JOURNAL OF MAGNETIC RESONANCE 95, 214–220 (1991)

A Common Sense Approach to Peak Picking in Two-, Three-, and Four-Dimensional Spectra Using Automatic Computer Analysis of Contour Diagrams

DANIEL S. GARRETT, ROBERT POWERS, ANGELA M. GRONENBORN, AND G. MARIUS CLORE

Laboratory of Chemical Physics, Building 2, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892

Received June 3, 1991

The analysis of multidimensional NMR spectra has been a challenging problem since the earliest two-dimensional experiments were reported as stacked plots (*1*). The first step in analysis involves obtaining a list of chemical-shift coordinates for the cross peaks. Initially, simple programs were used to generate contour plots of two-dimensional NMR data, which were analyzed manually. As the utility and versatility of multidimensional NMR spectroscopy grew, several automated methods of peak picking have been developed. This Communication describes a new peak-picking algorithm which is based on contour diagrams and designed for the automated interpretation of higher dimensional 3D and 4D spectra.

The oldest and most robust method of analysis is the manual interpretation of 2D contour plots. The strength of manual peak picking results from the relative ease with which the human eye can discriminate real peaks from artifacts and noise. As the proteins studied have become larger, the number of spectra to be analyzed and the number of cross peaks within a spectrum have increased dramatically, with the result that significantly more time and energy are required for the tedious manual peak-picking step. Interactive graphics software, which dynamically maintains a list of peak positions, has to some extent helped with this time-consuming step, particularly with regard to bookkeeping. Although more time will be saved with the automation of peak picking, manual inspection of spectra with an interactive graphics program will always be necessary to verify and edit automated results.

Approaches to automated peak picking can be divided into three types: (a) threshold-based methods; (b) multiplet-symmetry-based methods; and (c) peak-shape-based methods.

The simplest automated peak-picking algorithm is based primarily on the intensity of local extrema exceeding a threshold value (*2*). Uninteresting regions of the spectrum, such as t_1 noise ridges, are defined to avoid selecting peaks along these artifacts. Since some real peaks have very low intensities, the threshold must be set close to the noise level, which unfortunately results in a very large number of local extrema being picked due to the noise. Thus, by itself the threshold method fails by selecting too many or too few peaks, but is ideally suited as a filter for more sophisticated methods.

0022-2364/91 $3.00

The multiplet-symmetry-based method relies on the local extrema of real peaks exhibiting a known pattern. In ideal COSY-type spectra the multiplets show D_2, C_{2v}, and C_2 symmetries (*3–8*). Spectral overlap and coalescence of multiplet fine structure from broad linewidths, however, result in a lower symmetry for the multiplet (*9*). Further, as many important 2D experiments (*10*) and all the newly developed heteronuclear 3D and 4D experiments (*11–14*) do not contain any multiplet fine structure, the generality of symmetry-based methods is very restricted.

The peak-shape-based methods use the intensity distribution around local extrema to discriminate real peaks from noise or artifacts. This is a logical extension of the multiplet symmetry method (*3*), which hinted at using the C_{2v} symmetry of individual subpeaks in a multiplet to identify real peaks, but ignored this information. In this paper we describe a novel peak-shape algorithm called the *c*ontour *a*pproach to *p*eak *p*icking (CAPP) and discuss it in relation to another recently described method termed the *s*ynergetic approach *t*oward the *e*valuation of *l*ocal maxima in *l*ow-symmetry spectr*a* (STELLA) (*15*).

The STELLA algorithm requires that several local extrema be identified as real peaks or artifacts by interactively selecting individual rectangular regions around the peaks. The intensity distributions within each region are separately stored in a matrix for computing match values to the intensity distribution around unknown local extrema. An unknown local extremum is labeled as a real peak or artifact or remains unknown, depending on the distribution of match values to the predefined real peaks and artifacts. Basically, the spectroscopist is teaching the STELLA program the shape of real peaks and artifacts by example. The success in discriminating real peaks from artifacts relies on the fact that real peaks exhibit common features which differ from those of artifacts.

Another, probably superior, approach is one where the spectroscopist explicitly defines the peak shape of interest, eliminating the need for selecting a few dozen example peak shapes. This is the avenue taken by the CAPP algorithm presented in this paper, which models a contour diagram as a set of ellipses. Ideally, the contours resulting from an isolated peak of a properly phased spectrum are a set of concentric ellipses. Even though digital resolution, spectral overlap, and phase distortion perturb the contours from the ideal case, the approximation of each contour as a single ellipse provides a mechanism for defining the desired shape of real peaks. The CAPP method is analogous to manual peak picking in many ways since both are based on the same contour diagram and can be described pictorially. Further, CAPP and manual peak picking analyze 2D, 3D, and 4D NMR spectra in the same way, the latter two as a series of 2D slices.

The four steps involved in the CAPP algorithm are as follows: (a) generation of the contour diagram; (b) calculation of the ellipses that best fit the contours; (c) location of potential noise ridges along each axis; and (d) definition and location of real peaks from the ellipses.

In the first step of CAPP, the contours are calculated on a logarithmic intensity scale for each 2D slice using a minimum threshold level and level multiplier. The position of each contour point is linearly interpolated between adjacent NMR data points which bracket the current contour level. Each contour is maintained separately as a complete unbroken path to facilitate calculating the ellipses. The plane containing

the contour points has axes labeled X and Y instead of F_1 and F_2 to avoid nomenclature problems between 2D, 3D, and 4D NMR spectra. The third and fourth dimensions when present are labeled the Z and A axes, respectively, and are analyzed as separate 2D slices.

In the second step of CAPP, the set of ellipses which best fit the contours is calculated. A single ellipse is calculated for each contour in two steps: an approximate center and radii along X and Y are first calculated, and the ellipse parameters are then refined using the simplex method (*16*). The ellipse center (X_0, Y_0) is approximated as the average of the contour points on a single contour. The X and Y radii (r_X, r_Y) are approximated by the average distance of the extreme X and Y contour points from (X_0, Y_0). The approximate ellipse parameters are then optimized by simplex minimization (*16*) of the RMS deviation between each contour point and the closest point on the ellipse. The ellipse point, (X_{ell}, Y_{ell}), which is closest to the contour point (X_1, Y_1), is calculated in two steps (Fig. 1). First, an approximate ellipse point is calculated by the intersection of the ray in Eq. [1] and the ellipse in Eq. [2] as shown in Fig. 1A:

$$X = X_0 + L(X_1 - X_0),$$

$$Y = Y_0 + L(Y_1 - Y_0), \qquad L \geq 0, \qquad [1]$$

$$\frac{(X - X_0)^2}{r_X^2} + \frac{(Y - Y_0)^2}{r_Y^2} = 1, \qquad [2]$$

where L is an arbitrary parameter in the ray definition. Second, the approximate ellipse point (X_{ell}, Y_{ell}) is optimized by minimizing the distance between (X_1, Y_1) and the normal (Fig. 1B) to the ellipse at (X_{ell}, Y_{ell}) using a combination of Newton–Raphson and bracketed bisection techniques (*16*).

In the third step of CAPP, the ellipses are searched to locate potential ridges along both axes. The centers of concentric ellipses are averaged to define peaks for locating ridges. Initially, ridges which run along the X axis are sought by sorting the peaks by their Y chemical shift. The set of peaks whose Y chemical shifts are within a given range of each other are used to define a ridge along X if one of two conditions are met: (a) the sum of the radii along X exceeds a minimum length (usually 20% of sweep width along X), or (b) the set contains a minimum number of peaks (usually 15) with X chemical shifts extending over a minimum length (usually 20% of sweep width along X). The center, height in number of contour levels, and width of the ridge are calculated to define each ridge that was located. In a similar fashion, ridges along the Y axis are located and defined. This method of locating ridges is particularly useful in 3D and 4D experiments, where some weak ridges are not observed in every slice.

In the fourth step of CAPP, the ellipses which model desired peak shapes are required to meet four conditions: (a) the RMS deviation between contour points and ellipse must be less than a predetermined cutoff value; (b) the radius of the ellipse along each axis must lie within defined limits; (c) the ratio of ellipse radii must be between defined limits; and finally (d) the percentage circumference deviation between the ellipse and the contour must be less than a cutoff value. Any ellipse which does not meet all of these conditions is not considered in defining real peaks. The centers of the remaining

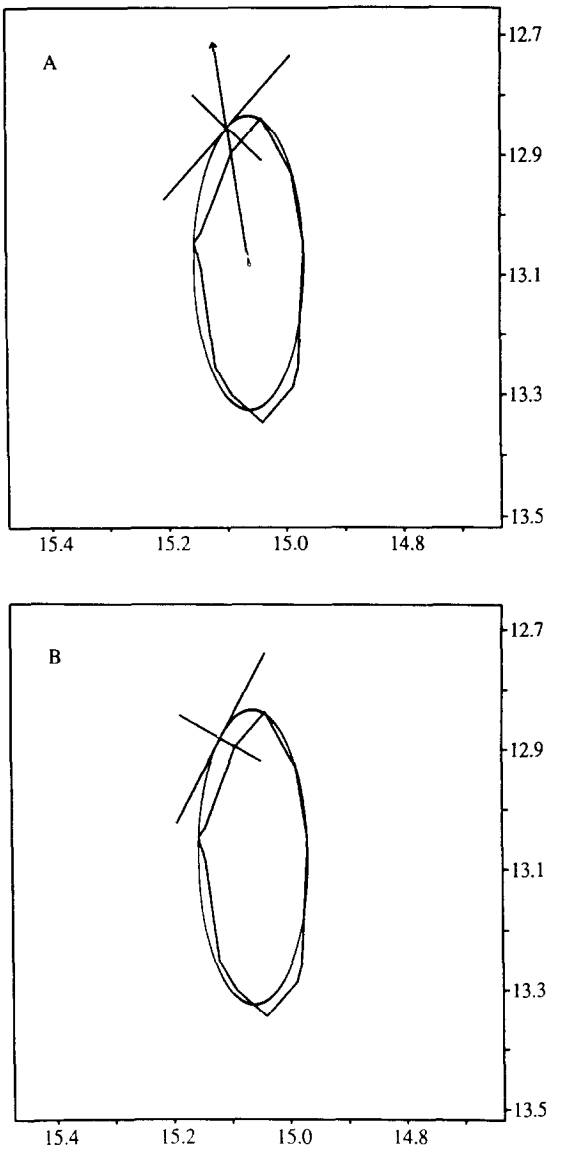

FIG. 1. Two-step calculation of the closest ellipse point. (A) A point on the cross-peak contour is first approximated by the intersection of the ray (indicated by the arrow) with an ellipse and then (B) is optimized by Newton–Raphson and bracketed bisection techniques until the distance from the contour point to the normal at the ellipse is under the desired error limit (see text for further details).

concentric ellipses are averaged to determine peak centers. A peak is kept as a real peak provided that: (a) the peak center is not on a ridge, or the highest contour level defining the peak is greater than the ridge; (b) the peak is defined by a minimum number of ellipses (usually 2); and (c) in the case of 3D and 4D spectra, the intensity

of the peak on the slice being analyzed is greater than the intensity on the previous and next 3D and/or 4D slice.

The parameters which control each step in CAPP are optimized from default values. One or two typical slices from a 3D experiment are used for optimization. Since run times are approximately five to ten seconds per slice on a Sun Sparc workstation, parameter optimization is completed in under thirty minutes. Typically, a 64-slice 3D experiment is automatically peak picked in five minutes by CAPP. To verify and edit the CAPP peak-pick table, we have also written an interactive graphics-based program, known as the *p*rimitive *i*nteractive *p*eak *p*icker (PIPP). This program, like CAPP, is written in C and makes use of the X11 graphics library and the Open-Look toolkit on a SPARC 1+ workstation.

The purpose of CAPP is to save spectroscopist time by automatically discriminating real peaks from artifacts and noise. Since the spectroscopist could have exclusively used PIPP for manual peak picking, CAPP will save time only if the number of peaks picked that are correct exceeds the number that are incorrect. The efficiency of CAPP can be assessed by the difference between the number of correctly and incorrectly picked peaks divided by the total number of real peaks. An efficiency of 100% indicates perfect peak picking, whereas a negative efficiency indicates that the spectroscopist would have been better off using PIPP alone without CAPP.

CAPP has been tested on 2D, 3D, and 4D spectra. The results of running CAPP on five 3D triple-resonance spectra are summarized in Table 1. The local maxima quoted in Table 1 show the number of picks that would be obtained with a threshold-based method. The other entries were obtained via inspection and editing of CAPP results with PIPP. The high positive efficiencies in Table 1 demonstrate that CAPP is very successful in discriminating real peaks from artifacts and noise.

Nevertheless, the presence of both false positive and negative errors indicates some systematic problems inherent in the use of CAPP, which can be attributed to four different sources. First, weak ridges whose length or number of peaks does not exceed the user-defined value are not located and may give rise to false positive picks. Since ridge contours tend to be narrower than real peaks, they are preferentially eliminated on the basis of the shape of their modeled ellipses. In addition, cross peaks involving side-chain NH_2 groups in the 3D HNCA, HNCO, and HN(CO)CA spectra, which are not useful in assigning the backbone resonances, are also preferentially eliminated. Unfortunately, some real peaks may have shapes that are similar to unlocated ridge or NH_2 shapes. The second kind of error results from the compromise between minimizing ridge and NH_2 picks while maximizing correct picks, leading to false positive and false negative errors, respectively. Third, real peaks that are on a ridge of comparable or greater intensity are not picked and give rise to false negative errors. These peaks are found manually in PIPP by observing that their widths are greater than that of the ridge. Fourth, real peaks which are not resolved in the contour diagram are also not picked and give rise to false negative errors.

The fourth systematic error involving unresolved peaks is a greater problem in crowded 2D spectra than in 3D or 4D spectra. The major source of this particular error lies in the modeling of contours which enclose multiple local maxima, as a single ellipse. Two solutions are currently being explored to overcome this limitation: in particular, multiple ellipse and multiple 2D Gaussian models.

TABLE 1

Summary of CAPP Results

3D experiment type[a]	Local maxima	False positive	False negative	Correct picks	Efficiency[d] (%)
HNCA[b]	5,040	10	24	159	81
HNCO[b]	2,667	10	10	110	83
HN(CO)CA[b]	2,482	6	17	89	78
HCACO[c]	7,166	3	13	151	90
HCA(CO)N[c]	4,186	4	17	67	75
Total	21,541	33	81	576	82

[a] All 3D experiments were recorded on a Bruker AM-600 at 26°C on a 1.1 m*M* sample of uniformly (>95%) $^{15}N/^{13}C$-labeled RNase H domain of HIV-1 reverse transcriptase in 100 m*M* sodium phosphate, pH 5.4, dissolved in 90% H_2O/10% D_2O for the HNCO (*17*), HNCA (*17*), and HN(CO)CA (*18*) experiments and in 99.996% D_2O for the constant-time HCACO (*19*) and HCA(CO)N (*19*) experiments. This domain has 138 residues, a molecular mass of 15 kDa, and a Trp → Ala substitution at position 113 introduced by site-directed mutagenesis (*20, 21*). The data were analyzed as a series of F_2–F_3(^{1}H) planes with the ^{15}N dimension in F_1 for the HNCA, HNCO, and HN(CO)CA experiments and the $^{13}C^{\alpha}$ dimension in F_1 for the HCACO and HCA(CO)N experiments. The spectral widths in the ^{15}N, $^{13}C^{\alpha}$, and ^{13}CO dimensions were 29.16, 33.13, and 12.05 ppm, respectively, with the carrier positions placed at 118.5, 56, and 177 ppm, respectively. The spectral width in the ^{1}H dimension was 13.44 ppm with the carrier at 4.76 ppm for the HNCO, HNCA, and HN(CO)CA experiments. For the HCACO and HCA(CO)N experiments, the ^{1}H spectral width was 8.33 ppm with the carrier at 4.76 ppm. For the HNCO, HNCA, and HCA(CO)N experiments, the number of points acquired in the various dimensions was 32 complex in F_1 (^{15}N), 64 complex in F_2 (^{13}CO or $^{13}C^{\alpha}$), and 1024 real in F_3 (^{1}H). For the HCACO and HCA(CO)N experiments, the number of points acquired was 32 complex in F_1 ($^{13}C^{\alpha}$) and 64 complex in F_2 (^{13}CO) for the HCACO experiment and 32 complex in F_2 (^{15}N) for the HCA(CO)N experiment, and 512 real in F_3. Zero filling was used in all dimensions, and for the HCACO and HCA(CO)N experiments, linear prediction by means of the mirror-image technique (*22*) was used to extend the data further. The final 3D spectra consisted of 64 × 128 × 1024 data points for the HNCO, HNCA, and HN(CO)CA experiments and 128 × 128 × 512 points for the HCACO and HCA(CO)N experiments. All the spectra were processed on a Sun Sparc workstation using in-house routines for Fourier transformation (*23*) and linear prediction (*22*), together with the commercially available software package NMR2 (New Methods Research, Inc., Syracuse, New York).

[b] Sixty-four F_2–F_3 slices after zero filling.

[c] One hundred twenty-eight F_2–F_3 slices after zero filling and linear prediction.

[d] Efficiency calculated (correct picks − false positives)/(correct picks + false negatives).

In summary, an efficient and robust automatic peak-picking program based on the contour diagram has been presented which allows rapid tabulation of peaks in 2D, 3D, and 4D spectra. The combination of CAPP and PIPP allows an entire 3D spectra to be easily peak picked in under three hours.

ACKNOWLEDGMENT

This work was supported by the AIDS Targeted Anti-Viral Program of the Office of the Director of the National Institutes of Health (G.M.C. and A.M.G.). This program is available on tape upon request from the authors.

REFERENCES

1. J. JEENER, B. H. MEIER, P. BACHMANN, AND R. R. ERNST, *J. Chem. Phys.* **71,** 4546 (1979).
2. C. CIESLAR, G. M. CLORE, AND A. M. GRONENBORN, *J. Magn. Reson.* **80,** 119 (1988).

3. B. U. MEIER, G. BODENHAUSEN, AND R. R. ERNST, *J. Magn. Reson.* **60,** 161 (1984).
4. P. PFÄNDLER, G. BODENHAUSEN, B. U. MEIER, AND R. R. ERNST, *Anal. Chem.* **57,** 3510 (1985).
5. P. PFÄNDLER AND G. BODENHAUSEN, *J. Magn. Reson.* **70,** 71 (1986).
6. K. P. NEIDIG, H. BODENMUELLER, AND H. R. KALBITZER, *Biochem. Biophys. Res. Commun.* **125,** 1143 (1984).
7. S. GLASER AND H. R. KALBITZER, *J. Magn. Reson.* **74,** 450 (1987).
8. K. P. NEIDIG, R. SAFFRICH, M. LORENZ, AND H. R. KALBITZER, *J. Magn. Reson.* **89,** 543 (1990).
9. D. NEUHAUS, G. WAGNER, M. VASAK, J. H. R. KAGI, AND K. WÜTHRICH, *Eur. J. Biochem.* **151,** 257 (1985).
10. R. R. ERNST, G. BODENHAUSEN, AND A. WOKAUN, "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," Clarendon Press, Oxford, 1987.
11. S. W. FESIK AND E. R. P. ZUIDERWEG, *Q. Rev. Biophys.* **23,** 97 (1990).
12. G. M. CLORE AND A. M. GRONENBORN, *Ann. Rev. Biophys. Biophys. Chem.* **20,** 29–63 (1991).
13. G. M. CLORE AND A. M. GRONENBORN, *Progr. NMR Spectrosc.* **23,** 43 (1991).
14. G. M. CLORE AND A. M. GRONENBORN, *Science* **252,** 1390 (1991).
15. G. J. KLEYWEGT, R. BOELENS, AND R. KAPTEIN, *J. Magn. Reson.* **88,** 601 (1990).
16. W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING, "Numerical Recipes in C: The Art of Scientific Computing," Cambridge Univ. Press, Cambridge, 1988.
17. L. E. KAY, M. IKURA, R. TSCHUDIN, AND A. BAX, *J. Magn. Reson.* **89,** 496 (1990).
18. M. IKURA AND A. BAX, *J. Biomolec. NMR* **1,** 99 (1991).
19. R. POWERS, A. M. GRONENBORN, G. M. CLORE, AND A. BAX, *J. Magn. Reson.* **94,** 209 (1991).
20. S. P. BECERRA, G. M. CLORE, A. M. GRONENBORN, A. R. KARLSTRÖM, S. J. STAHL, S. H. WILSON, AND P. T. WINGFIELD, *FEBS Lett.* **270,** 76 (1990).
21. R. POWERS, G. M. CLORE, A. BAX, D. S. GARRETT, S. J. STAHL, P. T. WINGFIELD, AND A. M. GRONENBORN, *J. Mol. Biol,* in press.
22. G. ZHU AND A. BAX, *J. Magn. Reson.* **90,** 405 (1990).
23. L. E. KAY, D. MARION, AND A. BAX, *J. Magn. Reson.* **84,** 72 (1989).