Journal of Biomolecular NMR, 5 (1995) 1–10
ESCOM

J-Bio NMR 272

The program XEASY for computer-supported NMR spectral analysis of biological macromolecules

Christian Bartels, Tai-he Xia, Martin Billeter, Peter Güntert and Kurt Wüthrich*

*Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg,
CH-8093 Zürich, Switzerland*

Received 23 January 1995
Accepted 16 February 1995

Keywords: NMR structure determination; Interactive computer graphics for support of NMR analysis; Peak picking and integration; Sequence-specific NMR assignments for biological macromolecules

Summary

A new program package, XEASY, was written for interactive computer support of the analysis of NMR spectra for three-dimensional structure determination of biological macromolecules. XEASY was developed for work with 2D, 3D and 4D NMR data sets. It includes all the functions performed by the precursor program EASY, which was designed for the analysis of 2D NMR spectra, i.e., peak picking and support of sequence-specific resonance assignments, cross-peak assignments, cross-peak integration and rate constant determination for dynamic processes. Since the program utilizes the X-window system and the *Motif* widget set, it is portable on a wide range of UNIX workstations. The design objective was to provide maximal computer support for the analysis of spectra, while providing the user with complete control over the final resonance assignments. Technically important features of XEASY are the use and flexible visual display of 'strips', i.e., two-dimensional spectral regions that contain the relevant parts of 3D or 4D NMR spectra, automated sorting routines to narrow down the selection of strips that need to be interactively considered in a particular assignment step, a protocol of resonance assignments that can be used for reliable bookkeeping, independent of the assignment strategy used, and capabilities for proper treatment of spectral folding and efficient transfer of resonance assignments between spectra of different types and different dimensionality, including projected, reduced-dimensionality triple-resonance experiments.

Introduction

With the widespread use of NMR structure determination for proteins and nucleic acids (Wüthrich, 1986), much interest has been focused on the development of computer facilities to support the labour-intensive steps of such projects (e.g., Neidig et al., 1984; Pfändler et al., 1985; Glaser and Kalbitzer, 1987; Hoch et al., 1987; Novic et al., 1987; Billeter et al., 1988; Cieslar et al., 1988; Grahn et al., 1988; Eads and Kuntz, 1989; Kleywegt et al., 1989,1991,1993; Kraulis, 1989; Stoven et al., 1989; Weber et al., 1989; Ikura et al., 1990; Pfändler and Bodenhausen, 1990; Van de Ven, 1990; Eccles et al., 1991; Garret et al., 1991; Gray and Brown, 1991; Bernstein et al., 1993; Wehrens et al., 1993; Hare and Prestegard, 1994; Meadows et al., 1994). In principle, either fully automated computational procedures or graphics-based interactive approaches can be envisaged. Success with fully automated procedures has so far been limited, mainly because of ambiguities arising from experimental artefacts, but important help in the determination and bookkeeping of peak positions, intensities and assignments has resulted from interactive routines. In our laboratory the program EASY was developed for this purpose (Eccles et al., 1991). Further use of the EASY package is limited by the fact that the program was implemented to operate within the *Sunview* window environment, and its capabilities were tailored for work with 2D

*To whom correspondence should be addressed.

Abbreviations: 1D, 2D, 3D, 4D, one-, two-, three-, four-dimensional; NOE, nuclear Overhauser enhancement; NOESY, nuclear Overhauser enhancement spectroscopy; TOCSY, total correlation spectroscopy; COSY, correlation spectroscopy; TPPI, time-proportional phase incrementation.

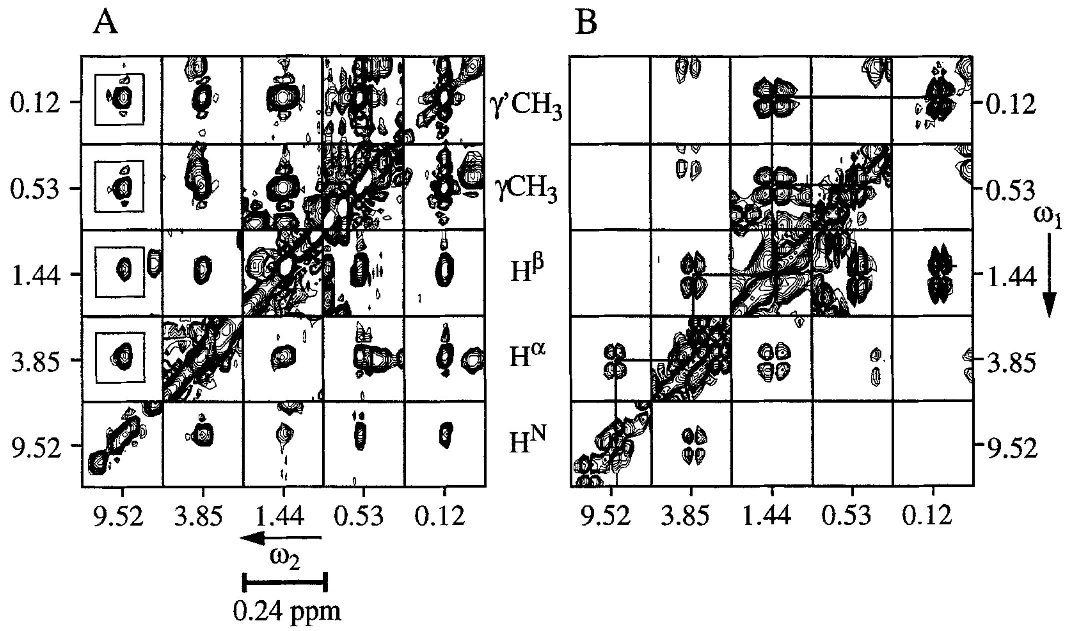


Fig. 1. Spectral reduction by XEASY for spin system identification in homonuclear 2D proton TOCSY (A) and COSY (B) spectra. The spectra are from the mutant α-amylase inhibitor Tendamistat(R19L), a protein with 74 residues (O'Connell et al., 1994). The TOCSY cross peaks with the amide proton of Val[5] are all at 9.52 ppm along ω_2, and regions containing these peaks were zoomed at the ω_1 frequencies of 0.12, 0.53, 1.44, 3.85 and 9.52 ppm; to this end, these frequencies were defined by simply clicking the mouse at the positions of the four cross peaks marked by a square. Using the information on resonance frequencies thus obtained, the additional TOCSY cross peaks of the spin system were similarly zoomed. The operator can readily overlay the corresponding regions from the COSY spectrum (B) on the screen to identify the spin system type and to uniquely assign the individual atoms of the spin system. The figure contains the resonance frequencies in ppm along ω_1 and ω_2, and the size of the zoomed regions in ppm.

NMR spectra. The presently described new interactive software package XEASY is portable on a wide range of UNIX workstations that use the X-window system (Scheifer et al., 1988) and its capabilities have been tailored for work with 3D and 4D as well as 2D NMR spectra. This paper gives a brief survey of the novel features implemented in XEASY and describes its practical use. The program and a detailed manual can be obtained from the authors.

Methods

The program package XEASY is written in the C programming language and it uses the *Motif* widget set (Heller, 1991) for the user interface. It is menu driven, but supports also two-letter commands as keyboard short-cuts. In our software environment, the analysis of NMR spectra with XEASY is a central step in the structure determination of biomacromolecules. XEASY accepts the initial input of frequency-domain spectra either from the program PROSA (Güntert et al., 1992) or from standard Bruker and Varian processing software. In the later stages of the spectral analysis, it takes additional input on possible resonance assignments from ASNO (Güntert et al., 1993). Its output is sent to DIANA and supporting programs (Güntert et al., 1991) for the structure calculation. To ensure efficient transfer of data, XEASY uses identical file formats as the surrounding software written in our laboratory, but it is suitable for efficient data transfer with other software packages as well. In the following we describe selected features that are of special interest for the operation of XEASY.

Display of relevant spectral regions

With XEASY 1D and 2D regions from one or several spectra are displayed, which may have different dimensionality. Special features support proper alignment of the displayed regions from different spectra for detailed comparison of peak positions and peak shapes. These routines allow, for example, interactive removal of momentarily irrelevant regions from the display, efficient stepping through the planes of 3D and 4D spectra, or reduction of spectra for display of multiple small regions at selected frequencies (Fig. 1). Several of these display modes will be illustrated in the Applications section below.

Automated sorting of spectral strips

In XEASY the inherent technical difficulties in efficient handling of 3D and 4D spectra are tackled using the concept of decomposition of the higher dimensional spectra into 2D strips (e.g., Driscoll et al., 1990; Wüthrich et al., 1991). The strip positions are defined by the coordinates of peaks which may be obtained from more readily accessible spectra, usually 2D [^{15}N,^{1}H]-COSY or 2D [^{13}C,^{1}H]-COSY, and the strips extend in one dimension (the 'vertical dimension' in the common presentation) over the entire spectral sweep width and in the second dimension over a user-defined, usually much smaller width. Starting from a given strip, s, assignments can in general be made by identification of strips that are characteristically related to s. This can be achieved by automated ranking of all other strips, k, by the closeness of coincidence of one or several peaks with related peaks in strip s, as expressed by the quantity d_k (Eq. 1). In this assignment process, XEASY is used for two different purposes. Firstly, it performs the automated ranking of the strips k. Secondly, the low-ranking strips, i.e., the strips that are most likely to correspond to the desired assignment, can be displayed in order to allow the final decision to be made interactively by the user. For the automated ranking of the strips, the previously introduced dot-product correlation function (Bartels and Wüthrich, 1994) was supplemented with a preliminary screening step, for the following reason. In most resonance assignment procedures (the previously considered sequential assignment using 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY is one of the exceptions), it is possible to identify a single frequency, f, in the reference strip s such that identification of a peak with the corresponding resonance frequency in one of the strips k can provide the desired assignment. Experience has shown that for such assignment steps the dot-product correlation function may, for obvious reasons (Bartels and Wüthrich, 1994), erroneously identify strips k as being low-ranking, although they do not contain any peak near the frequency f. Such potential pitfalls are circumvented without loss of any of the potentialities of the dot-product approach by using Eq. 1.

$$d_k = \begin{cases} 1 + |f - p_k| & \text{if } |f - p_k| > \Delta \\ 1 - \left| \dfrac{\vec{v}_k \cdot \vec{v}_e}{|\vec{v}_k||\vec{v}_e|} \right| & \text{if } |f - p_k| \le \Delta \end{cases} \qquad (1)$$

If f is the anticipated frequency of a peak in the desired strip k, the condition $|f - p_k| > \Delta$ eliminates those strips defined by peaks for which the position p_k differs by more than the user-specified parameter Δ from f. The parameter Δ is usually set to a value that is larger than the expected error in the determination of either the frequency f or the peak positions p_k, and, depending on the type of assignment to be made, it may also be set either to 0 or to ∞ (see Applications section). The remaining strips with $|f - p_k| \le \Delta$ are sorted according to the dot-product correlation function (Bartels and Wüthrich, 1994), which is a measure of the similarity between the peak pattern $\vec{v}_k$ observed in strip k and the peak pattern $\vec{v}_e$ expected for the strip that would provide the desired assignment. The peak patterns $\vec{v} = (i_1, i_2, \ldots, i_n)$ are n-dimensional vectors, with n equal to the number of data points along the vertical dimension. The vectors $\vec{v}$ are derived from the experimentally observed peak intensities in the strips, i_j^{ex}, according to

$$i_j = A_m \frac{i_j^{ex}}{|i_m^{ex}| + i_b} \qquad (p_l \le j < p_h) \qquad (2)$$

where i_m^{ex} is the experimental intensity at the local maximum m of the absolute peak intensity, p_l and p_h are the positions of the two adjacent local minima of the absolute intensity, and i_b is a parameter that is typically set to 15 times the standard deviation of the noise. The constant A_m gives a weight to every local maximum and can thus be used to emphasize or suppress subsets of peaks (for details, see Bartels and Wüthrich, 1994).

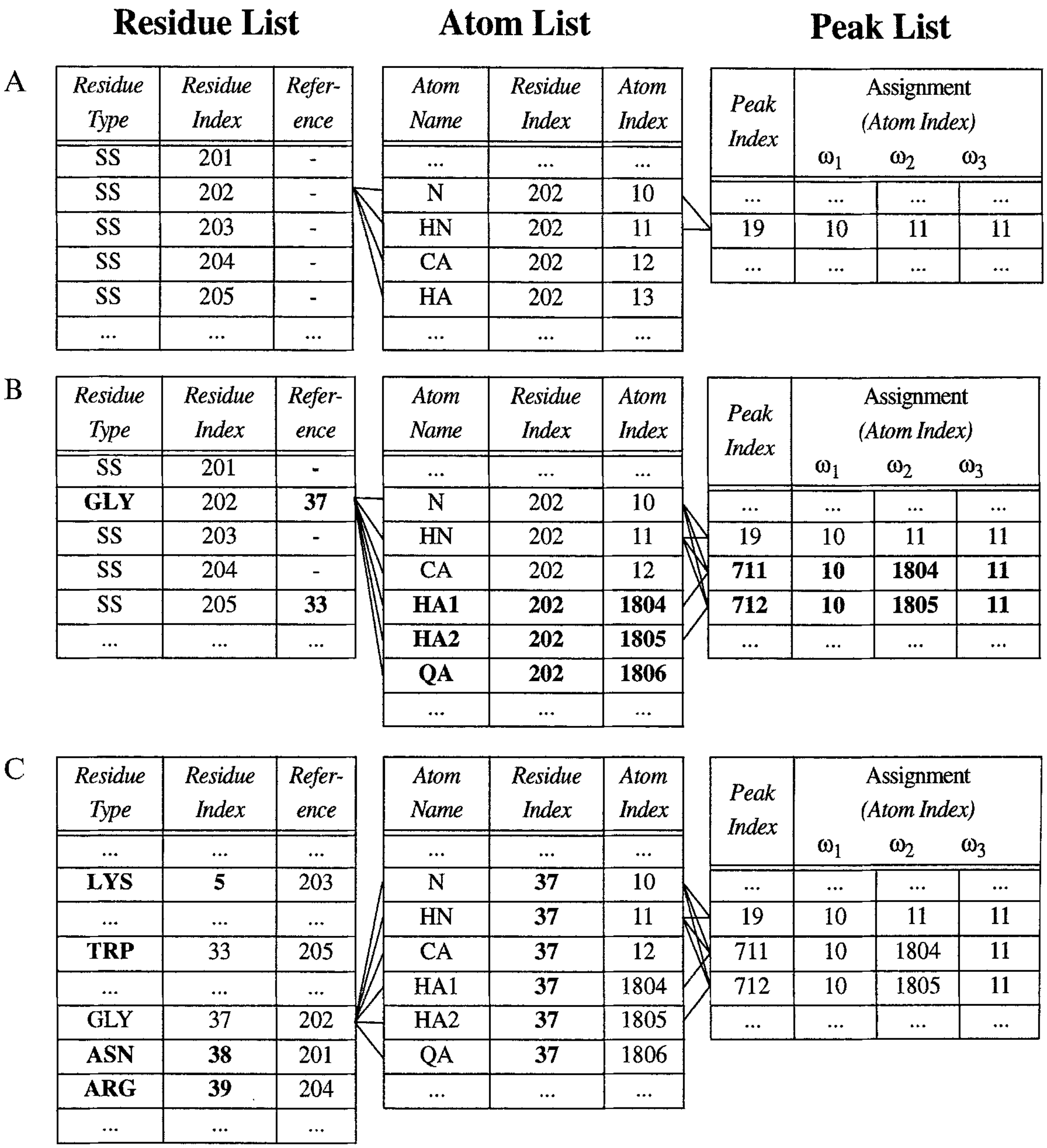
Determination of likely residue types

Routines have been implemented that identify likely amino acid types for a given sequence position, based on the set of NMR frequencies observed for this residue. The user specifies the set of frequencies by clicking the mouse on the peaks that belong to the residue in question, and XEASY will automatically provide a list of amino acid types, sorted by the similarity of the expected chemical shifts to the number of specified frequencies and their values. In the present implementation, the reference chemical shifts were taken from Gross and Kalbitzer (1988) for ^{1}H and from Richarz and Wüthrich (1978) for ^{13}C. To determine the optimal match to the set of observed frequencies, a standard algorithm for identification of maximal weighted matchings (e.g. Jungnickel, 1990) has been implemented.

Bookkeeping of resonance assignments

The data structures used by XEASY to store and update assignment information in the course of a structure determination are illustrated in Fig. 2. Three lists are used, where the residue list contains all the residues to be assigned, the atom list contains the corresponding atoms, and the peak list contains the spectral positions and assignments of picked NMR peaks. Two indices (*Residue Index* and *Atom Index*) relate the entries in the three lists, as is shown in Fig. 2 by lines connecting the different lists.

Figure 2A shows the data structure at the outset, when neither the amino acid type corresponding to a given *Residue Index* nor the sequence number are known, and only a few NMR peaks have been assigned. The residue type SS ('generic spin system') reflects the lack of information on the amino acid type. The *Atom Names* in the atom list are taken from a fragment library in which the atom names belonging to the different possible residue types are specified. For SS, only the HN atom and the backbone atoms that are common to all amino acid residues are defined. Figure 2B depicts the data structure in the course of the resonance assignment. The residue with *Residue Index* 202 has been identified as Gly^{37}, and there-



A

Residue Type	*Residue Index*	*Reference*
SS	201	-
SS	202	-
SS	203	-
SS	204	-
SS	205	-
...	...	...

Atom Name	*Residue Index*	*Atom Index*
...	...	...
N	202	10
HN	202	11
CA	202	12
HA	202	13
...	...	...

Peak Index	Assignment *(Atom Index)*		
	ω_1	ω_2	ω_3
...	...	...	...
19	10	11	11
...	...	...	...

B

Residue Type	*Residue Index*	*Reference*
SS	201	-
GLY	202	**37**
SS	203	-
SS	204	-
SS	205	**33**
...	...	...

Atom Name	*Residue Index*	*Atom Index*
...	...	...
N	202	10
HN	202	11
CA	202	12
HA1	**202**	**1804**
HA2	**202**	**1805**
QA	**202**	**1806**
...	...	...

Peak Index	Assignment *(Atom Index)*		
	ω_1	ω_2	ω_3
...	...	...	...
19	10	11	11
711	**10**	**1804**	**11**
712	**10**	**1805**	**11**
...	...	...	...

C

Residue Type	*Residue Index*	*Reference*
...	...	...
LYS	**5**	203
...	...	...
TRP	33	205
...	...	...
GLY	37	202
ASN	**38**	201
ARG	**39**	204
...	...	...

Atom Name	*Residue Index*	*Atom Index*
...	...	...
N	**37**	10
HN	**37**	11
CA	**37**	12
HA1	**37**	1804
HA2	**37**	1805
QA	**37**	1806
...	...	...

Peak Index	Assignment *(Atom Index)*		
	ω_1	ω_2	ω_3
...	...	...	...
19	10	11	11
711	10	1804	11
712	10	1805	11
...	...	...	...

Fig. 2. Data structure used by XEASY for bookkeeping of resonance assignments. A residue list, an atom list and a peak list are used. Lines between the three lists connect entries related by the *Residue Index* and the *Atom Index*, respectively. (A) Situation at the outset of the resonance assignment; SS is a generic amino acid residue with atom names attributed only to the backbone fragment that is common to all residues. The assignment given here corresponds to that obtained by picking the diagonal peaks in a 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY spectrum. (B) Situation during the resonance assignment procedure, reflecting the finding that residue 202 corresponds to Gly^{37} in the protein sequence, and that the complete Gly^{37} spin system has been identified. In the column *Atom Name*, QA defines a pseudoatom (Wüthrich et al., 1983). Furthermore, without displaying the supporting entries in the atom list and the peak list, the residue list indicates that residue 205 is in position 33 in the protein sequence. The temporary sequence numbers of the residues are listed in the *Reference* column. (C) Situation after completion of the resonance assignments: the *Residue Type* and *Residue Index* entries now refer to the amino acid sequence of the protein, where the supporting entries in the atom list and the peak list are given only for Gly^{37}; the original *Residue Index* is kept in the column *Reference*. Bold entries in (B) and (C) emphasize changes in the lists compared to situations (A) and (B), respectively. See text for further details on the use of this data structure.

fore the *Residue Type* has changed to GLY and the sequence number 37 is stored in the *Reference* column. When the residue type is specified, XEASY automatically adds all the atoms that belong to this residue type to the atom list, including pseudoatoms (Wüthrich et al., 1983) where applicable. These new atoms can be used to assign additional peaks to the residue. Figure 2C shows the situation after completion of the resonance assignments, where the different lists contain the information most suitable for further cross-peak assignments and for the structure calculation, i.e., the residue list corresponds to the amino acid sequence of the protein and the *Residue*

Index entries contain directly the sequence numbers of the amino acid residues. This is achieved by sorting the residue list according to the sequence numbers in the *Reference* column, and by subsequent exchange of the entries in the columns *Residue Index* and *Reference*.

Transfer of picked peaks and resonance assignments between different spectra

When working with 2D [^{1}H,^{1}H]-NMR spectra, transfer of peak assignments requires due attention to possible slight differences between chemical shifts of corresponding signals in different spectra. To this end, we retained a routine from the EASY program that ensures simultaneous chemical shift adjustments for all peaks that originate from the same atom (Eccles et al., 1991). For work with spectra of different dimensionalities and sweep widths, XEASY stores 'folding numbers' in addition to peak positions, which indicate how many times a peak has been folded. This allows one to retain the folding information when transferring picked peaks between different spectra or when picking additional peaks with known assignment, and to display the unfolded chemical shifts whenever this is needed. In the example of Fig. 3A, the filled peak at (ω_1(^{13}C) = 70 ppm, ω_2(^{1}H) = 3.9 ppm) is not folded, i.e., it has the folding numbers (0, 0). When this peak is transferred to a spectrum with a spectral range in the ^{13}C-dimension from 30 to 55 ppm, its position changes to (ω_1(^{13}C) = 45 ppm, ω_2(^{1}H) = 3.9 ppm) for complex data (States et al., 1982) and its folding numbers change to (−1, 0).

For the transfer of peaks into a spectrum with different dimensionality, XEASY provides a routine that defines the dimensions of the source spectrum from which to copy the shifts, and assigns peak positions and folding numbers for each dimension of the destination spectrum. In the peak transfer of Fig. 3, the peak at (ω_1(^{13}C) = 70 ppm, ω_2(^{1}H) = 3.9 ppm) in the 2D spectrum is at (ω_1(^{13}C) = 45 ppm, ω_2(^{1}H) = 3.9 ppm, ω_3(^{1}H) = 3.9 ppm) in the 3D spectrum, with folding numbers (−1, 0, 0).

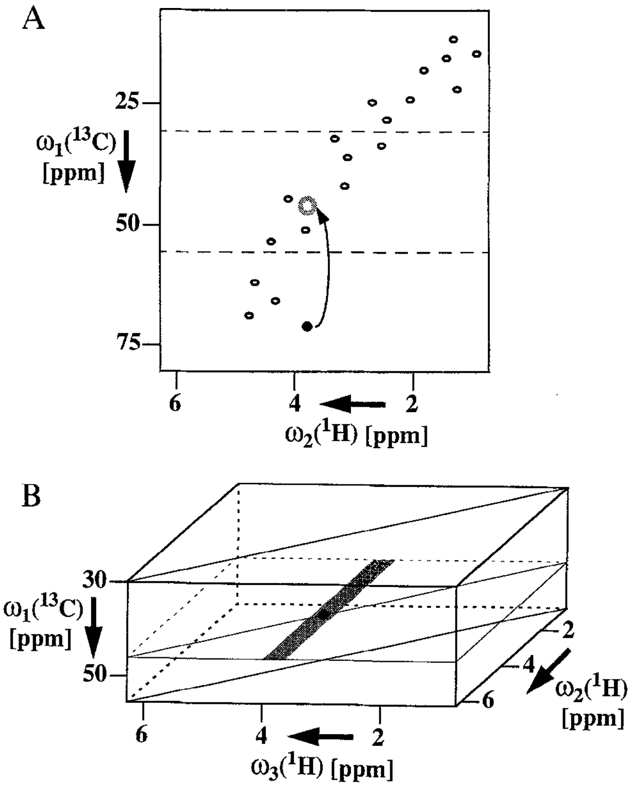


Fig. 3. Schematic representation of an XEASY transfer of peaks between a 2D spectrum and a 3D spectrum with different sweep widths. (A) 2D [^{13}C,^{1}H]-COSY spectrum with some cross peaks. The range bounded by the two dotted lines corresponds to the reduced sweep width used in the ^{13}C dimension of the 3D spectrum (B), and the arrow points to the position where the filled peak would be observed in a spectrum recorded with the indicated reduced sweep width. (B) 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY spectrum. The filled peak at (ω_1(^{13}C) = 70 ppm, ω_2(^{1}H) = 3.9 ppm) in (A) is transferred by copying its position, folding number and assignment from ω_2(^{1}H) of the 2D spectrum to both proton dimensions of the 3D spectrum, and by adapting the folding numbers to the sweep widths in the 3D spectrum; it is assumed that the spectrum was recorded with the method of States et al. (1982) (see text for details). In practice, peak transfer from 2D [X,^{1}H]-COSY to 3D spectra is used to identify strips such as that indicated by the shaded region (see text).

Applications

This section describes applications of XEASY for support of sequential assignment, spin system identification and NOESY cross-peak assignment. In selecting these illustrations it has been assumed that a 2D [^{15}N,^{1}H]-COSY spectrum is first analyzed to obtain peak positions needed for work with 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY (see Bartels and Wüthrich, 1994) and triple-resonance experiments (for a review on triple-resonance experiments see, for example, Bax and Grzesiek, 1993), and that 2D [^{13}C,^{1}H]-COSY is used in an analogous way to provide peak positions needed for the analysis of 3D HCCH-TOCSY spectra. In turn, the complete spin system identifications obtained with 3D HCCH-TOCSY are needed as a basis for the analysis of ^{13}C-resolved [^{1}H,^{1}H]-NOESY. In each example we proceed in three steps: (i) definition of a set of strips which is suitable for the particular assignment step and represents the whole data set; (ii) automated strip sorting to produce lists of related strips; and (iii) visual display of the strips grouped together in step (ii) to interactively decide on the final, unique resonance assignments. For each example, the quality of the automated strip sorting in step (ii) is represented by a statistical analysis of the ranks found for the strips representing the desired assignment. To produce realistic statistics and simulate the fact that the frequency f in Eq. 1, when defined interactively, usually deviates somewhat from the desired position p_k, the frequencies f are randomly chosen from normal distributions centered about p_k. The standard deviation of this distribution is specified in each case.

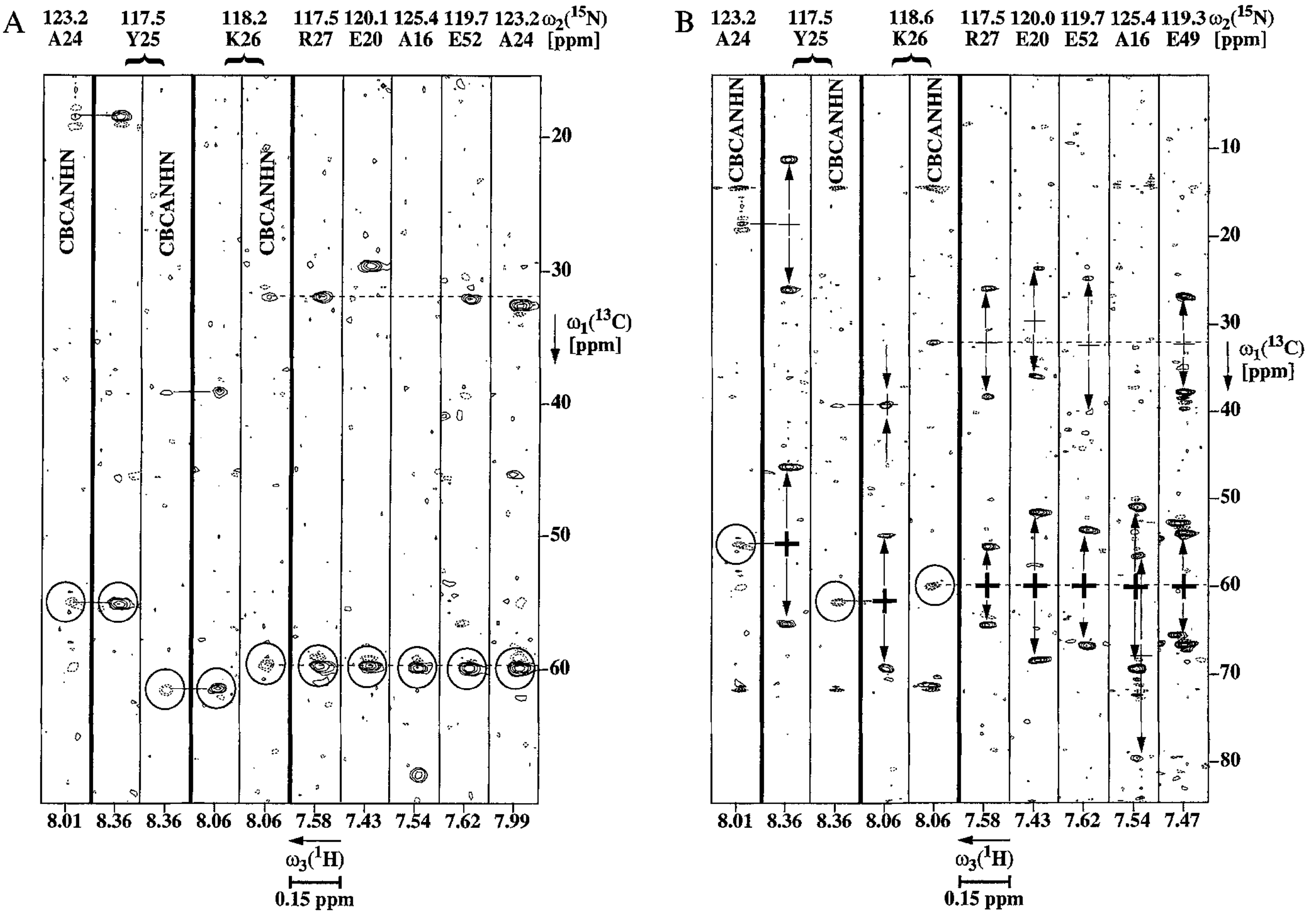


Fig. 4. Illustration of sequence-specific assignments with combined use of a 3D CBCANHN (Grzesiek and Bax, 1992b) spectrum and either a 3D CBCA(CO)NHN (Grzesiek and Bax, 1992a) (A) or reduced-dimensionality 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN (Szyperski et al., 1994a) spectrum (B). The data were recorded with [^{15}N,^{13}C]-labeled DnaJ(2–108) (Szyperski et al., 1994b). Strips extending over the entire spectral width in the ^{13}C dimension are defined for the amide groups of the residues indicated above each strip, where strips from the CBCANHN spectrum are labeled as such. In ω_2 the strips are located at the $^{15}N_i$ frequency listed at the top, and in ω_3 they are centered about the frequency listed below each strip. Both snapshots (A) and (B) were taken after completing the assignments for the tripeptide segment Ala^{24}–Tyr^{25}–Lys^{26}, when the assignment for residue 27 was in progress. In the first four strips on the left, horizontal lines indicate the connectivities used to establish the sequential relationships. The peaks in the fifth strip define the C^α frequency (marked by a broken horizontal line near 60 ppm) and the C^β frequency (marked by a broken horizontal line near 32 ppm) of Lys^{26}. The additional strips 6 to 10 are the lowest-ranked candidates for the sequential neighbour of Lys^{26} proposed by XEASY. (A) The intraresidual $N_i - C_i^\alpha - H_i^N$ peaks in the CBCANHN strips and the sequential $N_i - C_{i-1}^\alpha - H_i^N$ peaks in the CBCA(CO)NHN strips are encircled. The candidate strips 6 to 10 were sorted by increasing distance of their sequential peak to the C^α frequency of Lys^{26} and the similarity of their peak patterns with the pattern observed in the fifth strip (Eq. 1, $\Delta = 0.5$ ppm). (B) In the strips from the 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN spectrum, bold and thin crosses mark the C_{i-1}^α and the C_{i-1}^β frequency, respectively, which were defined as described in the text (in strip 4 the peak pair of the C^β atom is degenerated to a single peak due to accidental interference with the ^{1}H carrier frequency). The candidate strips 6 to 10 were sorted based on the distance of the bold crosses from the C^α frequency of Lys^{26} (Eq. 1, $\Delta = 0$ ppm). Although the ranking of candidate strips in the two experiments (A) and (B) is slightly different, Arg^{27} was correctly identified in both cases based on the relative distances of the observed C^β frequency in the candidate strips from the C^β frequency of Lys^{26}.

Sequence-specific assignment with combined use of 3D CBCANHN and 3D CBCA(CO)NHN spectra or, alternatively, with 3D CBCANHN and 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN spectra

(i) All entries in the peak list obtained from complete peak picking of the 2D [^{15}N,^{1}H]-COSY spectrum are loaded into a [^{15}N,^{1}H]-plane at an arbitrarily selected ^{13}C frequency of the CBCA(CO)NHN spectrum. For each of these entries a ^{13}C-^{1}H strip extending over the entire spectral width in the ^{13}C dimension is displayed, within which the user picks the positions of the observed sequential $N_i - C_{i-1}^\alpha - H_i^N$ and $N_i - C_{i-1}^\beta - H_i^N$ peaks. At the outset of the spectral analysis, these two peaks cannot a priori be individually identified based on their chemical shifts or their sign. Therefore, each strip is attributed to both of the two sequential peaks, so that the ^{13}C frequencies of these peaks replace the arbitrarily selected initial ^{13}C frequency as values of p_k used for the strip sorting (Eq. 1). In an analogous manner, strips are defined in the CBCANHN spectrum.

(ii) The user defines the frequency f in Eq. 1 to be the C_i^α frequency observed in a selected reference strip of the

CBCANHN spectrum. XEASY then ranks possible sequential neighbour strips, taken from the CBCA(CO)NHN spectrum, according to increasing distance of their sequential cross peaks from f and the overall similarity of their peak patterns with the pattern observed in the reference strip (Eq. 1). In the example of Fig. 4A, setting $\Delta = 0.5$ ppm and choosing the frequencies f from normal distributions about p_k with standard deviation 0.2 ppm, 69% of the actual sequential neighbour strips had rank 1, 93% had a rank ≤ 5 and the poorest ranking was 8.

(iii) 1D cross sections through the sequential and intraresidual peaks can be displayed for comparison of the peak line shapes with each newly found candidate strip. Using information from other spectra on the spin system types (see, for example, Fig. 5), the user aligns a stretch of sequentially neighbouring residues to the amino acid sequence.

Figure 4B illustrates that XEASY is well equipped for handling reduced-dimensionality triple-resonance experiments (Szyperski et al., 1993a,b,1994a; Simorre et al., 1994) as well. The advantage of reduced-dimensionality experiments is that n + 1 frequencies can be correlated by a peak pair in an n-dimensional spectrum. For example, peak pairs in a 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN spectrum correlate four frequencies in three dimensions. They establish relations between the polypeptide backbone and multiple frequencies from a given amino acid side chain, and thus provide a convenient starting point for sequence-specific assignment, where the 3D $H^{\alpha/\beta}C^{\alpha/\beta}$(CO)NHN spectrum is used in place of the 3D CBCA(CO)NHN experiment. To enable a straightforward correlation with strips from the 3D CBCANHN experiment, XEASY provides the facility to convert the difference frequencies in the reduced-dimensionality experiment to chemical shift positions. To this end, the user simply clicks the mouse at the positions of the two peaks of each pair and thus adds entries to the peak list (Fig. 2), with the positions in the ^{15}N and ^{1}H dimensions corresponding to the N_i and H_i^N frequencies, and in the ^{13}C dimension either to the C_{i-1}^{α} or the C_{i-1}^{β} frequency. Using these entries in the peak list and setting $\Delta = 0$ ppm in Eq. 1, the assignment is carried out as described above for the case of the CBCA(CO)NHN experiment (Fig. 4B). The loss of peak shape information when compared to the 3D CBCA(CO)NHN experiment does not prevent the sorting routine to produce a useful ranking of the candidate strips. For the example of Fig. 4B, choosing the frequencies f from normal distributions with standard deviation 0.2 ppm, 38% of the sequentially neighbouring strips were attributed rank 1, 87% had a rank ≤ 5 and the worst ranking was 12.

Spin system identification using 3D [HCCH]-TOCSY

(i) For each peak in the diagonal plane with identical proton chemical shifts, an $\omega_2(^1H) - \omega_3(^1H)$ strip is defined that represents the corresponding proton (Fig. 3). The previously picked peaks in the 2D [^{13}C,^{1}H]-COSY spectrum are transferred into the 3D [HCCH]-TOCSY spectrum, where spectral regions around each peak can be displayed for interactive adjustment of the peak positions in the 3D spectrum before the strips attributed to the aforementioned diagonal peaks are defined.

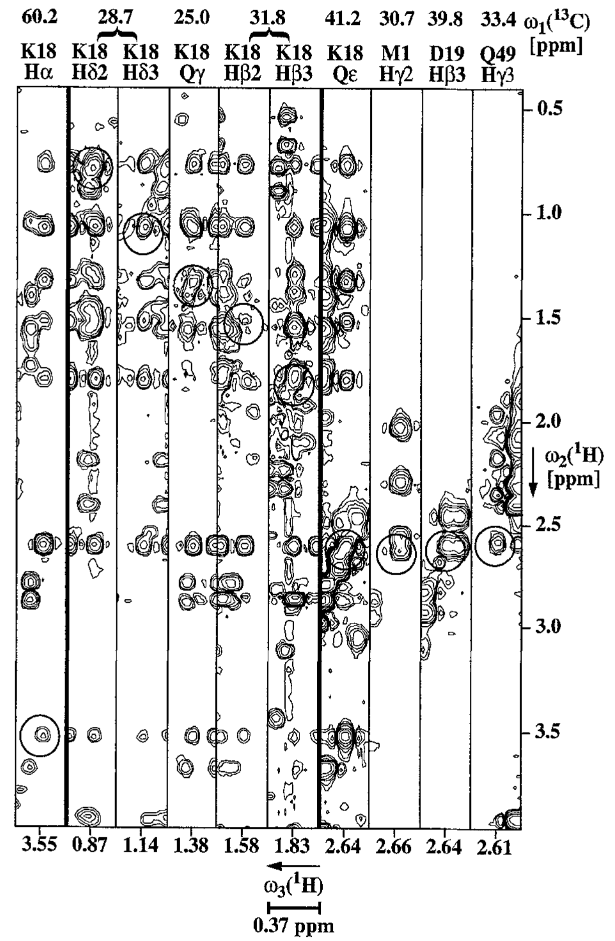


Fig. 5. Spin system identification in an [HCCH]-TOCSY spectrum (Bax et al., 1990) using XEASY, illustrated with the [^{15}N,^{13}C]-labeled mixed disulfide between *E. coli* glutaredoxin and glutathione (Bushweller et al., 1994). The assignment above each strip is for the residue and the proton type corresponding to the encircled diagonal peak. In ω_1 the strips are at the ^{13}C positions of these encircled peaks, and along ω_3 they are centered about the proton frequency of these peaks. The frequencies of the encircled peaks prior to any folding operation (see also Fig. 3) are listed above and below each strip. From left to right we refer to the strips as 1 to 10. Strip 1 corresponds to H^{α} of Lys^{18} and is used as a reference to search for the remaining strips of the Lys^{18} spin system. At the stage of the analysis illustrated here, strips 2–6 had already been found to correspond to the same spin system as strip 1. Strips 7–10 are candidates provided by XEASY for the strip of εCH_2, based on the small distance between the diagonal peaks (encircled) and the remaining cross peak in strip 1, and the similarity of their peak patterns with the pattern observed in strip 1 (Eq. 1). From the complete peak pattern, the strip at ($\omega_1 = 41.2$ ppm, $\omega_3 = 2.64$ ppm) could then be identified as belonging to the same spin system as the reference strip, either from the value of the dot-product correlation function (Eq. 1), or interactively by visual inspection.

(ii) The user selects an arbitrary strip s as a reference. The peaks in this reference strip define all or, in less favourable cases, some of the frequencies of the proton spin system of one of the amino acid residues in the protein. For each proton with frequency f represented by a cross peak in the reference strip, candidates for the strip containing the corresponding diagonal peak are identified and ranked using Eq. 1 (Fig. 5). With $A_m = 1$ for all local maxima m, i_b set to 15 times the standard deviation of the noise, $\Delta = 0.03$ ppm (Eqs. 1 and 2) and selection of the frequencies f from normal distributions with standard deviation 0.02 ppm, 61% of the correct strips were attributed rank 1, in 88% of the cases they had a rank ≤ 5, and no correct strip had a rank worse than 20 in the mixed disulfide of glutaredoxin and glutathione (Fig. 5). The assignments that were eventually found to correlate with poorly ranked strips all originated from heavily overlapped spectral regions.

(iii) By visual inspection of the group of low-ranking strips obtained in step (ii), those strips k are identified for which the diagonal peak represents a proton that belongs to the same spin system as the reference s. The ensemble of all these strips normally enables identification of all frequencies belonging to the spin system, and the corresponding amino acid types are automatically identified by XEASY. To resolve possible remaining ambiguities, the strips from the standard 'long-range' 3D [HCCH]-TOCSY spectrum can be displayed jointly with the corresponding regions from either a 3D [HCCH]-COSY spectrum or a 3D [HCCH]-TOCSY spectrum recorded with a short mixing time, for combined inspection of the two spectra.

NOESY cross-peak assignment in 3D ^{15}N- or ^{13}C-resolved [^{1}H,^{1}H]-NOESY spectra

(i) As described for the [HCCH]-TOCSY spectrum, strips representing different individual protons are defined for each peak in the [^{1}H,^{1}H]-diagonal planes of the heteronuclear-resolved NOESY spectra.

(ii) One of the two correlated protons is given by the assignment of the ^{1}H-^{1}H diagonal peak of the strip in which the cross peak is observed. The second proton defines the frequency f used in Eq. 1 to sort the candidate strips (Fig. 6). In the example of Fig. 6, with $A_m = 10.0$ for the ^{1}H-^{1}H diagonal peaks and 1.0 otherwise, $\Delta = 0.03$ ppm, and choosing the frequencies f from normal distributions with standard deviation 0.02 ppm, the correct strip was attributed rank 1 in 31% of the cases, rank ≤ 5 in 82% and the worst rank was 21 (out of 418 defined strips).

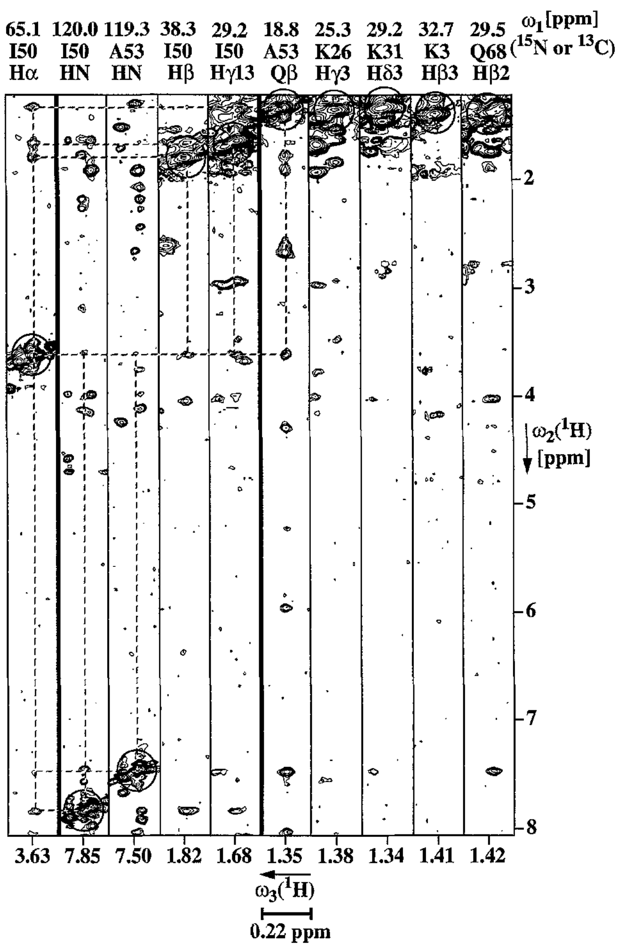


Fig. 6. Illustration of cross-peak assignments in 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra (Fesik and Zuiderweg, 1988; Ikura et al., 1989). The example is from the [^{15}N,^{13}C] doubly labeled N-terminal domain of the protein DnaJ (Szyperski et al, 1994b). Strips for individual protons are defined as in Fig. 5, where the strips 1 and 4–10 are from the ^{13}C-resolved [^{1}H,^{1}H]-NOESY spectrum, and strips 2 and 3 from ^{15}N-resolved [^{1}H,^{1}H]-NOESY. Diagonal peaks are encircled. The snapshot was taken after the assignment of the cross peaks between Ile^{50} H$^{\alpha}$ and Ile^{50} H^{N}, Ala^{53} H^{N}, Ile^{50} H$^{\beta}$ and Ile^{50} H$^{\gamma 13}$ was completed, and assignment of the cross peak at 1.35 ppm in the strip of Ile^{50} H$^{\alpha}$ was in progress. Cross peaks involving Ile^{50} H$^{\alpha}$ are connected to the corresponding diagonal peaks in the other strips by broken lines, i.e., the protons corresponding to the ^{1}H-^{1}H diagonal peaks in strips 2–5 define the assignment of the coupled hydrogen atom manifested in the cross peak in the strip of Ile^{50}. Strips 6–10 are candidates provided by XEASY for the assignment of the cross peak at 1.35 ppm. They are sorted by increasing distance of their diagonal peak (encircled) to this cross peak and the similarity of their peak patterns with the pattern observed in strip 1 (Eq. 1). Since only the strip of Ala^{53} Qβ lies exactly at 1.35 ppm and also shows a cross peak to Ile^{50} H$^{\alpha}$, visual inspection readily confirmed the validity of the automatic strip ranking by XEASY.

(iii) Two different situations may be encountered. In a 'complete' data set, which contains both the 'forward' and 'return' NOE connectivities, candidate strips containing a peak that correlates to the predetermined first proton are identified, and the cross peaks are assigned in the vertical dimension to the proton corresponding to the ^{1}H-^{1}H diagonal peak (Fig. 6). In the absence of 'return' NOEs to the first proton, for example because only a ^{15}N-labeled sample is available or the ^{13}C-resolved NOESY spectrum has been recorded in D_2O, one searches for strips that contain previously assigned cross peaks at the position of the second proton. With XEASY this can be done by defining strips for each cross peak that was picked in the spectrum, and then using Eq. 1 again to sort these strips.

Discussion and Conclusions

The large amount of data that needs to be processed for the NMR structure determination of a biological macromolecule warrants continued efforts to provide ever more efficient and reliable computer support. As an illustration, in our structure determination of the complex of cyclosporin A bound to its receptor cyclophilin, which has a total of 176 amino acid residues (Spitzfaden et al., 1994), about 2000 meaningful distance constraints were derived from more than 2700 NOESY cross peaks assigned in 2D [^{1}H,^{1}H]-NOESY, 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY, 3D ^{13}C-resolved [^{1}H,^{1}H]-NOESY and 4D [^{13}C, ^{13}C]-resolved [^{1}H,^{1}H]-NOESY. To arrive at these NOESY cross-peak assignments, a several-fold larger number of peaks from the different spectra was analyzed. Independent of the NMR techniques used, there is always a certain percentage of assignment decisions that are nontrivial due to degeneracies of two or several resonance frequencies, bleaching of signals that overlap with the water signal, limited signal-to-noise ratio, or artefacts such as baseline distortions or nonrandom noise. Furthermore, the spectral analysis is intrinsically complicated by the fact that the magnetization transfer pathways may depend on parameters that are not known at the start of a structure determination, such as short nonbonded interatomic distances or internal rate processes affecting relaxation rates. The program XEASY has been laid out both for handling of large data sizes and for dealing with uncertainties arising from experimental artefacts. At any stage of the spectral analysis the amount of data may be temporarily reduced to help the user concentrate on the relevant parts and to efficiently expand on already available assignments. The danger of incorrect assignments due to spectral artefacts is reduced by limiting automated searches to the identification of short lists of *likely* assignments, which are then further analyzed interactively. A series of projects with systems ranging in size from 40 to 200 residues and involving the interpretation of a variety of spectra (2D, 3D, 4D; homonuclear, heteronuclear; proteins, DNA) for purposes of spin system identification, sequence-specific assignment, NOESY cross-peak assignment and integration, or rate constant determination for amide proton exchange or other dynamic processes, has already confirmed the robustness and versatility of XEASY in practice.

Acknowledgements

Financial support was obtained from the Kommission zur Förderung der wissenschaftlichen Forschung in a joint project with Bruker-Spectrospin AG (project 2223.1), Ciba-Geigy AG and Silicon Graphics, Inc. We thank J.H. Bushweller, M. Pellecchia and T. Szyperski for the use of their spectra in Figs. 4–6, W. Antuch, M. Ottiger and T. Szyperski for constructive criticism, and R. Marani for the careful processing of the manuscript.

References

Bartels, C. and Wüthrich, K. (1994) *J. Biomol. NMR*, **4**, 775–785.

Bax, A., Clore, G.M. and Gronenborn, A.M. (1990) *J. Magn. Reson.*, **88**, 425–431.

Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.

Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.

Bushweller, J., Billeter, M., Holmgren, A. and Wüthrich, K. (1994) *J. Mol. Biol.*, **235**, 1585–1597.

Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.

Driscoll, P.C., Clore, G.M., Marion, D., Wingfield, P.T. and Gronenborn, A.M. (1990) *Biochemistry*, **29**, 3542–3556.

Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.

Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.

Fesik, S.W. and Zuiderweg, E.R.P. (1988) *J. Magn. Reson.*, **78**, 588–593.

Garret, D.S., Powers, R., Gronenborn, A.M. and Clore, G.M. (1991) *J. Magn. Reson.*, **95**, 214–220.

Glaser, S. and Kalbitzer, H.R. (1987) *J. Magn. Reson.*, **74**, 450–463.

Grahn, H., Delaglio, F., Delsuc, M.A. and Levy, G.C. (1988) *J. Magn. Reson.*, **77**, 294–307.

Gray, B.N. and Brown, L.R. (1991) *J. Magn. Reson.*, **95**, 320–340.

Gross, K.-H. and Kalbitzer, H.R. (1988) *J. Magn. Reson.*, **76**, 87–99.

Grzesiek, S. and Bax, A. (1992a) *J. Am. Chem. Soc.*, **114**, 6291–6293.

Grzesiek, S. and Bax, A. (1992b) *J. Magn. Reson.*, **99**, 201–207.

Güntert, P., Qian, Y.Q., Otting, G., Müller, M., Gehring, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 531–540.

Güntert, P., Dötsch, V., Wider, G. and Wüthrich, K. (1992) *J. Biomol. NMR*, **2**, 619–629.

Güntert, P., Berndt, K.D. and Wüthrich, K. (1993) *J. Biomol. NMR*, **3**, 601–606.

Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.

Heller, D. (1991) *Motif Programming Manual*, O'Reilly & Associates, Inc., Sebastopol, CA.

Hoch, J.C., Hengyi, S., Kjær, M., Ludvigsen, S. and Poulsen, F.M. (1987) *Carlsberg Res. Commun.*, **52**, 111–122.

Ikura, M., Kay, L.E., Tschudin, R. and Bax, A. (1989) *J. Magn. Reson.*, **86**, 204–209.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Jungnickel, D. (1990) *Graphen, Netzwerke und Algorithmen*, Bibliographisches Institut Wissenschaftsverlag, Mannheim.

Kleywegt, G.J., Lamerichs, R.M.J.N., Boelens, R. and Kaptein, R. (1989) *J. Magn. Reson.*, **85**, 186–197.

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 111–130.

Kleywegt, G.J., Vuister, G.W., Padilla, A., Knegtel, R.M.A., Boelens, R. and Kaptein, R. (1993) *J. Magn. Reson. Ser. B*, **102**, 166–176.

Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Neidig, K.P., Bodenmüller, H. and Kalbitzer, H.R. (1984) *Biochem. Biophys. Res. Commun.*, **125**, 1143–1150.

Novic, M., Oschkinat, H., Pfändler, P. and Bodenhausen, G. (1987) *J. Magn. Reson.*, **73**, 493–511.

O'Connell, J.F., Bender, R., Engels, J.W., Koller, K.P., Scharf, M. and Wüthrich, K. (1994) *Eur. J. Biochem.*, **220**, 763–770.

Pfändler, P., Bodenhausen, G., Meier, B.U. and Ernst, R.R. (1985) *Anal. Chem.*, **57**, 2510–2516.

Pfändler, P. and Bodenhausen, G. (1990) *J. Magn. Reson.*, **87**, 26–45.

Richarz, R. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2133–2141.

Scheifer, R.W., Gettys, J. and Newman, R. (1988) *X Window System C Library and Protocol Reference*, Digital Press, Bedford, MA.

Simorre, J.P., Brutscher, B., Caffrey, M.S. and Marion, D. (1994) *J. Biomol. NMR*, **4**, 325–333.

Spitzfaden, C., Braun, W., Wider, G., Widmer, H. and Wüthrich, K. (1994) *J. Biomol. NMR*, **4**, 463–482.

States, D.J., Haberkorn, R.A. and Ruben, D.J. (1982) *J. Magn. Reson.*, **48**, 286–292.

Stoven, V., Mikou, A., Piveteau, D., Guittet, E. and Lallemand, J.Y. (1989) *J. Magn. Reson.*, **82**, 163–168.

Szyperski, T., Wider, G., Bushweller, J.H. and Wüthrich, K. (1993a) *J. Biomol. NMR*, **3**, 127–132.

Szyperski, T., Wider, G., Bushweller, J.H. and Wüthrich, K. (1993b) *J. Am. Chem. Soc.*, **115**, 9307–9308.

Szyperski, T., Pellecchia, M. and Wüthrich, K. (1994a) *J. Magn. Reson. Ser. B*, **105**, 188–191.

Szyperski, T., Pellecchia, M., Wall, D., Georgopoulos, C. and Wüthrich, K. (1994b) *Proc. Natl. Acad. Sci. USA*, **91**, 11343–11347.

Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.

Weber, P.L., Malikayil, J.A. and Müller, L. (1989) *J. Magn. Reson.*, **82**, 419–426.

Wehrens, R., Lucasius, C., Buydens, L. and Kateman, G. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 245–251.

Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Wüthrich, K., Spitzfaden, C., Memmert, K., Widmer, H. and Wider, G. (1991) *FEBS Lett.*, **285**, 237–247.