap. Furthermore, any symmetry violation with regard to this center is a valuable criterion to decide how well the peak is separated from other peaks.

The symmetry violation of a set of data points d_{ik} with regard to a given center, as expressed by residuals r_{ik}, is calculated by subtracting a symmetrized set of data points, d'_{ik},

$$r_{ik} = d_{ik} - d'_{ik}. \qquad [4]$$

The symmetrized values d'_{ik} are calculated as the minima of d_{ik} and the values at all symmetry-related positions relative to the given center $\vec{c}$ (the idea of this symmetrization is related to the algorithm of Baumann *et al.* (*21*), except that the procedure is applied to a limited number of data points around a potential peak, rather than to a complete spectrum). Because $\vec{c}$ is generally not exactly on a data point the symmetry-related positions will also be between data points, and spline interpolation is used for approximating the values at their positions. The residuals r_{ik} can be used for judging the symmetry. If the given point is the center of an exactly symmetric region, all r_{ik} are zero. For the search of centers $\vec{c}$ it might appear obvious to take a standard norm of r_{ik}, such as least squares, and to use this number as a symmetry violation. This would work well as long as the peaks are clearly separated or have very different intensities, but fails as soon as there is strong overlap of peaks with similar intensity. In this case the calculated center would often lie between the real peak centers, rather than at the position of one of the peaks. It turned out to be a much more stable and reliable approach to use a measure $\sigma(\vec{c})$ that favors the remaining intensity r_{ik} to be as smooth as possible:

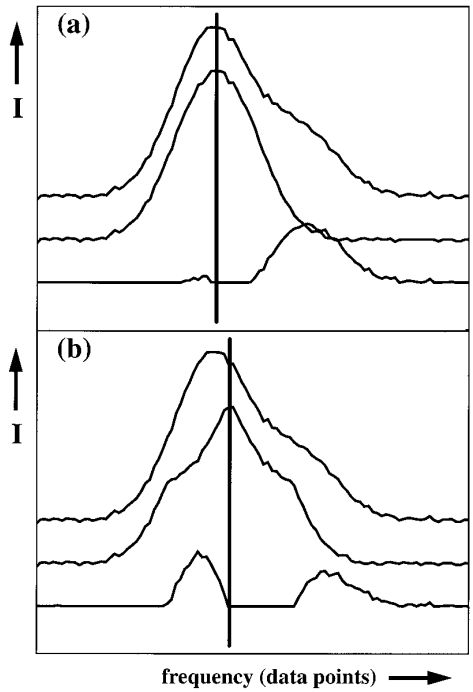

FIG. 1. Decomposition of a synthetic 1D data set with 100 data points. The data set consists of two peaks of amplitude 1.0 and 0.4, with a distance of 20 data points between the peak centers and linewidths of 20 data points each. Normally distributed random noise with a standard deviation of 0.02 was added. (a) Decomposition into two peaks with the use of the criterion described in the text (Eq. [5]). The top curve shows the synthetic data set. The vertical line shows the center determined for the main peak, the middle curve its shape after symmetrization, and the bottom line the difference between the top and middle lines. (b) Decomposition using the sum of differences (*24*) as the criterion for evaluation of symmetry.

$$\sigma(\vec{c}) = \sum_{i,k} |r_{ik} - r_{i-1,k}| + |r_{ik} - r_{i,k-1}|. \qquad [5]$$

Additional terms are used to avoid that the center "walks away" too far from the local maximum. These are based on the maximal expected splitting of peaks, which has to be specified by the user. Once the minimization of $\sigma(\vec{c})$ with $\vec{c}$ as a parameter has been completed, $\vec{c}$ is taken as the center of the potential peak and $\sigma(\vec{c})$ is used as a criterion for how well the potential peak is separated from other peaks.

A comparison with the standard approach based on minimizing the sum of squares of r_{ik} (*24*) illustrates the high reliability with which the position of the strongest signal in a cluster of overlapping lines can be identified using the criterion of Eq. [5] (Fig. 1). In the same situation, the sum-of-squares approach (*24*) yields a shifted position and a distorted line-

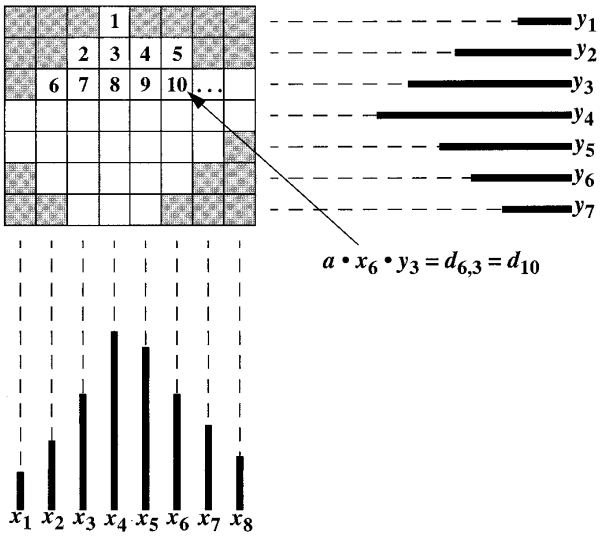

FIG. 2. Illustration of the calculation of unknown lineshapes x_j and y_k and the amplitude a from a given set of data points, $d_{jk} \equiv d_i$, using an overdetermined system of equations. Each data point in the peak area, numbered by the index i, leads to one equation. Data points outside the peak area are shaded.

shape for the strongest signal, and, consequently, fails to correctly locate the second, weaker signal in the cluster (Fig. 1b).

Uniformity of Peak Shape

Under most circumstances, the shape of a peak can be expressed as the product of a lineshape in each dimension, multiplied with an amplitude. In 2D spectra, with data points d_{jk} that have lineshapes x_j ($j = 1, \ldots, n$) and y_k ($k = 1, \ldots, m$), and an amplitude a, a peak shape can be factorized as

$$a \cdot x_j \cdot y_k = d_{jk} \equiv d_i. \qquad [6]$$

To determine lineshapes, Eq. [6] can be viewed as a system of equations for the $n + m + 1$ unknowns a, x_j, and y_k, with the number of equations being equal to the number of data points in the region (Fig. 2). To simplify the further notation, the spectral data points are treated as a one-dimensional vector $d_i = d_{jk}$ by using an index, i, that numbers the data points within the two-dimensional region (Fig. 2). These definitions can readily be extended to 3D and higher-dimensional spectra.

The system is nonlinear, but can be solved with a simple iterative method that assumes some starting values and then alternates between calculating new values for x_j while holding the values y_k fixed, and calculating new values for y_k while holding the values for x_j fixed. Methods that either just take single slices out of the data matrix (*13*) or sum up data points for obtaining the lineshapes (*22*) result in lineshapes that approximate the experimental data less precisely than the present algorithm.

With this procedure, lineshapes and amplitudes of potential peaks are calculated. The error (difference of the combination of lineshapes and the data points) expresses how uniform the shape of the peak is. This uniformity has proven to be a good criterion for discriminating between separated peaks and peaks with strong overlap. The lineshapes obtained from solving Eq. [6] will be used in the following steps. Note that no analytic lineshape such as Gauss/Lorentz is assumed.

Identification of Signals

Based on the criteria of symmetry and uniformity explained above, individually resolved signals can now be identified by performing calculations on the entire connected regions determined in the segmentation step. However, complex spectra have only few signals without overlap with other signals, and the individual connected regions that result from segmentation may contain many thousand data points and several hundred peaks. For this reason, symmetry and uniformity errors are at first only calculated for subregions around each local maximum that extend to all connected data points with values larger than half of the amplitude of the maximum. This is done for all local maxima, in order of decreasing amplitude. If the subregion around a maximum contains a previously processed (larger) maximum, the maximum is labelled as "not separated," and symmetry and uniformity are not calculated.

All maxima of a region that have been analyzed in this way are sorted according to the following three criteria: (i) Separated maxima precede not separated maxima. (ii) Of any two separated maxima, the one with the smaller relative error in symmetry and uniformity comes first. (iii) Of two maxima that are not separated, the one with the larger amplitude comes first. In the resulting list, maxima corresponding to well separated peaks are then at the top, and those corresponding to strongly overlapping peaks at the bottom, and they are further processed in this order. For each maximum, symmetry and uniformity of the surrounding sub-segment containing the data points with at least half the amplitude are again calculated, since these may have changed due to the subtraction of other signals (see below). The amplitude threshold for determining a sub-segment is then decreased step by step, until either the symmetry or the uniformity error increases significantly (typically by a factor of 1.2), additional maxima lie within the region, or the noise level is reached. If the peak amplitude is large enough, a new entry in the list of recognized peaks is generated. The final lineshapes are calculated from the data points symmetrized in the way described in the previous section, thereby reducing the influence of overlap. If the lineshapes do not extend down to the noise level because a higher threshold was used for the final determination of the sub-segment, they are extended by a fit with a mixed Gauss/Lorentz function. Using the lineshapes and amplitudes thus obtained, the peak is then subtracted from the data points and the lineshapes are entered into the list of all lineshapes.

Error estimates are made for all calculations and used for

various purposes, such as assigning a quality factor to each recognized peak and for increasing the threshold for the minimally tolerated amplitude. In this way it is possible to account for the error in the data that is produced by subtracting other peaks.

2.5. Grouping of Similar Lineshapes

Resolving overlap will require all lineshapes from the entire spectrum. In general, the identification of separated peaks as described in subsection 2.4 will produce each lineshape multiple times. This may cause problems for the next step of the procedure, and therefore it is necessary to produce a list that contains each lineshape only once. To this end the difference between two lineshapes is calculated as the root mean square difference (weighted according to the estimated error) between the values of the shape. An additional term for the difference between the center of the shapes is added. The shapes are then grouped with a clustering algorithm, which starts with each shape in a separate cluster. Then the two clusters with the smallest difference are repeatedly merged into one cluster until either a given threshold on the difference, or the desired number of clusters, which corresponds to the number of expected chemical shifts, is reached.

Lineshapes that are attributed to the same group by the clustering algorithm are combined into one lineshape, using a weighted average in which the shapes with smaller estimated errors obtain a higher weight. Combining several lineshapes into one also increases the precision of these shapes.

2.6. Resolving Spectral Overlap and Peak Integration

Potential Peaks

The results of the previous steps are lists of identified peaks and lineshapes for each frequency dimension. In all regions that do not consist exclusively of well-separated peaks, the overlapping peaks must now be resolved. For this purpose, a list of potential signals is created from the known lineshapes. For each frequency dimension the lineshapes within the range of the region are considered. Assuming that all lineshapes have been found, each peak must be a combination of one lineshape from each frequency dimension, and the set of all potential peaks can be constructed by taking all these combinations of lineshapes. This procedure creates all potential peaks in the bounding box of the region. Because regions are normally not rectangular, the potential peaks with centers outside the actual region are excluded. If a combination of lineshapes corresponds to a previously recognized peak, the potential peak is marked as "already found."

If x_j and y_k are two lineshapes in a 2D NMR spectrum, their outer product is the expected peak shape, s_i,

$$s_i \equiv s_{jk} = x_j \cdot y_k. \qquad [7]$$

The expected peak shapes are treated as one-dimensional vectors in the same way as the spectral data points in Eq. [6].

Subdivision

Because the aforementioned regions in complex spectra can be very large, large regions are first subdivided into more manageable parts. The algorithm described in the next paragraph achieves such a subdivision along a path through data points with the lowest possible intensity.

As a first step, a connected region around each local maximum is constructed by using a priority queue (*23*) that holds a number of data points ordered by their intensities and an index of a region that they are assigned to. Initially, the queue is filled with the local maxima. In addition to the queue, a list is generated that stores neighborhood relations between subregions, which is initially empty. The first entry in the priority queue is then repeatedly removed, marked with the index of the region, and all its unmarked neighbors are added to the queue. When a neighbor is encountered that was previously marked as belonging to another subregion, an entry with the two regions is made in the neighbor list. This procedure is continued until the queue is empty.

In a second step, the neighbor list is sequentially processed. The most strongly connected subregions will be at the front of the list, because in the previous step the data points were processed in the order of their intensities. Neighboring regions are then merged until the resulting region reaches the maximally desired size.

Approximation with a Selection of Peaks

The most obvious approach for resolving overlap would be to simply assume an unknown amplitude for each potential peak, and solving the overdetermined linear system of equations for least squares (Fig. 2). With ideal data, one would obtain zero for the absence of peaks, and amplitudes larger than zero for actual peaks. In reality, one obtains many small contributions instead of a few large ones, and even negative amplitudes are a quite common result. It is therefore preferable to first approximate the data with only a selection of potential peaks, and then add additional potential peaks when necessary, i.e., only where unexplained intensity remains. With a set of potential signals S, the index k going over all data points d_k, and s_{ik} being the shape of a potential signal i from S, Eq. [8] is the overdetermined system of equations for obtaining the unknown amplitudes a_i, which are imposed to be positive by the boundary conditions [9],

$$\sum_{i \in S} a_i \cdot s_{ik} = d_k \qquad \forall\, k \qquad [8]$$

$$a_i \geq 0 \qquad \forall\, (i \in S). \qquad [9]$$

The residuals obtained when solving this system are expected

to be localized where additional potential peaks must be selected. Unless all necessary potential peaks were selected, all intensity in the data cannot be explained. It is therefore not critical if the data are larger than the sum in Eq. [8], but it should never be significantly smaller. For these reasons, using an asymmetric norm instead of the normal least squares has proven to be highly valuable. The penalty of an error x is defined as

$$\varepsilon(x) = e^x - x - 1. \quad [10]$$

The error at data point k, taking account of the errors σ_{ik} of the peak shapes s_{ik} and the noise level δ, is defined in Eq. [11],

$$\Delta_k = \frac{\sum_{i \in S} (a_i \cdot s_{ik}) - d_k}{\sqrt{\delta^2 + \sum_{i \in S} (a_i \cdot \sigma_{ik})^2}}. \quad [11]$$

With an additional, strong penalty term for negative amplitudes (Eq. [12]; b is a weighting factor),

$$p_i = \begin{cases} b \cdot \left(\dfrac{a_i}{0.1 \cdot \delta} \right)^4 & \text{if } a_i < 0 \\ 0 & \text{otherwise} \end{cases} \quad [12]$$

the overdetermined system of Eq. [8] can be transferred to the problem of minimizing the function T in Eq. [13]:

$$T = \sum_k \varepsilon(\Delta_k) + \sum_{i \in S} p_i. \quad [13]$$

Gradients of this function can be calculated analytically, so that the optimization can be performed with the method of conjugate gradients (*24*).

The importance of using the asymmetric error function of Eq. [10] can be appreciated when comparing this approach with the work of Rischel *et al.* (*22*), who use a similar approach for peak integration and report that it is necessary to adapt amplitudes by visual inspection before peaks are subtracted.

Selection of Signals

The previous section explained how to calculate amplitudes for a given set of potential peaks. The key point is then to select the optimal set of peaks. A simple approach turned out to be most stable and reliable. For starting, one takes the set of already identified peaks. The amplitudes are calculated for this set, and the distribution of residuals (remaining intensity) is analyzed,

$$r_k = d_k - \sum_i a_i \cdot s_{ik}. \quad [14]$$

A match of this remaining intensity with each potential peak i that does not correspond to an already identified peak is then calculated as the cosine of the scalar product of the two vectors,

$$c_i = \frac{\sum_k r_k \cdot s_{ik}}{\sqrt{\sum_k r_k^2 \cdot \sum_k s_{ik}^2}}. \quad [15]$$

As long as there is a match above a user-given threshold (for example, 0.5), the potential peak with the best match is chosen, and added to the set of peaks. To make the procedure more reliable, an error estimate for c_i is also calculated, and taken into account by favoring potential peaks for which the error of the match calculation is small. Once there are no potential peaks with an acceptably good match left, all selected peaks that have a minimal amplitude (for example, 1.5 times the noise level) are added to the list of identified peaks. The match from Eq. [15] and the corresponding error are used for calculating the quality factor of the peak.

2.7. Symmetrization

Symmetrization of NMR spectra (*21*) is in current practice little used, since even spectra that are expected to have the same set of signals on both sides of the diagonal are often not fully symmetric, with significantly different intensities of peaks in symmetry-related positions of the frequency plane. Nonetheless, the fact that a peak was detected on both sides of the diagonal is a strong indication that it was correctly recognized. For this reason, an optional symmetrization step on the peak list can be performed in the AUTOPSY approach that may modify the qualities that were calculated in the recognition steps but does not directly remove any peaks. Since only peaks above a certain quality factor (for example, 0.5) are normally selected for further evaluation, such symmetrization may still result in the elimination of many artifactual peaks.

For each peak with quality factor q_1 (in the range 0 to 1) the peak is searched that is closest to its position mirrored at the diagonal. The quality factor of this peak is denoted as q_2; its distance from the mirrored position is d. If d is less than a given maximal distance $d_{\max}$, the new quality factor q_1' is calculated as

$$q_1' = 1 - (1 - q_1) \cdot \left(1 - \left(1 - \frac{d}{d_{\max}} \right) \cdot q_2 \right). \quad [16]$$

If no symmetric signal is found within the maximal distance, the intensity at the mirrored position is checked. It is well possible that no signal was found even though there is non-vanishing signal intensity, for example, because the position in question overlaps with the water line. The local noise level also needs to be considered. If s is the intensity at the symmetric

position, δ the corresponding noise level, and a the amplitude of the peak in question, the modified quality factor is

$$q_1' = \begin{cases} q_1 \cdot \dfrac{s + \delta}{a} & \text{if } s + \delta < a \\ q_1 & \text{otherwise.} \end{cases} \quad [17]$$

3. IMPLEMENTATION OF AUTOPSY

All functions needed for the AUTOPSY operations were implemented in a program-independent library. This library is structured in layers, where all functions of interest can be called up directly. It is possible to call low-level functions, such as symmetry calculation, as well as high-level functions, such as finding all separated peaks in a region. The library consists of around 6000 lines of ANSI C source code. All functions were implemented for spectra with an arbitrary number of dimensions; a maximal number of dimensions is provided at the compile time.

To keep the library general, input of spectra is handled over callback functions. Such functions were written for data that is already in memory, and for files in the BRUKER (See Bruker applications software) and XEASY (*3*) format.

Using the aforementioned library of functions, a complete peak picking program can be written with a few function calls, and the code can readily be incorporated into existing programs. A program with a comfortable user interface and flexible possibilities for various peak picking strategies was written where all the steps can be executed as single commands and therefore be combined freely. The program also has simple display possibilities for spectra and peak lists. Before writing out a peak list the user can make selections on the peak list that are based on criteria, such as peak position, linewidths, and quality factor. Most of the source code for this interactive program was taken from the molecular graphics program MOLMOL (*25*).

The program AUTOPSY is available from the authors. For details see http://www.mol.biol.ethz.ch/wuthrich/software/autopsy.

4. APPLICATION OF AUTOPSY WITH 2D NOESY SPECTRA

To evaluate the results that can be obtained using AUTOPSY, the program was applied to a 2D NOESY spectrum recorded for the structure determination of the killer toxin from the yeast *Williopsis mrakii* (*26*), a protein with 88 amino acids. Quantitative evaluation of the results of automated peak picking routines is difficult to obtain. Here we chose to use the calculated peak list as input for the automated NOE assignment procedure NOAH (*27*) implemented in the program package DYANA (*28*). The quality of the structure resulting from this procedure is compared with the quality of the structure obtained with manual peak picking and manual NOE assignment.

Input Spectrum and Peak Picking

The 2D NOESY spectrum in H_2O was recorded at 750 MHz, and processed to a size of 4096 data points in the direct ω_2 dimension and 2048 data points in the indirect ω_1 dimension. The same spectrum had been used for the manual interpretation by Antuch *et al.* (*26*).

Automated peak picking was done with the AUTOPSY procedure. The whole spectrum was used, i.e., no parts (such as the water line or the diagonal) were manually excluded. A minimal size of 6 data points in ω_2 and 3 data points in ω_1 was given for each peak. For the first step of identification of separated peaks a minimal amplitude of 2.0 times the local noise level was used. For the second step of identification of further peaks using lineshape decomposition, the minimal amplitude was 1.5 times the noise level. Symmetrization of the peak list was done with the procedure described in subsection 2.7. The peak picking calculation took less than 2 hours on a Silicon Graphics Indigo^2 with a MIPS R10000 processor (175 MHz). The program located 7871 possible peaks, of which the ones with calculated linewidths of less than 10 Hz, or a quality factor of less than 0.5, were excluded. The remaining 3789 peaks were used for the further analysis.

Figure 3 shows plots of a few representative spectral regions. It can be seen that the results are reliable for spectral regions with little (Fig. 3a) or moderately strong (Fig. 3b) overlap. Where very strong overlap occurs (Fig. 3c), the spectrum would be difficult to interpret even for an experienced spectroscopist, and the results of automated peak picking still look meaningful. Almost no peaks are identified on the very strong water line (Fig. 3d), and although the diagonal was successfully excluded, peaks close to the diagonal were still found (Fig. 3e).

Automated Assignment and Structure Calculation

The automatically determined peak list with 3789 entries, together with the chemical shift list obtained from the sequential assignment, was used as input for the automated assignment program routine NOAH (*27*) implemented in the program DYANA (*28*). After 25 assignment cycles, unique assignments were found for 2761 peaks. Using only these peaks, a final structure calculation resulting in a bundle of 20 conformers was then performed with DYANA, using torsion angle dynamics and the standard simulated annealing protocol. Figure 4 shows the resulting structure in comparison with the structure obtained from manual peak picking and manual assignment, where the RMSD between the well defined parts (residues 4–39 and 47–87) of the two mean structures is 0.92 Å. Both target function values and RMSDs are very similar for both structures (Table 1).

Compatibility with Known Structure

As an additional measure for the quality of the automatically determined peak list, its compatibility with the published struc-

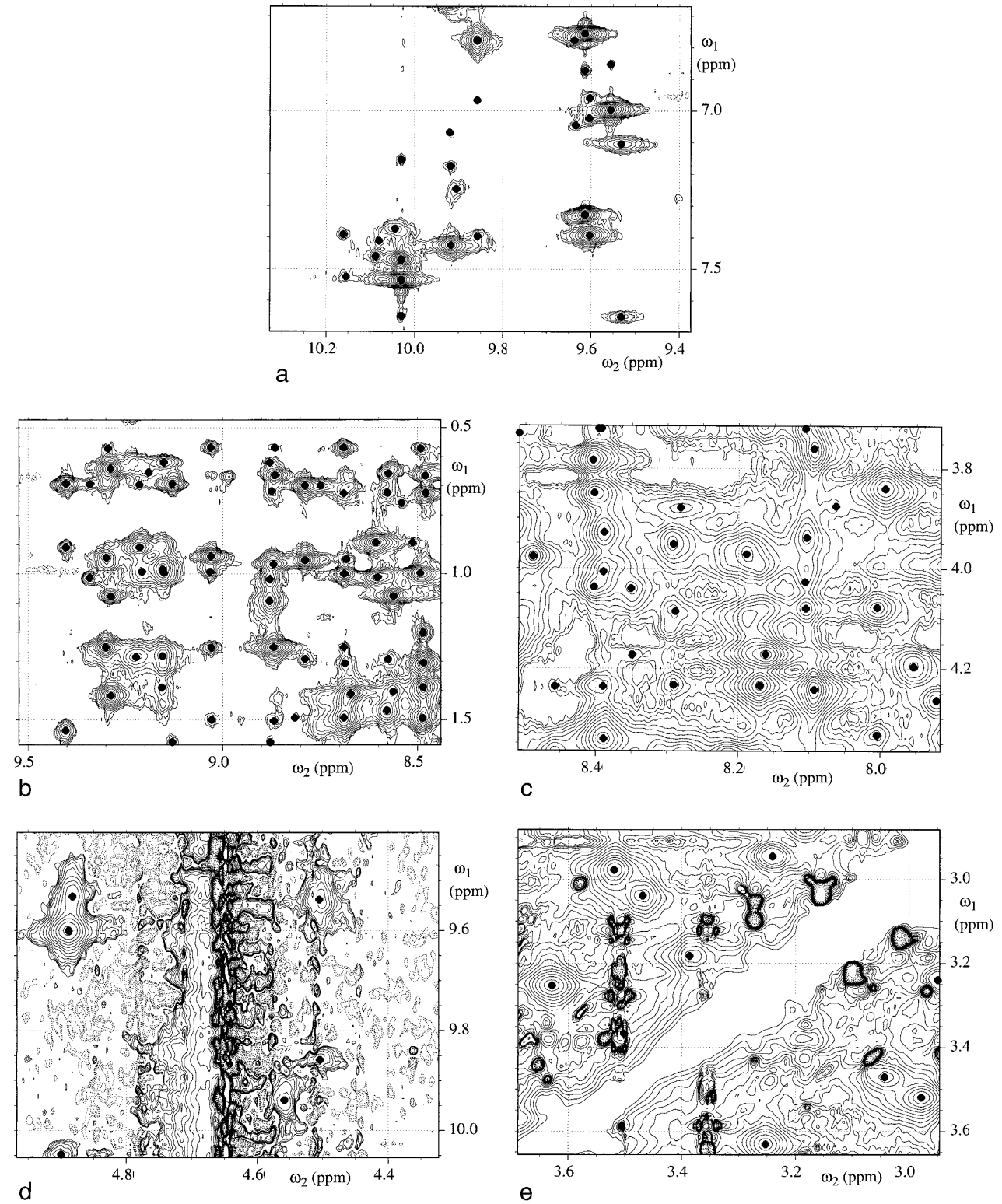

FIG. 3. Representative regions from the 2D NOESY spectrum of the protein WmKT. Automatically identified signal positions with quality factors > 0.5 and linewidths > 10 Hz are identified as dots. (a) Region with little peak overlap; (b) and (c) regions with increasing overlap; (d) region containing the water line; (e) region containing the diagonal.

ture (*26*) was calculated. Using a tolerance of ±0.01 ppm for the chemical shifts, there is a possible assignment to a proton pair with a distance of less than 6.0 Å for 3299 of the 3789 peaks. For 188 peaks there is no possible assignment, and the remaining additional 302 peaks have no possible assignments that would be compatible with the structure.

The 490 peaks that were thus found to be incompatible with the structure were manually analyzed and classified. A total of 218 of them were unambiguously confirmed as real peaks, 44 as peaks for which the chemical shift could not be precisely determined, 111 as questionable cases, and 117 as erroneous peak identifications.

Possible explanations for the fact that 218 unambiguously identified NOE cross peaks do not have a compatible assignment include that no sequential assignments were obtained for 10 protons in WmKT, effects of spin diffusion, and the presence of impurities in the sample.

Precision of Integrals

For all peaks that are present in both the automatically and manually determined peak lists, the integrals were quantitatively compared. Using $I^{1/6}$ (corresponding to the distance constraint used for the structure calculation) for an integral of size I, 62% of all integrals differ by less than 5% from the corresponding integrals in the manually determined peak list, 85% coincide within 10%, and almost 99% coincide within 25%. The few large differences occur mostly in cases where in the manual interpretation the intensity was distributed to several peaks, while only one peak was recognized by the automated procedure.

TABLE 1

Comparison of the Automatically Determined Structure of the Yeast Killer Toxin WmKT with the Structure Obtained from Manual Spectrum Interpretation

Quantity	Automatic[a]	Manual[b]
Number of assigned peaks	2761[c]	1698[d]
Number of upper distance limits	1237	1053
Range of final target function values[e]	1.7–3.2 Å^2	1.9–4.3 Å^2
RMSD (4–39, 47–87)[f]	0.57 Å	0.59 Å

[a] The 20 conformers calculated with DYANA (*28*) from data obtained by automatic assignment of the automatically determined peak list (see text).

[b] The 20 conformers calculated with DIANA (*31*) from data obtained by manual spectrum interpretation (*26*).

[c] Spectrum evaluated on both sides of the diagonal.

[d] Spectrum evaluated on one side of the diagonal.

[e] Range of residual violations in the last run of the individual structure calculations with DIANA or DYANA, respectively.

[f] Root mean square deviation of the backbone atoms of residues 4 to 39 and 47 to 87 relative to the average atom coordinates.

FIG. 4. Automatically determined structure of WmKT (dark) superimposed onto the structure obtained by manual interpretation of the spectrum (bright) (*26*). Shown are 10 conformers of each structure, the superposition is for best fit of the backbone atoms of residues 4–39 and 47–87. Image generated with MOLMOL (*25*).

5. DISCUSSION

The results in the previous section show that the AUTOPSY approach performs reliable automated peak picking for 2D NOESY spectra, and other complex 2D NMR spectra can be similarly analyzed. The outcome depends critically on the quality of the input spectrum. It is essential that careful data processing is performed, in particular baseline and phase correction, and that the resolution is high enough so that meaningful lineshapes can be extracted. The much smaller digital resolution of 3D spectra combined with reduced signal/noise ratio and increased incidence of artifacts, especially in ^{13}C-edited spectra, poses additional problems for successful application of AUTOPSY. While the results can probably be improved by processing spectra to larger sizes than normally used for manual interpretation, a more successful method for reliable automated peak picking for 3D NMR spectra may combine the AUTOPSY approach with analysis of additional input, such as chemical shifts taken from 2D spectra. With AUTOPSY, the NOAH routine (*27, 29*) implemented in DYANA (*28*) for combined automated NOESY cross peak assignment and three-dimensional structure calculation, and the program GARANT for automated sequence-specific assignments (*30*), a set of automated tools for all labor-intensive steps of NMR structure determination based on 2D spectra is now available. Future work will concentrate on combining these individual tools into a functioning and manageable software entity, and to implement additional routines for expanding the automated analysis to three- and possibly higher-dimensional NMR spectra.

ACKNOWLEDGMENTS

Financial support was obtained from BRUKER/Spectrospin AG, Fällanden, Switzerland. We thank Mrs. M. Geier for the careful processing of the manuscript.

REFERENCES

1. K. Wüthrich, "NMR of Proteins and Nucleic Acids," Wiley, New York (1986).
2. K. Wüthrich, "NMR in Structural Biology," World Scientific, Singapore (1995).
3. C. Bartels, T. Xia, M. Billeter, P. Güntert, and K. Wüthrich, The program XEASY for computer-supported NMR spectral analysis of biological macromolecules, *J. Biol. NMR* **6,** 1–10 (1995).
4. K. P. Neidig, M. Geyer, A. Görler, C. Antz, R. Saffrich, W. Beneicke, and H. R. Kalbitzer, AURELIA, a program for computer-aided analysis of multidimensional NMR spectra, *J. Biomol. NMR* **6,** 255–270 (1995).
5. S. A. Corne and P. Johnson, An artificial neural network for classifying cross peaks in two-dimensional NMR spectra, *J. Magn. Reson.* **100,** 256–266 (1992).
6. E. A. Carrara, F. Pagliari, and C. Nicolini, Neural networks for the peak-picking of nuclear magnetic resonance spectra, *Neural Networks* **6,** 1023–1032 (1993).
7. A. Rouh, A. Louis-Joseph, and J.-Y. Lallemand, Bayesian signal extraction from noisy FT NMR spectra, *J. Biomol. NMR* **4,** 505–518 (1994).
8. C. Antz, K.-P. Neidig, and H. R. Kalbitzer, A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a mulitivariate discriminant analysis, *J. Biomol. NMR* **5,** 287–296 (1995).
9. G. J. Kleywegt, R. Boelens, and R. Kaptein, A versatile approach toward the partially automatic recognition of cross peaks in 2D ^{1}H NMR spectra, *J. Magn. Reson.* **88,** 601–608 (1990).
10. D. S. Garret, R. Powers, A. M. Gronenborn, and G. M. Clore, A common sense approach to peak picking in two-, three-, and four-dimensional spectra using automatic computer analysis of contour diagrams, *J. Magn. Reson.* **95,** 214–220 (1991).
11. W. Denk, R. Baumann, and G. Wagner, Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines, *J. Magn. Reson.* **67,** 617–636 (1986).
12. T. A. Holak, J. N. Scarsdale, and J. H. Prestegard, A simple method for quantitative evaluation of cross-peak intensities in two-dimensional NOE spectra, *J. Magn. Reson.* **74,** 546–549 (1987).
13. H. Gesmar, P. Fæster Nielsen, and J. J. Led, Simple least-squares estimation of intensities of overlapping signals in 2D NMR spectra, *J. Magn. Reson. B* **103,** 10–18 (1994).
14. J. W. Brown and W. H. Huestis, Quantification of two-dimensional NOE spectra via a combined linear and nonlinear least-squares fit, *J. Biomol. NMR* **4,** 645–652 (1994).
15. K.-H. Sze, I. L. Barsukov, and C. K. Roberts, Quantitative evaluation of cross-peak volumes in multidimensional spectra by nonlinear-least-squares curve fitting, *J. Magn. Reson. A* **113,** 185–195 (1995).
16. B. U. Meier, G. Bodenhausen, and R. R. Ernst, Pattern recognition in two-dimensional NMR spectra, *J. Magn. Reson.* **60,** 161–163 (1984).
17. K. P. Neidig, R. Saffrich, M. Lorenz, and H. R. Kalbitzer, Cluster analysis and multiplet pattern recognition in two-dimensional NMR spectra, *J. Magn. Reson.* **89,** 543–552 (1990).
18. D. Jeannerat and G. Bodenhausen, Separation of overlapping multiplets and contraction of substructures within multiplets using symmetry properties, *J. Magn. Reson. A* **119,** 139–144 (1996).
19. P. Solé, F. Delaglio, and G. C. Levy, A segmentation technique for automated contour selection in 2D NMR spectroscopy, *J. Magn. Reson.* **80,** 517–519 (1988).
20. J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, "Computer Graphics: Principles and Practice," pp. 689–693, Addison–Wesley, Reading (1990).
21. R. Baumann, G. Wider, R. R. Ernst, and K. Wüthrich, Improvement of 2D NOE and 2D correlated spectra by symmetrization, *J. Magn. Reson.* **44,** 402–406 (1981).
22. C. Rischel, P. Osmark, and F. M. Poulsen, Resolving overlaps in two-dimensional NOE spectra by cross-peak subtraction, *J. Magn. Reson. B* **110,** 80–81 (1996).
23. R. Sedgewick, "Algorithms," pp. 380–386, Addison–Wesley, Reading (1988).
24. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C," Cambridge Univ. Press, Cambridge, United Kingdom (1988).
25. R. Koradi, M. Billeter, and K. Wüthrich, MOLMOL: A program for display and analysis of macromolecular structures, *J. Mol. Graphics* **14,** 51–55 (1996).
26. W. Antuch, P. Güntert, and K. Wüthrich, Ancestral $\beta\gamma$-crystallin percursor structure in a yeast killer toxin, *Nature Struct. Biol.* **3,** 662–665 (1996).
27. C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich, Automated combined assignment of NOESY spectra and three-dimensional protein structure determination, *J. Biomol. NMR* **10,** 351–362 (1997).
28. P. Güntert, C. Mumenthaler, and K. Wüthrich, Torsion angle dynamics for NMR structure calculation with the new program DYANA, *J. Mol. Biol.* **273,** 283–298 (1997).
29. C. Mumenthaler and W. Braun, Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry, *J. Mol. Biol.* **254,** 465–480 (1995).
30. C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich, GARANT—A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra, *J. Comp. Chem.* **18,** 139–149 (1997).
31. P. Güntert, W. Braun, and K. Wüthrich, Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA, *J. Mol. Biol.* **217,** 517–530 (1991).