A Novel Contour Plot Algorithm
for the Processing of 2D and 3D NMR Spectra

J. Weber[1], F. Herrmann[2], P. Rösch[2] and A. Wokaun[1]

[1]*Physikalische Chemie II and* [2]*Lehrstuhl für Biopolymere*
University of Bayreuth, D - W - 8580 Bayreuth, Germany

Contents

I.   Introduction

Elucidation of the conformation of peptides and proteins in solution by NMR (nuclear magnetic resonance) methods requires efficient and versatile data processing capabilities. Especially for the handling and for a convenient display of large two- or three-dimensional data matrices, sophisticated codes are being developed.

A conventional 2D NMR spectrum typically consists of $512 * 4096$ 4 byte floating point data values, equivalent to 8 MByte. An impressive variety of 3D NMR experiments has been conceived and realized to this date (1-7). The introduction of the third dimension considerably increases the data size. At present, data matrices consisting of $256 * 256 * 512$ or of $128 * 128 * 2048$ floating point data values, i.e. 134 MByte, can be handled with medium-sized work stations (8).

Provided that highly resolved spectra of the molecule can be recorded, the next decisive step in structural investigation is the assignment of the spectra. In this process, a major task consists in identification of NOE connectivities. Short range connectivity information is required for the sequential assignment, while the long-range NOE contacts serve as the crucial input for the distance geometry and restrained molecular dynamics calculations (9-14).

Several efforts have been reported (15-19) to automate the assignment of 2D spectra. However, neither for 2D nor for 3D spectra of proteins, a decisive breakthrough in this computational problem has been made yet. As a prerequisite for any assignment algorithm, be it by eye or by computer, several requirements must be met, i.e. the capability of handling large data matrices, the identification of cross peaks, and the extraction of their specific coordinates. As a tool for the solution of this problem, a new contour plot algorithm, called the 'ribbon' method, is reported in this communication, which provides several advantages as compared to conventional grid search methods.

II.   The 'ribbon' contour plot algorithm

The algorithm extracts all pixels belonging to a contour, and stores the coordinates of the contour

in a sequence proceeding around the circumference of the peak. Furthermore, the peak volume and the center of mass coordinates are computed.

1. Peak identification

At the onset, the process of drawing a contour in the two-dimensional data matrix representing a 2D spectrum is reviewed. Once the intensity level has been set, a peak is defined as a set of data points fulfilling the condition that all points with intensities higher than the predetermined level are immediately adjacent along rows and/or columns of the 2D matrix. First, the original data matrix is searched for 'transitions' across the preset contour level. To avoid border problems, the data matrix is extended by one row and one column of zeroes on either side of the data. Starting from coordinates (0, 0), the 2D spectrum is searched for up-transitions (preceding point has a lower intensity than the contour, while intensity of the next point is higher than the contour level) and for down-transitions, both in horizontal and vertical direction. The two types of transition are stored as distinct flags.

All points belonging to one peak are extracted from the 'transition matrix.' This might be envisaged as a 'flood fill' were the peaks are protruding from an ocean of constant height, corresponding to the contour level. In a search along the rows, all points between the first up-transition and the corresponding down-transition are marked as 'horizontally connected' and stored as 'to be searched vertically.' Starting from these points, a vertical search along the columns is performed, and 'vertically connected' points are marked and stored as 'to be searched horizontally.' This procedure is repeated until no more points remain 'to be searched.' The connected points identified in this manner are written into a peak matrix.

2. The contour

The coordinates of the peak contour are collected in an ordered array (i.e., sequentially along the contour line starting at any point) by means of a virtual 'ribbon' that surrounds the peak in the corresponding matrix. First, the band is placed at the smallest rectangle surrounding the peak. The idea, illustrated in Figure 1, is to shrink the ribbon until it touches the contours of the peak everywhere. For

this purpose one has to define, for every segment of the elastic ribbon, the direction in which it is going to contract. With all other segments fixed, one point of the ribbon is moved inwards until a transition is found. If more than one step is required, new points are inserted into the ribbon from which a search perpendicular to the original one has to be performed subsequently.

Once the searching tip has arrived at a transition, it is anchored there, and develops sprouts in the two perpendicular directions. At this point it is important to test immediately whether it is possible to go 'around the corner' from the present contour point (Figure 2). Staying with the image of a flood fill, this test ensures that one reaches the inside of any lagoons or coves around the borders of the peak.

If, in the course of the search, any two points of the ribbon meet with opposite search directions, a connection is made, the points in between are deleted, and the rubber band shrinks accordingly. When all points with their specific search directions have been anchored in this manner, the ribbon adheres correctly around the contour of the peak, and *the points in the contour are connected sequentially,* which is of advantage for the graphical output.

One has to consider the case illustrated in Figure 3 that there might be a 'lake' hidden within the peak. If transitions are indeed found within the peak, the outer contour is stored and then removed, and the sign of the inner transitions is inverted (down becomes up and vice versa). Thereafter, a restart of the ribbon search method will correctly yield the inner contour.

Evidently one might think of a peak surrounded by another peak (an island within the lake in the picture of the flood fill, cf. Figure 3). This case is treated by storing and deleting the contour of the 'lake,' such that every interior peak is consequently found. Of course, the entire procedure must now be repeated over the entire integer matrix of 'transitions,' until all contours have been drawn.

Thus far the points of the contours are associated with the indices of the data points. Finally the accurate coordinates of every point of the contour are calculated by interpolation.
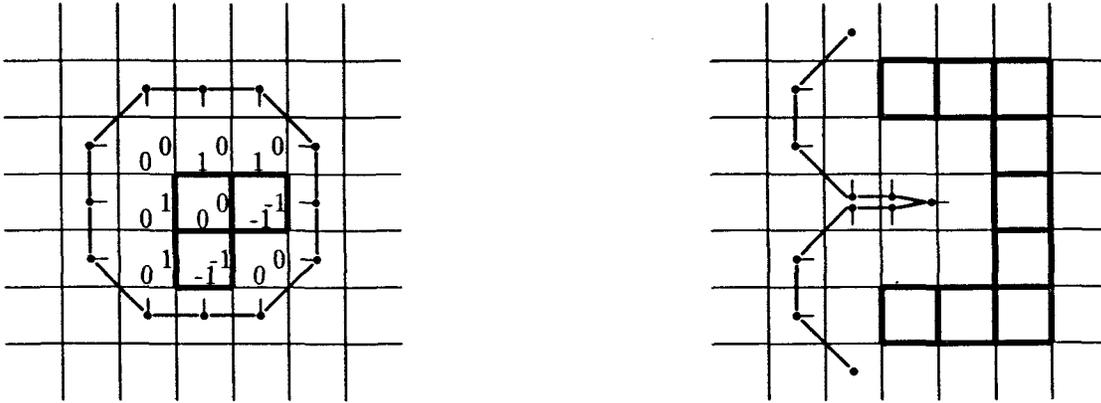
Figure 1: Principle of the 'ribbon' method used to define the contour of a peak. In (a), the fully expanded ribbon surrounds the 'peak' as represented by the integer 'transition matrix.' The search directions of the ribbon elements are indicated by horizontal or vertical dashes. After a two-step search (b), four new ribbon elements with perpendicular search directions have been inserted.
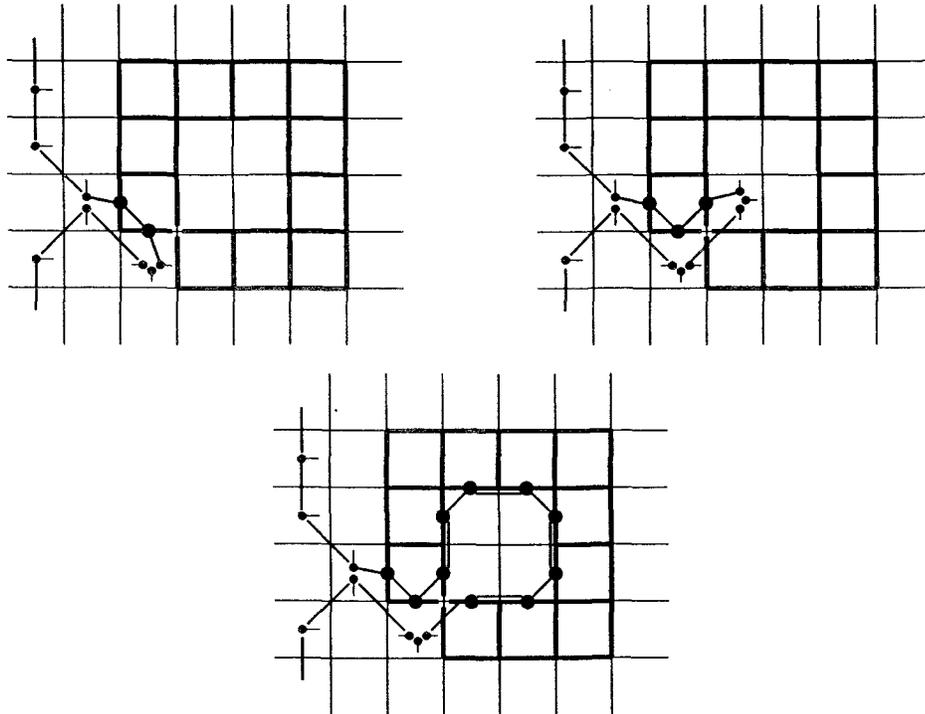

Figure 2: 'Corner test' used to reach the inside of a 'cove.' A triple of new ribbon elements is inserted after every successful step around a corner (a). The triple inside the cove shown in (b) serves to attach the ribbon to the inner wall (c).
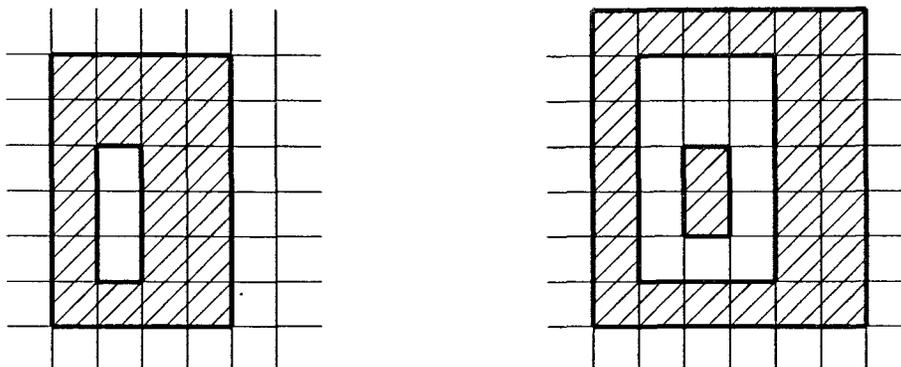
Figure 3: Complex peak structures. In the model of a flood fill, a peak may contain a 'lake' (a), or even a second peak ('island') within the lake (b).

III.   Discussion

A conventional grid search results in pairs of points to be connected, i.e. the individual elements of a line. In contrast, the present contour plot algorithm yields all the points contributing to the contour of a given peak in sequential order. These contours are easily stored, and can be conveniently displayed with graphics standards such as X-Windows (20), GL (21) or PEX (22).

As a consequence of this graphic advantage, it is possible to overlay the results of several different experiments on the screen or display device. Of particular interest is the option for a visual or graphical comparison of the spectra resulting from different types of experiments (i.e., pulse sequences). Such spectra will, in general, have been acquired using different spectrometer frequencies, carrier offsets, and sweep widths. Prior to a direct overlay, the various spectra must therefore be scaled individually; this option has been implemented in our program.

As an example a comparison of TOCSY and NOESY type experiments is shown in Figure 4.

With the present algorithm, an exact calculation of the peak volume (integral) is straightforward, as all points inside the contour are clearly identified. At the same time, the center of mass coordinates of the peak may be obtained. By comparison, several programs in current use (23 - 26) calculate the integral over the smallest rectangle surrounding the peak; the coordinates of the peak, set equal to the

center of the rectangle, consequently do not precisely correspond to the center of mass. The accuracy gained with the present algorithm is a promising starting point for automated data evaluation. The amount of data is considerably reduced to a list of coordinates and integrals of the peaks. Again, complementary information from various types of experiments may be used as an input. This possibility is being explored in ongoing work in our laboratories.

IV.   Environment for the reduction and representation of data from 2D and 3D NMR experiments - the 'NDEE' program package

The contour plot algorithm described is part of a program system that manages all operations which need to be performed during the reduction of a two- or three-dimensional NMR spectrum, i.e. baseline subtraction, fast Fourier transformation, 2D or 3D phase correction, and graphical display. These operations have been implemented in a program package termed 'NDEE.' The code was developed to meet the requirements of NMR research groups, i.e. processing of 2D and 3D data files, fast performance, portability, and ease of handling via a self-explanatory user interface. Particular attention was paid to the problems and needs met in the structural determination of biopolymers.
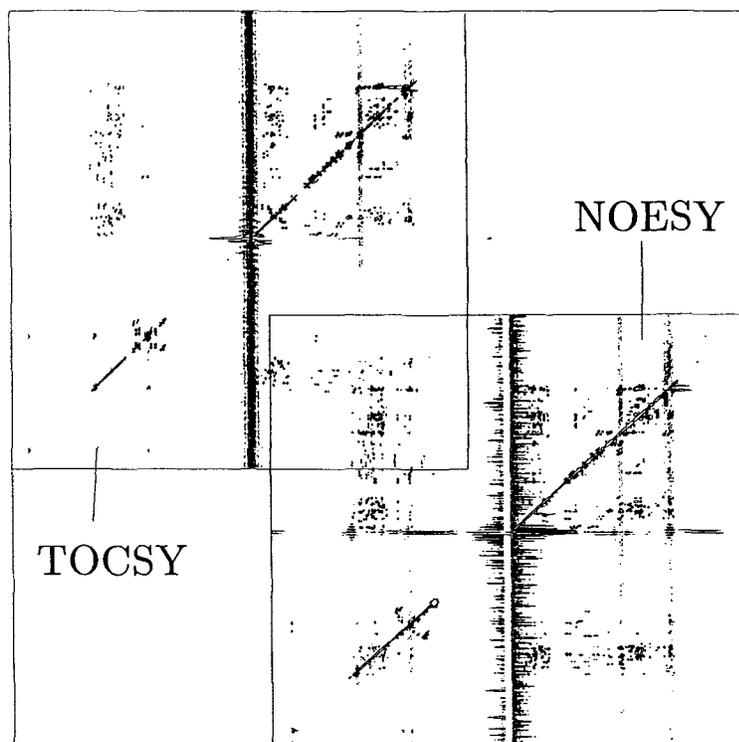
Figure 4: Comparison of TOCSY and NOESY spectra of pTH spectra (29) that have been obtained in independent experiments using different carrier offsets and sweep widths, may be directly compared after appropriate scaling. The two spectra are shown side by side in the black/white illustration for clarity; in a proper overlay on a color screen display differences are revealed immediately.

Portability was achieved by developing the code in standard C programming language, using the environment of the UNIX operating system. The graphics output was written for the standard X-Windows interface. This restriction with regard to a standard compiler, operating system, and graphics interface results in a straightforward compilation of the program on various platforms. NDEE was shown to run on ESV and Silicon Graphics workstations, DEC stations, SPARC stations, and a CRAY system without any complications. Thus the operation of the program is independent of any particular processing or display hardware. As the program may be used in a network file system, any X-Windows terminal may be used for data output.

Raw data input files from an NMR spectrometer are first converted into the NDEE format. The user intuitive surface was developed for easy control. The panels and windows of the surface are based on pure X-Windows function calls, which results in an uniform appearance on any suitable terminal. The use of the X-Motif library was avoided to achieve wider

portability, faster performance and a considerable reduction of the binary code.

A prominent feature is the 'multiplot' window in which several spectra may be overlayed after suitable scaling. In this manner, a series of contour plots may be directly compared on screen regardless of the origin of the spectra, spectrometer frequency, sweep width, etc.

The display of the data cube corresponding to a 3D NMR file was also programmed in X-Windows, to avoid the compulsory use of more expensive hardware equipped with PEX (22) or GL (21) graphics accelerators. As an example of a 3D spectrum, a part of the ^{15}N-HMQC TOCSY spectrum of EIAV-Tat (30) is shown in Figure 5. The unravelling of the overlapping multiplet patterns by 3D spreading is clearly visible.

In summary, in the NDEE program system the entire data analysis is carried out on screen. Processing starts with the raw data and ends with high quality graphics output suitable for reproduction including all necessary annotations. A conve-
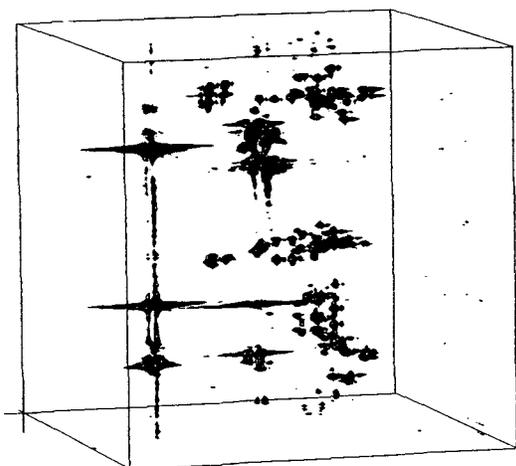
Figure 5: A 3D data cube from a ^{15}N-HMQC TOCSY experiment performed on EIAV-Tat (30).

nient interface to standard molecular dynamics programs has been realized. The on-screen edited NOE and J constraints are converted, by a single menu-prompted switch, into a formatted file for use with codes such as XPLOR (27) or GROMOS (28).

The contour plot algorithm implemented in the NDEE program package is a promising platform for tackling the problem of pattern recognition and automated assignment of protein spectra.

A demo version of the program is available from the author (F. Herrmann) and via the anonymous ftp of Bayreuth University (132.180.8.29).

V.   References

[1] G.W. Vuister, R. Boelens, *J. Magn. Reson.* **1987**, *73*, 328.

[2] C. Griesinger, O.W. Sørensen, R.R. Ernst, *J. Magn. Reson.* **1987**, *73*, 574.

[3] H. Oschkinat, C. Griesinger, P.J. Kraulis, O.W. Sørensen, R.R. Ernst, A.M. Gronenborn, G.M. Clore, *Nature* **1988**, *332*, 374.

[4] C. Griesinger, O.W. Sørensen, R.R. Ernst, *J. Magn. Reson.* **1989**, *84*, 14.

[5] S.W. Fesik, E.R.P. Zuiderweg, *J. Magn. Reson.* **1988**, *78*, 588.

[6] L.E. Kay, D. Marion, A. Bax, *J. Magn. Reson.* **1989**, *84*, 72.

[7] M. Ikura, L.E. Kay, A. Bax, *Biochemistry* **1990**, *29* 4659.

[8] R.E. Hoffman, G.C. Levy, *Prog. NMR Spect.* **1991**, *23*, 211.

[9] T. Havel, I.D. Kuntz, G.M. Crippen, *Bull. Math. Biol.* **1983**, *45*, 665.

[10] W. Braun, N. Gō, *J. Mol. Biol.* **1985**, *186*, 611 . [11] T.F. Havel, K. Wüthrich, *Bull. Math. Biol.* **1984**, *46*, 673 . [12] J.A. McCammon, S.H. Harvey, *Dynamics of proteins and nucleic acids* Cambridge University Press, New York **1987**.

[13] M. Karplus, G.A. Petsko, *Nature* **1990**, *347*, 631 . [14] W.F. van Gunsteren, A.E. Mark, *Eur. J. Biochem.* **1992**, *204*, 947.

[15] B.U. Meier, Z.L. Mádi, R.R. Ernst, *J. Magn. Reson.* **1987**, *74*, 565 . [16] H. Grahn, F. Delaglio, M.A. Delsuc, G.C. Levy, *J. Magn. Reson.* **1988**, *77*, 294 . [17] L. Emsley, G. Bodenhausen, *J. Am. Chem. Soc.* **1991**, *113*, 3309.

[18] D.S. Garret, R. Powers, A.M. Gronenborn, G.M. Clore, *J. Magn. Reson.* **1991**, *95*, 214.

[19] M. Kjær, F.M. Poulsen, *J. Magn. Reson.* **1991**, *94*, 659 . [20] *The X Window System Series*, Vol. *1-7*, O'Reilly & Associates, Inc., Sebastopol, **1988**.

[21] Graphics Library, Silicon Graphics, Inc.; California.

[22] PHIGS Extension to X, Evans & Sutherland Workstations Reference Manual.

[23] NMRZ, Tripos Inc.

[24] FELIX, Biosym Technologies Inc.

[25] UXNMR and AURELIA, Bruker Analytische Meßtechnik, Karlsruhe.

[26] EASY, *J. Biomol. NMR* **1991**, *1*, 111.

[27] A.T. Brünger, *Methods and Applications in Crystallographic Computing* (N. Isaacs, ed.) Oxford Press, Oxford, Great Britain, **1987**, 613.

[28] W.F. Van Gunsteren, R. Kaptein, E.R.P. Zuiderweg, *Nucleic acid conformation and dynamics* (W.K. Olson, ed.) pp. 79-92, Report of NATO/CECAM Workshop, Orsay, France, **1983**.

[29] U. Marx, S. Austermann, W.-G. Forssmann, F. Herrmann, P. Rösch, to be published.

[30] D. Willbold, P. Bayer, Rosin-Ardesfeld, A. Gazit, A. Yaniv, F. Herrmann, P. Rösch, to be published.