Journal of Biomolecular NMR, 1 (1991) 111–130
ESCOM

J-Bio NMR 017

Efficient analysis of protein 2D NMR spectra using the software package *EASY*

Craig Eccles, Peter Güntert, Martin Billeter and Kurt Wüthrich

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

Received 25 January 1991
Accepted 11 March 1991

Keywords: Nuclear magnetic resonance; Protein structure determination; Automated spectral analysis; Software package *EASY*; Sequence-specific assignments

SUMMARY

The program *EASY* supports the spectral analysis of biomacromolecular two-dimensional (2D) nuclear magnetic resonance (NMR) data. It provides a user-friendly, window-based environment in which to view spectra for interactive interpretation. In addition, it includes a number of automated routines for peak-picking, spin-system identification, sequential resonance assignment in polypeptide chains, and cross peak integration. In this uniform environment, all resulting parameter lists can be recorded on disk, so that the paper plots and handwritten notes which normally accompany manual assignment of spectra can be largely eliminated. For example, in a protein structure determination by 2D ^{1}H NMR, *EASY* accepts the frequency domain datasets as input, and after combined use of the automated and interactive routines it can yield a listing of conformational constraints in the format required as input for the calculation of the 3D structure. The program was extensively tested with current protein structure determinations in our laboratory. In this paper, its main features are illustrated with data on the protein basic pancreatic trypsin inhibitor.

INTRODUCTION

The use of nuclear magnetic resonance (NMR) spectroscopy for the determination of biomacromolecular structures in solution (Wüthrich, 1986) has by now been well established. Current research emphasizes advanced techniques of isotope labeling and NMR spectroscopy for use with larger, more complex systems (e.g., Kay et al., 1990; Otting et al., 1990), but it can be foreseen that much interest in the NMR method will continue to be focussed on structure determinations with polypeptides and small proteins in the size range up to molecular weights of about 10 000. Although such structure determinations are in no way routine projects, much of the work can be done with standard techniques, for example, using homonuclear 2D ^{1}H NMR, sequential ^{1}H NMR assignments (Billeter et al., 1982; Wagner and Wüthrich, 1982; Wider et al., 1982), identifi-

0925-2738/$ 5.00

cation of regular secondary structures by pattern recognition in the ^{1}H NMR data (Pardi et al., 1984; Wüthrich et al., 1984), and calculations of the three-dimensional (3D) structure with a combination of distance geometry, molecular dynamics and energy refinement (Braun, 1987; Clore and Gronenborn, 1989; Wüthrich, 1986, 1989). Therefore, an integrated computer-supported procedure for all steps of the spectral analysis up to the preparation of the input for the structure calculation is a promising and attractive avenue toward further improved efficiency of the determination of NMR structures of proteins. In the last few years, considerable effort has already been devoted to such projects. Different groups have developed routines for peak-picking (e.g., Pfändler et al., 1985; Glaser and Kalbitzer, 1987; Hoch et al., 1987; Meier et al., 1987), spin-system identification (e.g., Neidig et al., 1984; Pfändler et al., 1985; Oh et al., 1988; Eads and Kuntz, 1989; Kraulis, 1989; Kleywegt et al., 1989; Weber et al., 1989; Van de Ven, 1990), sequential resonance assignments (e.g., Billeter et al., 1988; Cieslar et al., 1988; Eads and Kuntz, 1989; Kleywegt et al., 1989; Van de Ven, 1990), and cross peak integration (e.g., Denk et al., 1986; Holak et al., 1987; Stoven et al., 1989; Kjær et al., 1990). This paper presents a new program package, *EASY* (for *E*TH *A*utomated *S*pectroscop*y*), which allows the spectroscopist to perform all these functions in an integrated environment.

The interactive parts of *EASY* are quite general for use with 2D NMR spectra, but the program has been designed primarily for the purpose of determining 3D structures of small proteins. Apart from the acronym, it was felt from the beginning that the program should be *easy* to use, and require a minimum of training. A user-friendly and consistent interface, on-line help and close to real-time performance for most functions were felt to be important if the program was to replace traditional methods of analysis. The program allows the user to interactively perform all steps involved in the preparation of the NMR input for the structure calculations (Wüthrich, 1986). In addition, a number of routines are provided which are 'automated' in the sense that the computer performs a systematic search of a dataset defined by the operator, and during this search makes decisions independent of the operator.

METHODS

This section provides a survey of the strategy followed for protein structure determinations when using the program *EASY*. It further gives a brief introduction to the functions implemented in the program for use in the different steps of this procedure. A more detailed, illustrated description of the ways in which these steps are executed is given in the *Results* section.

Protein structure determination using EASY

The flowchart in Fig. 1 foresees that a minimum of three spectral types are acquired, i.e., 2QF-COSY (two-dimensional two-quantum-filtered correlation spectroscopy) and TOCSY (two-dimensional total correlation spectroscopy) for delineation of spin systems, and NOESY (two-dimensional nuclear Overhauser enhancement spectroscopy) for sequential resonance assignments and collection of inter-proton distance constraints (Wüthrich, 1986, 1989). Additional information from other experiments, such as two-quantum spectroscopy, may be added to resolve difficult assignment problems. To enable efficient use of *EASY*, all spectra used in a structure determination should be obtained under the same conditions of pH and temperature to minimize chemical shift

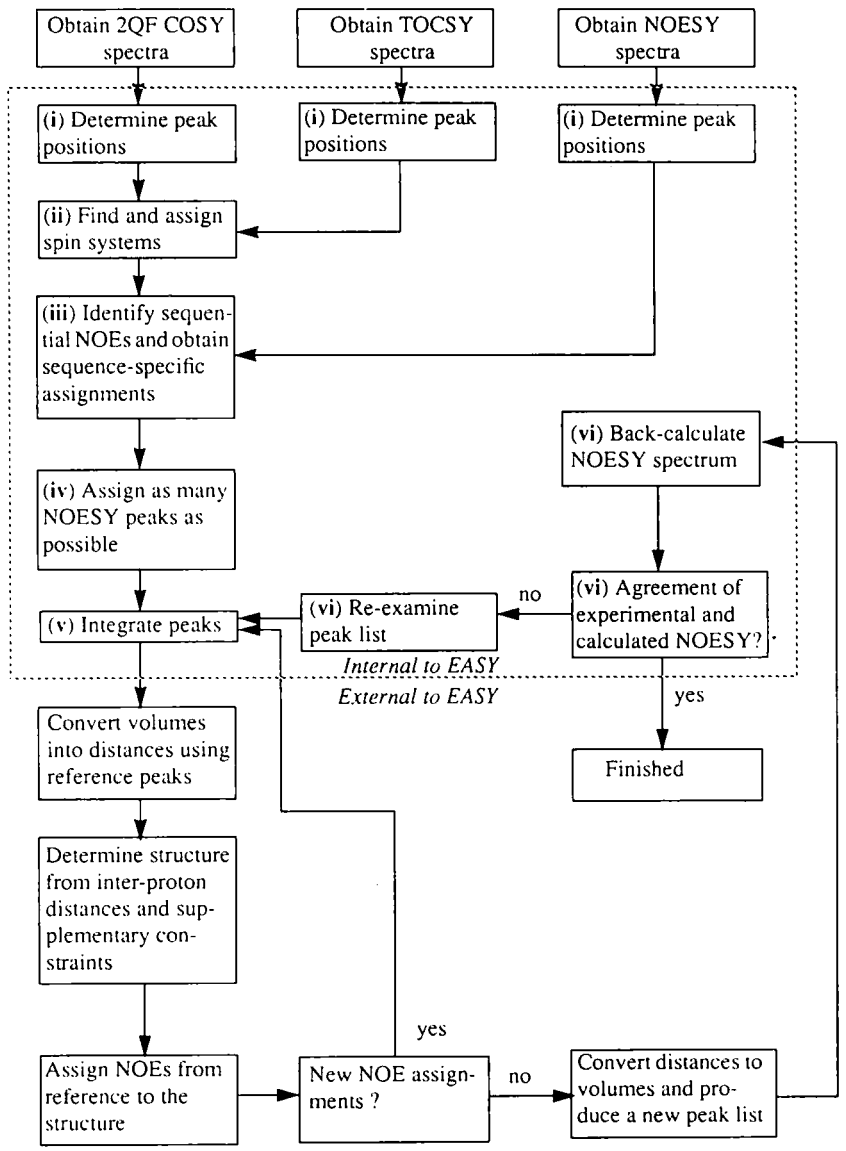

Fig. 1. Outline of a protein structure determination using the *EASY* program. The steps supported by *EASY*, numbered (i) - (vi), are further described in the *Methods* section.

differences. *EASY* accepts as input 2D NMR spectra *after* Fourier transformation, and the following steps are then performed.

(i) The positions of cross peaks are recorded. This abstraction of the datasets is advantageous, as it is much more compact than the original spectral format, it is more readily manipulated by the computer, and it provides a convenient form of bookkeeping.

(ii) ^{1}H spin systems of the individual amino-acid residues are identified. As a practical rule, 2QF-COSY and TOCSY can establish relations between different protons that are separated by,

at most, three chemical bonds, or by multiple steps of, at most, three bonds each, respectively. Each amino-acid residue is thus represented by a single spin system, the only exceptions being the aromatic side chains, asparagine and glutamine, which contain two separate spin systems that must be connected with the use of NOE (nuclear Overhauser effect) relations (Wüthrich, 1986). In the 2QF-COSY spectrum, a stepwise connectivity pathway may be followed from the amide proton to the C^{α} proton and onward into the side chain, whereas in the TOCSY spectrum all relations between protons belonging to the same spin system can be simultaneously represented. In any but the smallest proteins, the number of spin systems will be such that mutual overlap is common and unambiguous delineation of the connectivity pathways is difficult. Using TOCSY in conjunction with the 2QF-COSY data, one can usually considerably limit the number of ambiguous pathways. Once the spin systems have thus been disentangled, *EASY* assigns them to one of several spin-system categories based on the 2QF-COSY and TOCSY connectivity patterns and the chemical shifts of key protons (Wüthrich, 1986).

(iii) In the third step, the individual spin systems are assigned to particular amino-acid residues in the amino-acid sequence. Using the sequential NOEs (Billeter et al., 1982; Wüthrich, 1986), polypeptide segments of variable length are identified. This step is particularly suited to computer analysis, because of the large number of possible connectivity pathways which are present in a typical NOESY spectrum. Sequence-specific assignments are obtained by matching these polypeptide segments against the independently known amino-acid sequence of the protein. Once each spin systemhas been assigned to a specific residue, the assignment information is condensed into a proton list which contains the chemical shift of each assigned proton or methyl group in the protein.

(iv) Based on the aforementioned proton list, each cross peak in the NOESY spectrum can be assigned to a pair of interacting protons. Well-separated peaks are thus unambiguously assigned, whereas for peaks in crowded spectral regions *EASY* may produce a list of two or more possible assignments.

(v) Those NOESY cross peaks which are unambiguously assigned at this stage (typically 30–50% of the total number) are integrated and corresponding distance constraints inferred.

(vi) The data from (iv) and (v) are used as input for initial structure calculations. By reference to these preliminary structures, one can often eliminate a majority of the previously listed multiple possible assignments, since NOEs between protons separated by more than about 5Å in these structures are not usually visible (Wüthrich, 1986). The structure calculation is then repeated with the thus extended distance constraint set, and this procedure can be repeated several times. As a final check, *EASY* allows one to back-calculate from the structure a NOESY spectrum which can then be compared with the original dataset.

Figure 1 illustrates that *EASY* accepts as input the 2D NMR spectra after Fourier transformation, executes the steps (i) – (v), and can further support the assignment of additional NOEs in step (vi). *EASY* has been designed to fit in with existing software in our laboratory so that all relevant data can be passed on with minimal user intervention. Within *EASY*, a complete spectral analysis can be performed interactively, but in addition, automated routines are provided for individual steps. In practice, one makes use of the automated steps wherever possible, and the results are then checked and supplemented using the interactive tools provided.

Functions implemented in the program EASY

The program *EASY* has been installed on *Sun* workstations (Sun3, Sun4, SPARCstation) using the UNIX operating system SunOS level 4.0 or later. A minimum of 8 Mb of memory is required. It is written in the C programming language, and operates within the *Sunview* window environment using resizable windows and pop-up menus. The program consists of approximately 150 commands which are organized into eight menus. These menus can be displayed and commands selected either using a mouse or, for the more experienced user, by typing 2-letter codes. The following is a brief summary of the contents of each menu.

Spectra manipulation menu. 2D NMR spectra may be loaded, calibrated, clipped (to save memory space), generated from a peak list and line shape data, and the difference between two spectra may be calculated.

Peak analysis menu. Cross-peak positions may be recorded using automated or manual peak-picking routines. Assignment data may be added to the list of cross peak positions. Peak lists may be sorted by chemical shift, or have their coordinates transposed. Lists of expected peaks may also be generated from assignment data.

Spin system-analysis menu. Spin systems may be identified automatically using TOCSY and 2QF-COSY peak lists. Peaks may be manually added or removed from a spin system, and statistics about the results of the spin-system identification may be listed.

Sequential assignment menu. Intraresidual NOEs may be identified in NOESY peak lists using the data obtained in the spin system-analysis menu. NOE connectivity lists can be generated and edited. These can then be used to identify all sequential assignment pathways that cannot be excluded by reference to the amino-acid sequence. The final sequence-specific assignments are then obtained interactively using this data.

Proton list menu. Proton lists without chemical shifts can be generated from the amino-acid sequence. The data on the sequence-specific assignments of the spin systems can then be used to fill in the chemical-shift data.

Assign peaks menu. Cross peaks in a 2D NMR spectrum, in particular NOESY, can be assigned with an automated routine from the proton list that was generated in the proton list menu using the chemical-shift data.

Peak integration menu. Peak volumes can be calculated using a variety of different interactive and automated procedures. Overlapping cross peaks can be integrated using a line-shape-fitting algorithm (Denk et al., 1986).

Sequence menu. The protein sequence can be loaded, and statistical information about the number of residues and spin system-types can be listed.

In addition, each menu contains the following utility commands:

Window functions. A cross-section window is provided for simultaneous display of up to ten rows or columns from the spectrum. Rows and columns can be added, subtracted, and saved to files. An overview window displays the complete spectrum and outlines those regions that are currently being viewed. Within this window, new regions can be selected and previous ones restored. While contour plots are the conventional way of representing a spectrum, the *EASY* program relies mostly on a display which represents intensity using a color scale. A separate window is provided which allows real-time adjustment of the color intensity and the color threshold (the level below which black is displayed). A text window is provided to keep record of previously entered

commands and to display data. It also provides a number of facilities for text editing, printing and general data manipulation. There is also a special window for rapid modification of frequently changed parameters.

Zoom functions. A variety of functions are provided for viewing the spectrum in detail. Commands include the zooming of single or multiple regions, display of regions related by symmetry about the diagonal, shifting and enlarging of the currently viewed region, restoration of previously viewed regions, and save-to-disk of regions of interest. A number of regions from different parts of the same spectrum or from different spectra may be viewed simultaneously.

Drawing functions. The spectrum may be overlaid with rectangles and lines for the purpose of checking peak alignments. Contours and stacked plots may be rapidly drawn over a spectrum, e.g., a NOESY spectrum may be contoured over a 2QF-COSY intensity display.

Miscellaneous functions. These include spectral scaling, display of data-point intensities, noise-level measurement, display of a ppm grid, and the preparation of contour and stacked plot data for hardcopy printing.

RESULTS

This section describes the salient features of the program *EASY* in more detail, using examples taken from work on a structure refinement of the protein BPTI (basic pancreatic trypsin inhibitor) (Güntert, P., Orbons, L., Berndt, K. and Wüthrich K., to be published) as illustrations. BPTI is a small globular protein of 58 amino-acid residues, of which the ^{1}H NMR assignments and the 3D structure are well known (Deisenhofer and Steigemann, 1975; Wagner and Wüthrich, 1982; Wagner et al., 1987; Wlodawer et al., 1984, 1987).

The 2D ^{1}H NMR spectra are first converted into the file format required by the *EASY* program. This produces a logarithmic 8-bit file for display purposes, and a floating-point 16-bit file for numerical analyses. Both files are in serial rather than submatrix form. A transposed version of the 16-bit file can also be generated to provide rapid access to column data. The *EASY* program can operate without the 16-bit file, but numerical accuracy and plot quality are improved if it is available.

Peak list and proton list

In addition to the chemical-shift coordinates, the peak list can store other information. In a typical entry (1) produced while searching for spin systems

```
219    2.694    7.122    1    C    0.000e + 00    0.000e + 00  -  0  0
6  6  2  1                                                              (1)
41  41  7  9
```

the numbers across the top are, from left to right, the peak number, the chemical shift coordinates of the peak in ω_1 and ω_2, the peak color code (a number from 1 to 6) and the type of spectrum where the peak was observed (C, T, N for COSY or 2QF-COSY, TOCSY, NOESY). The remaining entries on the top line will be described below. The numbers on the next two lines identify the

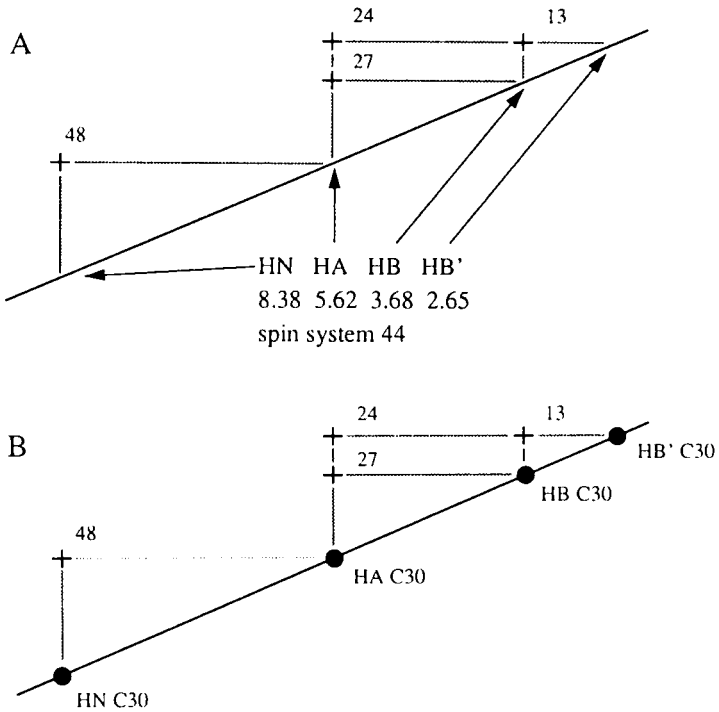

Fig. 2. The data structures used within *EASY*. (A) Relationship between the *peak* list (marked with crosses) and the *spin-system* list, which is made up of groups of protons that are connected by spin-spin coupling (indicated with horizontal and vertical lines). Four numbers are required to define a cross peak assignment. For example, peak 48 is attributed to spin system 44 in both ω_1 and ω_2, and within this spin system it is assigned to HN in ω_2 and HA in ω_1. The spin system 44 consists of protons at the following chemical shifts: 8.38 (HN); 5.62 (HA); 3.68 (HB); and 2.65 ppm (HB'). (B) Relationship between the *peak* list and the *proton* list (filled circles).

spin systems and the location of the protons within these spin systems (see Fig. 2 for the relations between the various lists used by *EASY*). The cross peak listed in (1) may therefore be assigned to protons 1 and 2 in spin system 6, or to protons 7 and 9 in spin system 41, or to both. Assigning a peak to more than one spin system allows analysis of overlapped peaks in crowded spectral regions, where several peaks with the same chemical shifts may have been picked only once. More details on the spin-system list will be given in the section on spin-system identification, in particular Eq. (6).

Once spin-system identification and sequence-specific assignment have been completed, it is more convenient to work with the more compact proton list in the place of the spin-system list (Fig. 2B). A typical peak list entry then has the form (2):

736 1.574 8.552 1 N 1.276e+04 4.500e+00 d 515 73 (2)

The first five entries are as described for (1). These are followed by two floating-point numbers representing the peak volume and its uncertainty in percent. Next, the method of integration is indicated by a lower-case letter (d stands for the method of Denk et al., 1986). The final two numbers

are ω_1 and ω_2 pointers to entries in the proton list (3), i.e., the cross peak 736 is assigned to proton 515 in ω_1 and proton 73 in ω_2.

In a typical entry in a proton list (3)

29 7.360 0.003 HD1 57 (3)

the first number is used to refer to a particular proton. The second field contains the chemical shift of this proton, and the third field gives the range of chemical shifts for peaks which have been assigned to this proton. The final two fields define the position of the hydrogen atom in the amino-acid residue and the residue number. The residue name can then be found by reference to the sequence. The proton list is generated automatically from the amino-acid sequence, and the chemical shifts can be copied over from the newly assigned spin-system list, or from published listings of the chemical shifts.

Peak-picking

In each spectrum, the peak positions should be obtained as accurately as possible to enable the automated assignment routines to function efficiently. Interactive and automated peak-picking modes are provided. In the interactive mode, the user moves the screen cursor over the center of the peak and selects the center with a mouse button. The peak position is then entered into the peak list. Additional information about the peak may be entered manually and displayed next to the peak on the screen.

Automated anti-phase peak-picking

Anti-phase cross peaks, such as those found in 2QF-COSY spectra, exhibit a symmetry about their center which can be detected by a simple algorithm (Meier et al., 1987). This algorithm relies on a symmetry function that is equal to the average peak height at the position of a signal with the expected symmetry, and is much smaller at other locations in the 2D spectral plane. Peak centers are then identified as those local maxima of this symmetry function which exceed a user-specified threshold. Once a peak has been identified with this symmetry criterion, the position of its center is determined more accurately by a search for the crossing point of horizontal and vertical lines of zero intensity in the vicinity of the symmetry maximum.

Automated anti-phase peak-picking was tested with a 2QF-COSY-spectrum of BPTI in H_2O. The transformed data had 2k × 1k points. A total of 1246 peaks were found, of which 957 could be accurately positioned using the above procedure. Because many of the 1246 peaks were found in the t_1-noise band at the water frequency, the peak list was checked for symmetry about the diagonal. Peaks were considered symmetrically picked if their transposed coordinates aligned to within ± 0.02 ppm. The 413 pairs of symmetry-related peaks were saved as a new peak list. The asymmetric peaks were interactively examined and, if appropriate, made symmetric and added back to the list of symmetric peaks. Otherwise they were dropped. Following this step and removal of peaks located within ± 0.02 ppm of the diagonal, the new peak list contained a total of 768 peaks (see Fig. 3 for an illustration). This result of automated peak-picking must be interactively checked because it only recognizes AX-type peak shapes. More complex fine structures arising from large passive coupling may therefore be picked as more than one peak. In the case of BPTI, very few peaks had to be added, and the peak list after manual checking contained 714 peaks.

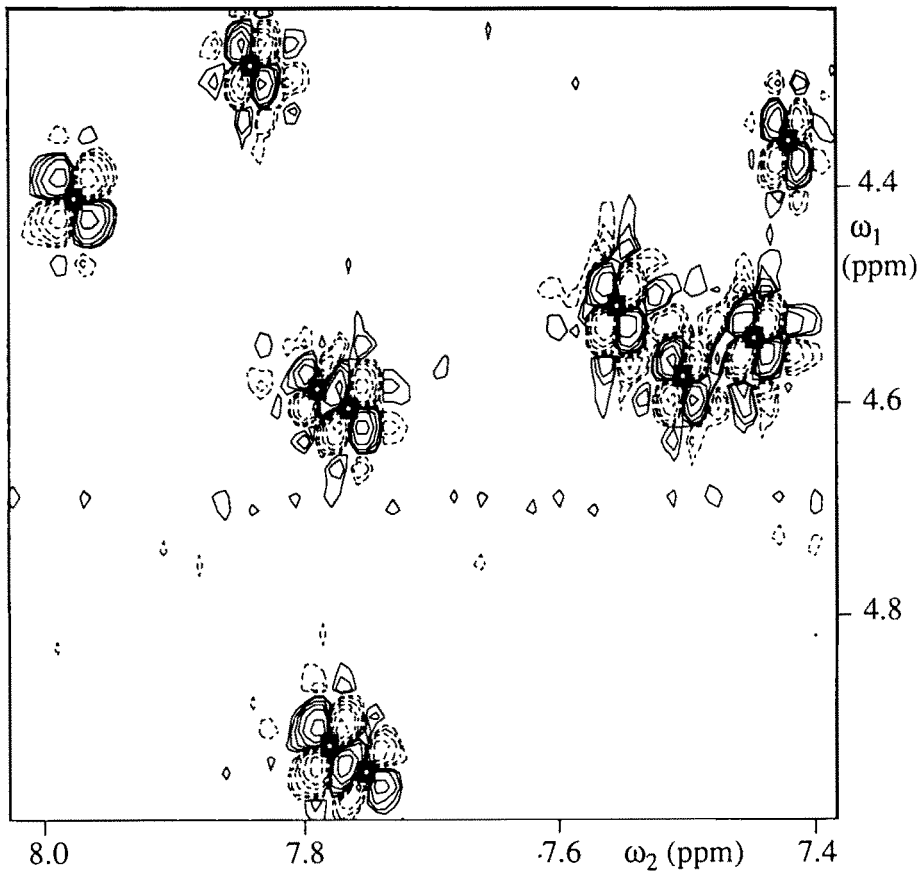

Fig. 3. Result of automated symmetric picking of anti-phase peaks in a 600 MHz 2QF-COSY spectrum of BPTI. The peak centers are identified by black squares.

Automated picking of in-phase peaks

The routine used by *EASY* for automated picking of in-phase peaks, such as those found in TOCSY and NOESY spectra, first searches for maxima in the spectrum which are above a predetermined threshold (usually between one and two times the noise level). The peaks thus found are then checked to see if their line width at the threshold is between a predefined minimum and a preset maximum value. The first check allows the routine to ignore noise spikes, while the second check provides some tolerance against poor base line correction or against artifacts such as those produced by the water signal. For spectra where the resolution is sufficient to resolve multiplet structure, another routine may be called which searches for peaks that are a certain distance apart in ω_2 but have similar locations in ω_1. These peaks will then be merged into a single peak at the average ω_2 shift. As an additional step, the program can shift the recorded peak positions to the center of mass of the corresponding cross peaks. This will only be done for peaks that are well separated from other peaks.

Automated picking of in-phase peaks was tested with a TOCSY spectrum of BPTI in H_2O. The complete transformed dataset had 2k × 1k points. A total of 257 peaks were picked in the region

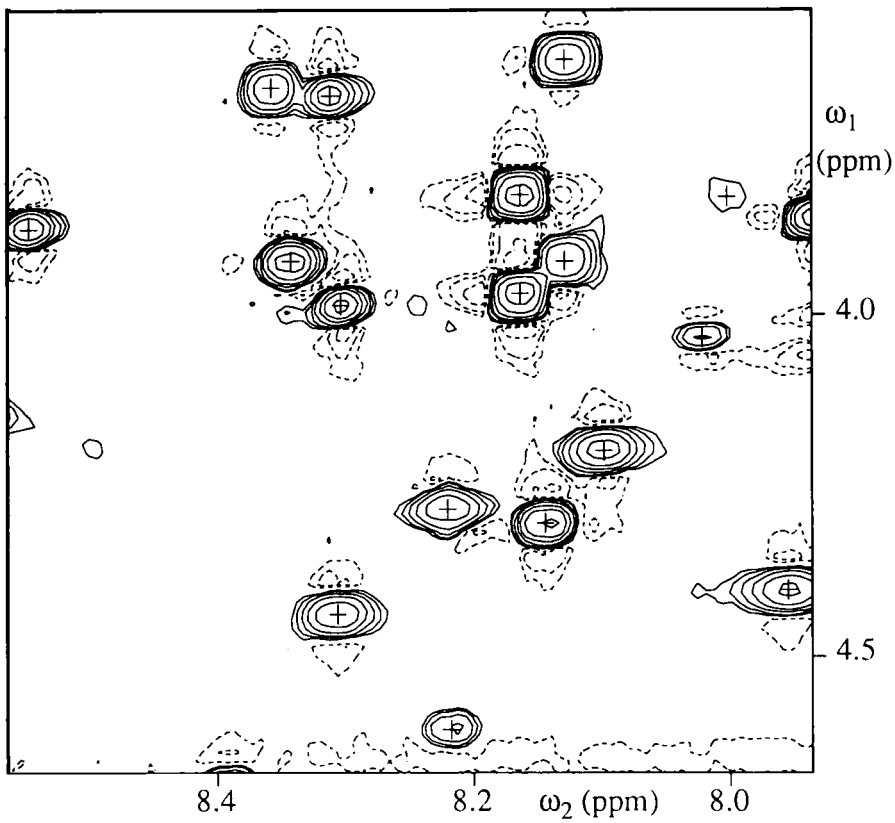

Fig. 4. Result of an automated picking of in-phase peaks in a 600 MHz TOCSY spectrum of BPTI in H_2O. The peak centers are indicated by crosses. They have been adjusted using the center-of-mass algorithm mentioned in the text.

($\omega_1 = 0.1$-5.8 ppm, $\omega_2 = 6.6$-10.7 ppm). Multiplet peaks were automatically collapsed, leaving 235 peaks. Thirteen very weak peaks were then added interactively to give a total of 248. The aforementioned center of mass adjustment was then applied to this dataset (Fig. 4).

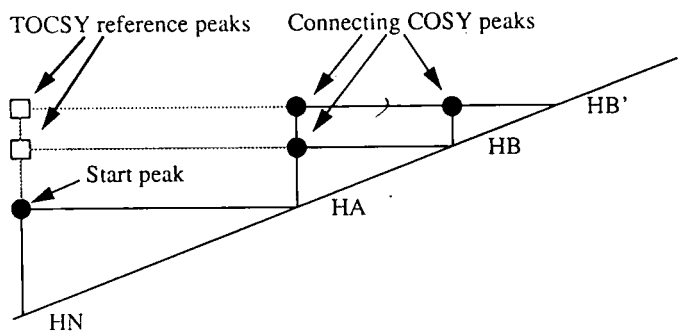

Fig. 5. Schematic illustration of automated spin-system identification by *EASY*. Filled circles represent 2QF-COSY cross peaks, squares are TOCSY cross peaks. First, the HN-HA cross peaks are identified in the 2QF-COSY and TOCSY peak lists. 'TOCSY reference peaks' are then all those at the HN shift along ω_2. Only 'connecting 2QF-COSY peaks' are then added to the spin system, which are all those that have a TOCSY reference peak at the same ω_1 chemical shift.

Spin-system identification

The strategy used by *EASY* for automated spin-system identification has many traits in common with the procedure described recently by Van de Ven (1990). The starting points are the amide-proton/C^{α}-proton COSY cross peaks, which are normally well separated from other types of cross peaks and show little overlap in small proteins. The corresponding TOCSY cross peaks can then be identified, and the search in ω_1 continued to identify TOCSY cross peaks relating the amide proton with protons in the amino-acid side chain (Fig. 5). Because the alignment between COSY and TOCSY lists is often quite poor (deviations of ± 0.03 ppm are not uncommon), uncertainties need to be considered when the search for connectivities is made.

Once the COSY and TOCSY connectivities have been established, the resulting groups of coupled spins are checked against a database to determine the spin-system type. This database includes the 20 common amino-acid residues. However, the results of the initial spectral analysis quite often represent only partial fragments of the amino-acid spin systems, which enables a distinction between different classes of residues but not an identification of individual amino-acid types. Therefore, the database also includes several classes of amino-acid residues, for example those with long side chains, or those containing a $>C^{\alpha}H\text{-}C^{\beta}H_2$- fragment giving rise to an AMX spin system (p. 133 in Wüthrich, 1986). In the present implementation of *EASY*, the following classes of residues are included in the database:

lsc	L, M, E, Q, K, R	('long side chains')
amx	C, D, N, S, F, Y, W, H	('AMX spin systems')
arm	F, Y, W, H	('aromatics')
CDN	C, D, N	
MEQ	M, E, Q	
K_R	K, R	
V_I	V, I	
lng	V, T, L, I, M, E, Q, K, R	('long')

(4)

A series of criteria are encoded in the database file which may be used to obtain a score from 0 to 10 for each of the spin-system types. As an illustration, (5) shows the database entry for the residue class *lsc*:

```
lsc  if number HA = 1, score contribution +20
     if number HA > 1, score contribution −30
     if number HB = 2, score contribution +20
     if number HG > 0, score contribution +20
     if any of HD to HZ present, score contribution +20
     if 2QF-COSY peak HB-HB exists, score contribution +10
     if 2QF-COSY peak HG-HG exists, score contribution +10
     if 0.3 < δ(HB) <2.7, score contribution +10
```

(5)

Similar entries can be made by the user for each spin-system type listed in (6). The exact score con-

tributions are somewhat arbitrary, and may be selected by the user. The program normalizes the score contributions so that the maximum score value that can be obtained is 10. The program tests for the number of proton types, whether certain protons are present, whether certain cross peaks exist, and whether certain characteristic chemical-shift ranges are met. The score is initially zero and is incremented by the specified score contribution if the spin system satisfies the particular criterion. Once all spin-system types have been considered, the results are written into the spin-system list. If a single spin-system type has a maximum normalized score greater than 5, it is written into the *residue class* field of the spin-system list. A typical entry into the spin-system list is shown in (6), where the attribution of proton names in the top row is mainly based on the COSY connectivities observed.

Spin system No. 19

proton:	HN	HA	HB	HB	HG	HG	HD
shift:	8.39	4.72	1.80	0.83	1.75	1.35	3.48

scores:					
Ala = 0	Arg = 0	Asn = 0	Asp = 0	Cys = 0	
Gln = 0	Glu = 0	Gly = 0	His = 0	Ile = 6	(6)
Leu = 0	Lys = 0	Met = 0	Phe = 0	Pro = 0	
Ser = 4	Thr = 0	Trp = 0	Tyr = 0	Val = 3	
arm = 0	CDN = 0	*lsc* = 9	*amx* = 4	MEQ = 6	
K_R = 0	V_I = 0	*lng* = 0			

residue class: *lsc*

Only a limited number of the spin-system types listed in (6) are included in the automated search (the other scores are set to zero) depending on the quality of the experimental data (pp. 130–135, in Wüthrich, 1986). Thus, because the spin-system search starts from HN-HA cross peaks, prolines are not identified and so there is no point in including Pro in the search, and for aromatic side chains the automated spin-system identification ends at $C^{\beta}H_2$ and so there is no point in searching separately for *arm* and CDN. To provide a practical example, for an automated spin-system search in BPTI the 9 species Gly, Ala, Val, Ile, Thr, Ser, K_R, *amx*, and *lsc* out of the total of 28 entries in the database (6) were used. Information on the remaining 19 species may subsequently be added manually, using results obtained from interactive complementation of the automated spin-system identifications.

The following experience was gained when using *EASY* for spin-system identification in the 2QF-COSY and TOCSY spectra of BPTI. The 58-residue amino-acid sequence of BPTI (7)

```
1                   10                  20                  30
R P D F C L E P P Y T G P C K A R I I R Y F Y N A K A G L C
                    40                  50                  58
Q T F V Y G G C R A K R N N F K S A E D C M R T C G G A
```
(7)

contains 53 spin systems that include an amide proton. Using the automatically picked and checked peak lists from the 2QF-COSY and TOCSY spectra, the automated spin-system search

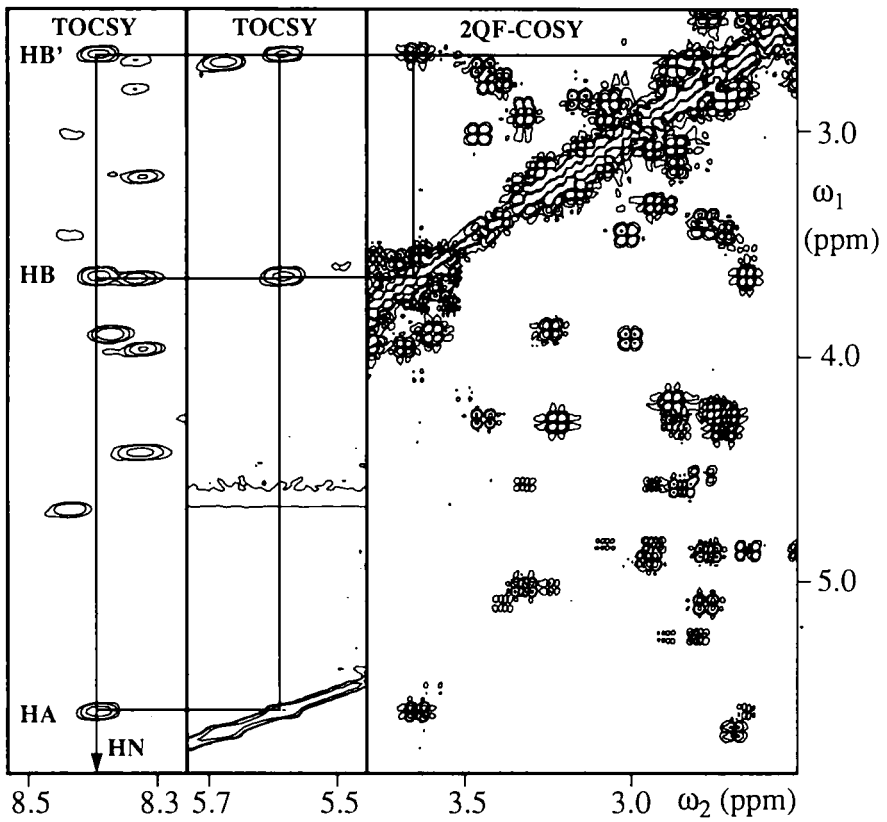

Fig. 6. Display of TOCSY and 2QF-COSY spectral regions for interactive checks on spin-system identifications. The spin system is AMPX as evidenced by the chemical shifts of the C^{β} protons, the presence of only three TOCSY peaks at ω_2(HN), and the presence of two HA-HB TOCSY cross peaks and one HB-HB' COSY cross peak. The two TOCSY peaks at ω_2(HA) in the centre strip verify that the two corresponding cross peaks in the HN region on the left belong to the same spin system.

on the level of the aforementioned 9 species yielded 27 spin-system identifications that were found to be correct by comparison with Table 1 in Wagner et al. (1987). These included 4 Gly, 3 Ala, Val 34, 2 Thr, Ser 47, 6 *amx*, 4 *lsc*, and 6 K_R (note that Ser and K_R are also included in the total number of *amx* systems and *lsc* systems, respectively). In addition, 15 spin-system assignments were incorrect, and the assignments for 23 spin systems remained ambiguous. The errors and ambiguities arose primarily from the presence of artifactual peaks in the HN-HA region, misinterpretation of resonance lines from side-chain amino or guanidinium protons of Lys and Arg as backbone amide proton signals, absence of HA-HB cross peaks in the 2QF-COSY spectrum, or chemical-shift degeneracy of β-methylene protons.

To improve the results of the automated search, a variety of commands and display modes are included for interactive editing of the spin-system list. For example, a multi-zoom command displays multiple spectral regions of interest simultaneously (Fig. 6), and the spectra can be overlaid with lines for visualizing spin-system identifications. Two special routines are contained in *EASY* to support the interactive work following the automated spin-system identification. One of these

is used with the TOCSY spectra recorded in H_2O, and searches for spin systems with different ω_2 chemical shifts in the HN region but otherwise identical shifts. Each pair of such spin systems would then typically be identified as one residue of Lys or Arg. The second routine identifies the aromatic side chains using a NOESY spectrum in D_2O (p. 121 in Wüthrich, 1986), and can also be applied for a search of Asn spin systems in a NOESY spectrum recorded in H_2O.

Overall, before starting the sequential assignments, all amide-proton-containing spin systems expected from the amino-acid sequence of BPTI (7) were identified on the level of the 11 species Gly, Ala, Val, Ile, Thr, Ser, MEQ, K_R, *lsc*, *arm* and CDN.

Sequence-specific resonance assignments

Obtaining sequence-specific assignments of amino-acid spin systems may be viewed as a two-step process (Wagner and Wüthrich, 1982; Wüthrich, 1983, 1986), in which polypeptide segments of variable length are assembled based on the observation of sequential NOEs, and sequence-specific assignments are then achieved by matching the sequence of these segments against the independently known amino-acid sequence. *EASY* includes facilities for interactively establishing sequential assignment pathways, as well as a routine for automated sequential assignments which largely combines the two aforementioned steps by making reference to the amino-acid sequence at each sequential addition of a spin system. In this automated approach, one makes use of the fact that all sequential NOEs $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ (for a definition, see pp. 117–118 of Wüthrich, 1986) involve at least one amide proton, the only exception being those across Xxx-Pro bonds (which are not considered in the automated approach). In the NOESY peak list, the intraresidual NOEs are first identified by reference to the previously compiled spin-system list (6), starting with the HN-HA cross peaks. An uncertainty of typically ± 0.02 ppm is allowed between the chemical shifts in the spin-system list and the NOESY data. Missing intramolecular HN-HA NOE cross peaks are reported to the user, and their locations have to be determined manually. Further intraresidual NOEs with HN may be assigned using a smaller uncertainty in ω_2 of typically ± 0.01 ppm, since the HN-HA cross peaks can be used as a reference.

Once the intraresidual NOESY cross peaks have been assigned, the program searches the NOESY spectrum at each HN chemical shift in the ω_1-direction for interresidual connectivities. Again, an uncertainty limit of about ± 0.01 ppm must be provided to allow for inaccurate peak-picking and imperfect calibration. This automated search yields a list of NOE connectivities from the amide proton of each spin system to protons in other spin systems, as is illustrated in (8) for the spin system 17 in the analysis of the BPTI spectra.)

Spin system Nr. 17

17 MEQ HN	– 4 CDN	HA	
	– 14 CDN	HA	
	– 19 *lsc*	HA	
	– 19 *lsc*	HB	(8)
	– 19 *lsc*	HB	
	– 30 *lsc*	HN	
	– 58 K_R	HD	

The automated sequential assignment routine then uses the NOEs between HN and either HN, HA or HB, and the amino-acid sequence to find an uninterrupted pathway through the NOE data along the sequence. In principle, such pathways exist for all polypeptide segments that are bounded by prolyl residues, e.g., in BPTI for the segments 3–7, 10–12 and 14–58 (7). With spin system 17 (which is Met, Glu or Gln) at the origin of a pathway (8), the automated search can be started at the sequence positions 7, 31, 49 or 52 (see (7)). As an illustration the result for the search with Glu 7 is shown in (9), where the top row lists positions in the amino-acid sequence, the first column the scores of each possible assignment pathway (see below), and all other entries refer to the spin-system list (6).

Score ↓	7	6	5	4	3	2	1	→	Sequence
100	17	30	52						
140	17	30	14	59	52				
200	17	19	14	59	52				
200	17	19	51	68	14				
200	17	30	14	59	1				
200	17	30	14	64	4				
200	17	30	14	64	59				
260	17	19	14	59	1				
260	17	19	14	64	4				
260	17	19	14	64	59				
260	17	30	14	64	1				
300	17	30	14	68	51				
320	17	19	14	64	1				
360	17	19	14	68	51			←	Correct pathway

(9)

The result (9) can be rationalized from the following facts: (i) It is assumed that the search leads from the start point toward the N-terminus ($d_{\alpha N}$ and $d_{\beta N}$ have this directionality when the search starts from HN but with d_{NN} this assumption may lead to an erroneous assignment; see Billeter et al., 1982). Using the NOE connectivity list (8) and the amino-acid sequence (7) it is clear that only the connectivities to spin systems 19 and 30 need to be considered. (ii) Using the NOE connectivity lists for the spin systems 19 and 30 and the amino-acid sequence, the possible assignment steps between residues 6 and 5 are identified. These are combined with all possible spin-system pairs for the step 7 to 6 to obtain all possible assignment pathways for the segment 7–5. In this way the search is continued to residue 3. Each of the resulting rows in (9) is then attributed a score (first column in (9)) based on the number of observed NOEs for the pathway, with the contributions from $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ being 60, 40 and 20 points, respectively. Ideally, the pathway with the highest score will then be the correct one. This automated procedure proved to deal successfully with erroneous assignment pathways that result when a longer-range NOE connectivity, for example, $d_{\alpha N}(i, i+3)$ is mistaken for a sequential relation $d_{\alpha N}$. In contrast, problems arise if critical connectivities are missing. In this case, the user must manually investigate the spectrum to look for missing or overlapping cross peaks.

The number of possible connectivity pathways is reduced by starting the automated search at

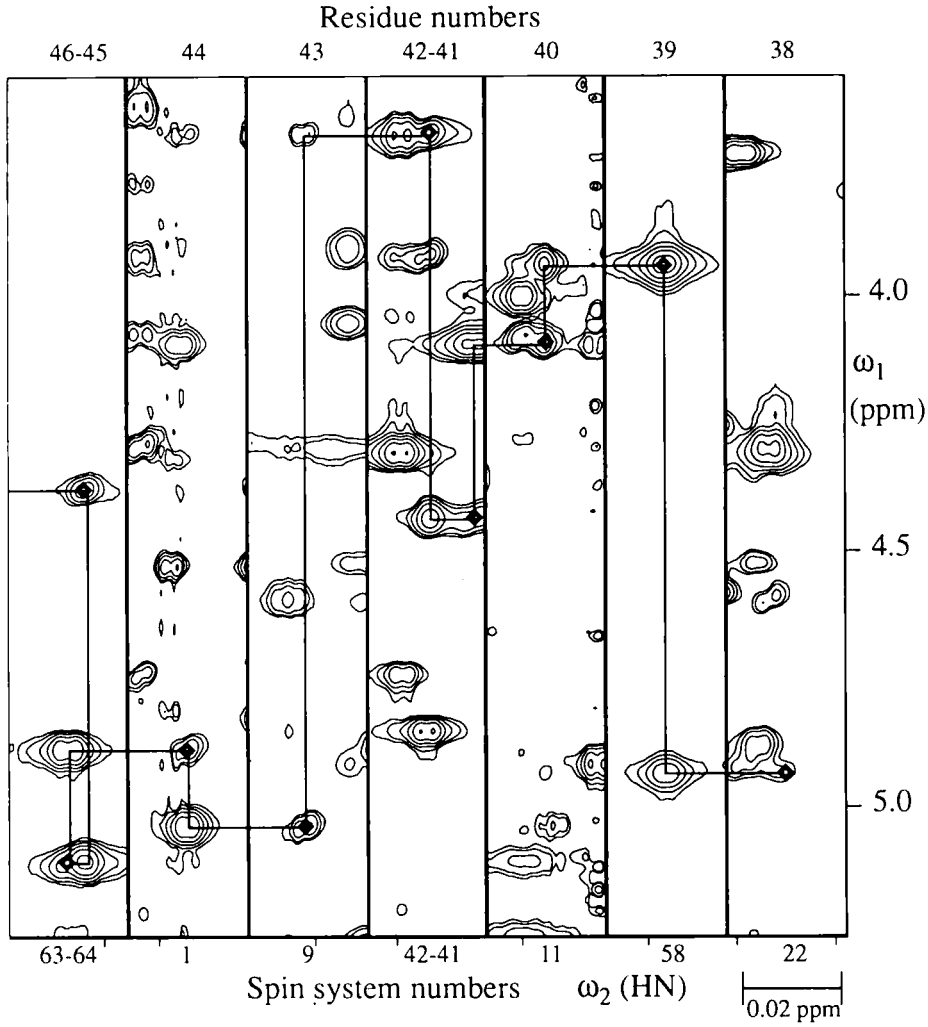

Fig. 7. Ordering by apparent sequence number of vertical strips from a NOESY spectrum which contain the cross peaks of individual spin systems at the ω_2-position of the backbone amide proton. This presentation is used by *EASY* to visualize sequential connectivities. The figure shows the final arrangement of the nine spin systems which constitute the polypeptide segment 38-46 of BPTI. The $d_{\alpha N}$ connectivities are outlined with lines, the intraresidual HN-HA cross peaks are identified with black squares.

residues or residue pairs that are unique on the level of the spin-system identification (pp. 131–134 in Wüthrich, 1986). Furthermore, a $d_{\beta N}$ connectivity will be used to establish or extend a pathway only if $d_{\alpha N}$ and d_{NN} are not found (Billeter et al., 1982), but in all instances the contributions from $d_{\beta N}$ to the score are used.

In the overall strategy for obtaining sequence-specific assignments, the principal role of the automated routines is to identify those among the large number of assignment pathways which have the highest probability of being correct. In order to identify the correct solution, a selection of spin systems may be ordered by apparent residue number rather than chemical shifts, and displayed in parallel strips (Fig. 7). In the final result of using *EASY* for NMR assignments of BPTI,

which included interactive addition of seven connectivities, 14, 3, 2, 37, 2 and 342 pathways were found, respectively, for the continuous polypeptide segments 3–7, 10–12, 14–17, 17–36, 36–37 and 38–58. In all cases, the pathway with the highest score was found to coincide with the previously established assignments (Wagner and Wüthrich, 1982; Wagner et al., 1987).

Input of ^1H–^1H distance constraints for structure calculations

Once the sequence-specific assignments of the spin systems are available, the identification of the amino-acid spin systems is completed interactively so that all entries (or, in practice, nearly all entries) in the proton list (3) can be made. Since up to this point only a fraction of the NOESY cross peaks were needed for the sequential assignments, the thus completed proton list is now used to assign the remaining NOESY cross peaks. These peaks are then integrated, and the resulting NOE intensities represent the final results obtained with *EASY*. In our laboratory these are used as the input for the programs CALIBA, HABAS and DIANA for protein structure calculations (Güntert et al., 1989, 1991).

NOESY cross-peak assignments

In principle, cross-peak assignments based on a complete proton list (3) are straightforward. In practice, however, this step of a protein structure determination is often more laborious than obtaining sequence-specific resonance assignments. A major practical difficulty arises because the proton list contains chemical shifts taken from 2QF-COSY, and chemical shift variations of ± 0.02 ppm between corresponding 2QF-COSY and NOESY spectra are not uncommon. This means that, based on the proton list from 2QF-COSY, one can easily have 50 possible assignments for one NOESY peak in a crowded region of the spectrum (Fig. 8A). In *EASY* an additional data structure, called the *assignment* list (10), is introduced to deal with these multiple possible assignments:

Peak 280	$\omega_2 \rightarrow$		
ω_1	41	349	(10)
$\downarrow$			
450	1	1	
134	1	1	

The entry (10) states that peak 280 has four possible assignments corresponding to protons 41 and 349 in ω_2, and 450 and 134 in ω_1 ('1' in the matrix identifies valid possibilities, '0' would imply that the possibility was ruled out). A simple routine implemented in *EASY* for the assignment of the NOESY peak list notes for every peak those protons which fall inside a rectangle with dimensions equal to the largest uncertainty expected between the proton list and the NOESY peak list (see Fig. 8A). Usually though, this results in very few peaks having an unambiguous assignment. The number of unambiguous assignments can be significantly increased by first identifying reference peaks within the NOESY spectrum for which the assignment is certain. These are mainly the previously assigned intraresidual and sequential NOEs. These reference peaks can then be used to update the proton chemical shifts so that the proton list matches the NOESY peak list. Once this has been done, additional unambiguous assignments may be found by looking for proper align-

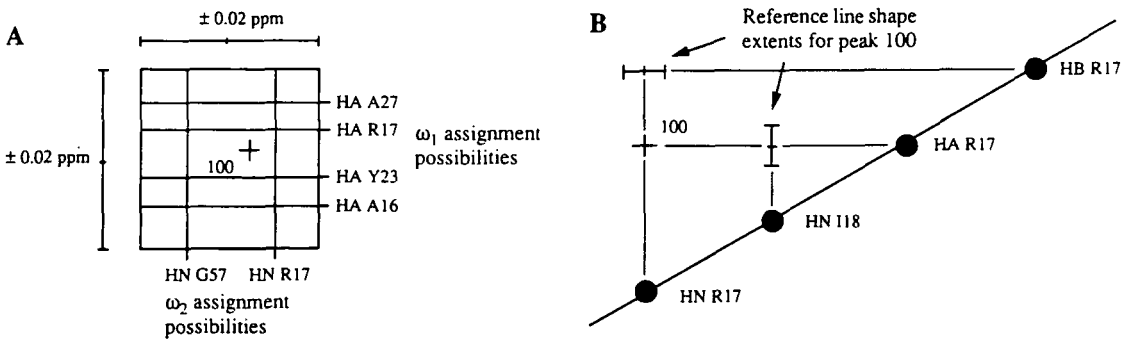

Fig. 8. (A) Multiple assignment possibilities arising from chemical-shift degeneracy within 0.02 ppm in the proton list. The peak numbered 100 is used as an illustration. The different possible assignments are stored in the *assignment* list for peak 100. (B) Illustration of the use of reference line shapes, again using peak 100 as an example. A list indicates those protons which are used to determine cross peak positions from which to extract line shape data. Note that only the extents of the line shapes ω_1 and ω_2 are stored, not the line shapes themselves.

ment between reference peaks and peaks with ambiguous assignments. If reference peaks for one of the assignment possibilities do not align, this possibility can be removed. This procedure is automated in *EASY*.

Finally, to present a more complete picture, reference should again be made to Fig. 1, which includes multiple cycles of NOESY cross peak assignments and structure calculations. Typically, unique assignments are initially obtained for less than 50% of all NOESY cross peaks. By reference to preliminary structures many ambiguities can subsequently be removed, a job that is done outside *EASY* (Fig. 1), but the results of which can be input to *EASY* for the peak integration.

Line-shape definition and peak integration

In a spectrum with little peak overlap, simple summation of data points above a certain minimal intensity level is sufficient to accurately determine the peak volume. However, in NOESY spectra of proteins containing several thousands of cross peaks, overlap is quite common and peak integration is not trivial. In the *EASY* program, a technique proposed by Denk et al. (1986) has been implemented for this purpose. This approach consists of three steps: (i) reference line shapes along ω_1 and ω_2 are determined for each proton in the 2D spectrum; (ii) the peaks are grouped together into clusters of overlapping peaks, where two peaks are said to overlap if the rectangles defined by their line shape extents intersect and (iii) the volume of each peak in a cluster is determined by a linear least-squares fit of the peak shapes constructed from the ω_1 and ω_2 reference line shapes to the experimental data points in the spectrum. In mathematical terms, the problem is one of adjusting the volumes V_p to minimize the expression (11):

$$\sum_{(\omega_1,\omega_2)} \left[S(\omega_1,\omega_2) - \sum_{p=1}^{m} V_p L_1^p(\omega_1) L_2^p(\omega_2) \right]^2 \qquad (11)$$

where: $S(\omega_1,\omega_2)$ is the spectral intensity at the point defined by the coordinates (ω_1,ω_2); V_p, the

volume of cross peak p; m, the number of cross peaks in the cluster; and L_i^p is the reference line shape for the ω_i resonance of cross peak p. This linear least-squares problem is solved using standard methods (Press et al., 1986). An uncertainty can be obtained for each peak volume by calculating the square-root of the above function over the peak region (not the total cluster). This can then be expressed as a percentage of the calculated volume.

Although reference line shapes can be identified interactively, improved efficiency is achieved with an automated routine for the line shape selection implemented in *EASY*. This routine first searches for isolated cross peaks which do not overlap with others. For a given proton, the most intense isolated peak is then chosen to provide the line shape (peaks located close to the diagonal can be excluded from the search as they often have severe base-plane distortion). Next, the routine records the distance from the peak maximum down to a predefined base level. These *line-shape extents* (Fig. 8B) are stored in a *line-shape list*. This list has an entry for each proton in the proton list, and for each entry ω_1 and ω_2 line-shape extents can be recorded.

CONCLUDING REMARKS

The program EASY supports the interpretation of 2D ^{1}H NMR spectra of proteins with a number of automated routines and a window-based environment for interactive work. Quite naturally, the automated routines are similar to those developed by others (see the *Introduction* for key references) to perform individual steps of the established assignment procedure (Wüthrich, 1986). In our experience, the principal asset of *EASY* lies in the fact that these automated routines are part of an integrated package with the facilities for interactive work, which enables one to pursue the spectral analysis all the way from the initial peak-picking to the preparation of the input for structure calculations.

In the evaluation of 2D ^{1}H NMR spectra with EASY, by far the most time-consuming part is the interactive work needed to check the results of the automated routines, or to perform operations that are not supported by automated routines. Compared with the amount of interactive work, the execution times for all automated routines described here are hardly a limiting factor. For example, the peak-picking of a 2QF-COSY spectrum of BPTI described in the section '*Automated anti-phase peak-picking*' took less than three minutes, and the automated integration of the cross peaks in the left half of a NOESY spectrum of BPTI (605 cross peaks) required less than 1 minute of CPU time on a Sun4 workstation.

Although the automated routines in *EASY* are specifically geared to the analysis of 2D ^{1}H NMR spectra of proteins, the interactive facilities are generally applicable for work with NMR spectra. An extended version of *EASY* for work with heteronuclear 3D NMR spectra is in development. Many of the facilities of the 2D version of the program can readily be adapted for this purpose. For example, a sequentially ordered arrangement of [^{1}H,^{1}H]-NOESY strips (Fig. 7) can also be assembled from a 3D heteronuclear-correlated [^{1}H,^{1}H]-NOESY spectrum.

ACKNOWLEDGEMENTS

Financial support by the Kommission zur Förderung der wissenschaftlichen Forschung, KWF (project 1615.1), Bruker-Spectrospin AG and the Schweizerischer Nationalfonds (project 31.25174.88) is gratefully acknowledged. We thank R. Baumann for technical assistance, Drs. L.

Orbons and K. Berndt for the use of unpublished spectra of BPTI, and R. Marani for the careful processing of the manuscript.

REFERENCES

Billeter, M., Braun, W. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 321-346.
Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400-415.
Braun, W. (1987) *Q. Rev. Biophys.*, **19**, 115-157.
Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119-127.
Clore, G.M. and Gronenborn, A.M. (1989) *CRC Crit. Rev. Biochem. Mol. Biol.*, **24**, 479-564.
Denk, W., Baumann, R. and Wagner, G. (1986) *J. Magn. Reson.*, **67**, 386-390.
Deisenhofer, J. and Steigemann, W. (1975) *Acta Crystallogr. sect. B* **31**, 238-250.
Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467-482.
Glaser, S. and Kalbitzer, H.R. (1987) *J. Magn. Reson.*, **74**, 450-463.
Güntert, P., Braun, W., Billeter, M. and Wüthrich, K. (1989) *J. Am. Chem. Soc.*, **111**, 3997-4004.
Güntert, P., Braun, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 517-530.
Hoch, J.C., Hengyi, S., Kjær, M., Ludvigsen, S. and Poulsen, F.M. (1987) *Calsberg Res. Commun.*, **52**, 111-122.
Holak, T.A., Scarsdale, J.N. and Prestegard, J.H. (1987) *J. Magn. Reson.*, **74**, 546-549.
Kay, L.E., Clore, G.M., Bax, A. and Gronenborn, A.M. (1990) *Science*, **249**, 411-414.
Kjær, M., Andersen, K.V., Ludvigsen, S., Hengyi, S., Madsen, J.C. and Poulsen, F.M. (1990) *Abstracts XIV ICMRBS, Warwick*, P3-21.
Kleywegt, G.J., Lamerichs, R.M.J.N., Boelens, R. and Kaptein, R. (1989) *J. Magn. Reson.*, **85**, 186-197.
Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627-633.
Meier, B.U., Mádi, Z.L. and Ernst, R.R. (1987) *J. Magn. Reson.*, **74**, 565-673.
Neidig, K.P., Bodenmüller, H. and Kalbitzer, H.R. (1984) *Biochem. Biophys. Res. Comm.*, **125**, 1143-1150.
Oh, B.H., Westler, W.M., Darba, P. and Markley, J.L. (1988) *Science*, **240**, 908-911.
Otting, G., Qian, Y.Q., Billeter, M., Müller, M., Affolter, M., Gehring, W.J. and Wüthrich, K. (1990) *EMBO J.*, **9**, 3085-3092.
Pardi, A., Billeter, M. and Wüthrich, K. (1984) *J. Mol. Biol.*, **180**, 741-751.
Pfändler, P., Bodenhausen, G., Meier, B.U. and Ernst, R.R. (1985) *Anal. Chem.*, **57**, 2510-2516.
Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes, The Art of Scientific Computing*, Cambridge University Press, Cambridge, pp. 509-515.
Stoven, V., Mikou, A., Piveteau, D., Guittet, E. and Lallemand, J.-Y. (1989) *J. Magn. Reson.*, **86**, 163-168.
Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633-644.
Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 347-366.
Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Gō, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611-639.
Weber, P.L., Malikayil, J.A. and Müller, L. (1989) *J. Magn. Reson.*, **82**, 419-426.
Wider, G., Lee, K.H. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 367-388.
Wlodawer, A., Walter, J., Huber, R. and Sjölin, L. (1984) *J. Mol. Biol.*, **180**, 301-329.
Wlodawer, A., Nachman, J., Gilliland, G.L., Gallagher, C. and Woodward, C. (1987) *J. Mol. Biol.*, **198**, 469-480.
Wüthrich, K. (1983) *Biopolymers*, **22**, 131-138.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.
Wüthrich, K. (1989) *Science*, **243**, 45-50.
Wüthrich, K., Billeter, M. and Braun, W. (1984) *J. Mol. Biol.*, **180**, 715-740.