J-Bio NMR 006

Computer-assisted assignment of 2D ^{1}H NMR spectra of proteins: Basic algorithms and application to phoratoxin B

Gerard J. Kleywegt, Rolf Boelens, Michel Cox, Miguel Llinás* and Robert Kaptein**

Department of Chemistry, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands.

Received 22 November 1990
Accepted 14 January 1991

Keywords: 2D NMR; Proteins; Semi-automatic assignment; Software package

SUMMARY

A suite of computer programs (*CLAIRE*) is described which can be of assistance in the process of assigning 2D ^{1}H NMR spectra of proteins. The programs embody a software implementation of the sequential assignment approach first developed by Wüthrich and co-workers (K. Wüthrich, G. Wider, G. Wagner and W. Braun (1982) *J. Mol. Biol.* **155**, 311). After data-abstraction (peakpicking), the software can be used to detect patterns (spin systems), to find cross peaks between patterns in 2D NOE data sets and to generate assignments that are consistent with all available data and which satisfy a number of constraints imposed by the user. An interactive graphics program called *CONPAT* is used to control the entire assignment process as well as to provide the essential feedback from the experimental NMR spectra. The algorithms are described in detail and the approach is demonstrated on a set of spectra from the mistletoe protein phoratoxin B, a homolog of crambin. The results obtained compare well with those reported earlier based entirely on a manual assignment process.

INTRODUCTION

In recent years, automation of the assignment of NMR spectra of biomacromolecules (in particular, of 2D ^{1}H NMR spectra of proteins) has developed into a holy grail pursued by scientists in many laboratories the world over. Since the assignment process tends to be tedious and labo-

*On leave from the Department of Chemistry, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.

** To whom correspondence should be addressed.

rious, attempts to automate it are worthwhile. Also, since it involves a considerable amount of 'bookkeeping', the problem would appear to be susceptible to automation, at least to some extent. Unfortunately, there are a number of complicating factors which render the entire exercise less than trivial, in particular the fact that computer software tends to be less apt at handling imperfect, incomplete and 'fuzzy' visual information than experienced NMR spectroscopists.

The combined efforts so far have resulted in a fair number of algorithms and programs (or program packages) that differ widely in philosophy and applicability. Kraulis (1989) has developed a program called *ANSIG* which is essentially an assignment-support system, or 'electronic drawing board', that alleviates the task of bookkeeping and provides extensive consistency checks which help prevent incorrect assignments. Wand and co-workers have developed a novel assignment strategy, the Main-Chain-Directed or MCD approach (Englander and Wand, 1987; DiStefano and Wand, 1987; Wand et al., 1989; Feng et al., 1989) which they are in the process of automating (Wand and Nelson, 1988). However, most efforts to date have been aimed at implementing (parts of) Wüthrich's sequential assignment strategy (Wüthrich et al., 1982; Wüthrich, 1986, 1989; Kaptein et al., 1988). In its basic form, this approach encompasses the following steps:

- detecting and classifying spin systems;
- detecting inter-spin-system connectivities;
- matching spin systems to specific residues of the protein under investigation (i.e., making sequence-specific assignments).

Automating this paradigm tends to lead to three analogous steps, usually preceded by a preprocessing stage in which the original data (2D NMR spectra) are transformed, yielding four steps in all:

- data abstraction (e.g. peakpicking, spectrum matching or aligning, possibly in conjunction with noise-reduction or symmetry-detection operations);
- pattern detection (i.e., finding possible spin systems);
- detection of inter-pattern correlations (i.e., inter-residue cross peaks in a 2D NOE spectrum);
- generation of the actual assignments.

For a discussion of the first of these four steps we refer to Kleywegt et al. (1990) and the references cited therein. There have been several reports on programs which basically or exclusively tackle the problem of pattern detection on the basis of, usually, COSY, RELAY and/or HOHAHA spectra (Meier et al., 1984, 1987; Neidig et al., 1984; Pfändler et al., 1985; Weber et al., 1989). The program of Billeter et al. (1988) deals with the final assignment step (assuming that the user has somehow accomplished the preceding tasks) in combination with extensive bookkeeping and consistency checking. At present, there are five program suites that tackle all three final steps of the automated assignment procedure in the Wüthrich tradition: the one described by Cieslar et al. (1988), the one reported by Eads and Kuntz (1989), Van de Ven's *PROSPECT* program (Van de Ven, 1990; Van de Ven et al., 1990), an expert system called *PEPTO* (Catasti et al., 1990) and our *CLAIRE* package (Kleywegt et al., 1989). It should be pointed out that, as far as we are aware, none of these packages is capable of fully automatically assigning real protein spectra. In other words, each of them requires some sort of input and feedback from the user at one or several stages of the assignment process.

In a previous communication (Kleywegt et al., 1989), we have briefly discussed the basic ideas implemented in the programs that constitute the *CLAIRE* package (*CLAIRE* is an acronym for 'CLuster of programs for the Assignment of Individual REsonances') as well as the results obtained with a test data set pertaining to crambin. Here we shall describe the details of the algorithms (including some refinements and extensions) as well as the application of our approach on an entirely experimental data set consisting of spectra of phoratoxin B, a protein that shows homology to crambin (Clore et al., 1987; Lecomte et al., 1987).

ALGORITHMS AND PROGRAMS

Design considerations

Given a set of patterns, NOE information, and the primary structure of a protein, one could design a program that generates a 'best' assignment for the entire protein or, alternatively, the set of all possible assignments that are consistent with the data. There are, however, serious disadvantages to these approaches. Any 'best' assignment may still contain errors and, hence, forces the user to perform extensive error-checking and, in the worst-case scenario, to assign the better part of the protein manually after all. On the other hand, the total set of consistent assignments could easily number in the hundreds, thus leaving it up to the user to find the correct needle in the haystack of possibilities. *CLAIRE* implements a rather evolutionary approach to the assignment process: it becomes an iterative process in which a few assignments are made in every cycle, which then, inherently, limit the span of possibilities for assigning the remainder of the protein. In addition, the user is in control of the assignment process at all times.

CLAIRE has been designed in a modular fashion, where each program has one or several well-defined tasks (e.g., detecting patterns, generating possible assignments, attempting to extend patterns) which it generally completes quite rapidly. Table 1 contains a list of some of the programs that are part of the *CLAIRE* package, their purposes, and typical runtimes. A clear advantage of *CLAIRE* is that the user need not ever edit a file other than output files and job-control files. The interactive programs *AUTOPI* and *CONPAT* provide the necessary interface for that. *AUTOPI* is used to set the values of the input parameters for all other programs and *CONPAT* is used to manipulate all attributes of the pattern data structure (vide infra) and to enter the amino acid sequence.

TABLE 1
OVERVIEW OF PROGRAMS IN *CLAIRE*[a]

Program	Purpose	Typical runtime [b]
AUTOPI	interactive parameter input for all other programs	n/a
CONPAT	interactive pattern editing	n/a
SPIN2D	pattern detection	4:00
FILPAT	pattern filtering	0:28
XREF2D	detection of inter-pattern NOEs and pattern reshuffling	0:22
ASSI2D	generation of possible assignments	0:13
CHECKA	generation of assignments for individual protons, etc.	0:32

[a] Only programs which are described in the text have been included. Our previous program *EDAPAT* (Kleywegt et al., 1989) is obsolete; *CONPAT* now provides a superset of its functionality.

[b] Runtimes (where applicable) are in CPU minutes:seconds on a VAXstation 3100 operating in batch mode.

The software was written in FORTRAN-77 and can be ported to various workstations with minor adaptations to the command script that is used to run the individual programs. However, the graphical facilities in *CONPAT* involve system-dependent graphics operations. Graphics binders have been written for the PLOT10 (Tektronix) and VWS (Digital Equipment Corp.) graphics libraries. *CLAIRE* constitutes an integral part of our laboratory's *TRITON* software package forthe processing and analysis of multi-dimensional NMR spectra. A stand-alone version of *CLAIRE* (running on VAX workstations with VMS and VWS) is available from the authors upon request.

Objective

Symbolically, a protein can be represented as an ordered set of N_{res} amino acid residues:

$$\{R_i\}\ ;\ i = 1, N_{res}\ ;\ R_i \in \{\text{Ala, Arg, ..., Val}\} \tag{1}$$

Each residue R_i contains a number $N_p(i)$ of protons:

$$R_i \leftrightarrow \{p_{ij}\}\ ;\ j = 1, N_p(i) \tag{2}$$

In a 1D NMR experiment, in principle, each proton gives rise to a signal in frequency space. In other words, there exists a mapping A such that:

$$A\ (p_{ij}) = c_{ij} \tag{3}$$

where c_{ij} is the chemical shift of proton p_{ij} (in ppm or channel numbers). The aim of the assignment process, then, is to determine the mapping A, i.e. to generate an assignment list. The analytical expression for A in terms of structural parameters is generally unknown. Another difficulty involves the fact that, in principle, A is not a one-to-one (bijective) mapping, i.e. more than one proton may be mapped onto the same point in frequency space (chemical-shift degeneracy). This is a fundamental problem that improvement of the experimental conditions cannot guarantee to remedy (even in the limit of 'zero' linewidth two protons may still have the same chemical shift). Another confounding problem is the fact that the mapping A is not necessarily defined for all protons p_{ij}, i.e. some protons may not be observable by NMR techniques (for instance, because they exchange too rapidly).

In practice, 1D NMR is unsuitable for the assignment of all (observable) protons in large molecules. Fortunately, a large number of 2D and 3D NMR techniques have been developed that offer additional information. For example, a cross peak in a COSY spectrum at position (c_A,c_B) tells us that there must be two protons p_{kA} and p_{kB} (both in residue R_k), which, in most cases, are no more than three chemical bonds apart, such that A (p_{kA}) = c_A and A (p_{kB}) = c_B. However, the absence of such a cross peak does not allow for any definitive conclusions. Moreover, noise and artefacts may look deceptively much like peaks and may also obscure or distort real peaks (particularly weak ones). These are all problems which are typically taken care of in the data-abstraction stage. Peaks may be extracted automatically (Cieslar et al., 1988; Kleywegt et al., 1990), manually (Eads and Kuntz, 1989; Kleywegt et al., 1990) or using a combined method (Kleywegt et al., 1990). Precise peak positions can be determined by calculating the 'centre of gravity' (Cieslar et al., 1988; Eads and Kuntz, 1989) or by performing an interpolation (Kleywegt et al., 1990). Unfortunately, no peakpicking procedure can be expected to yield 'exact' peak positions, which necessitates the introduction of some sort of 'tolerance' criterion.

In practice, the mapping *A* will be variously dependent on the experimental conditions or on noise, so that in two different spectra the values of *A* (p_{ij}) may differ slightly. Although this feature is sometimes exploited in order to resolve ambiguities in the assignment, for the purpose of computer-assisted assignment it introduces a number of problems. These may to some extent be remedied by using mathematical matching or alignment methods; based on our experience, however, correspondence of two different spectra within approximately one channel can be better achieved at the time of the experiment.

Our approach requires an input consisting of two sets of cross peaks:

$$\{X_{yi}\}\ ;\ i = 1,\ N_{pks}(y)\ ;\ y \in \{\text{HOHAHA, 2D NOE}\} \tag{4}$$

where $N_{pks}(y)$ is the number of elements (peaks) of the data set extracted from the spectrum of type y. Each peak has some descriptive information associated with it:

$$X_{yi} \leftrightarrow (c_{1yi},\ c_{2yi},\ I_{yi}) \tag{5}$$

where c_{1yi} and c_{2yi} are the peak's coordinates and I_{yi} is its intensity. Most *CLAIRE* programs use the peaks to construct two 'logical spectra' S_y. For two chemical-shift or channel numbers c_A and c_B, $S_y\,(c_A,c_B)$ is true if and only if

$$\exists\ X_{yi} |\ |\ c_{1yi} - c_A\ | \leq \tau \wedge |\ c_{2yi} - c_B\ | \leq \tau \wedge I_{yi} \geq M_y \tag{6}$$

where τ is the tolerance and M_y the minimum intensity for peaks from spectrum y to be considered. In other words, $S_y\,(c_A,c_B)$ is true if there is a sufficiently intense peak near (c_A,c_B). Both spectra are assumed to be square and symmetric with respect to the diagonal.

Using these two spectra, the software searches for patterns, yielding a set

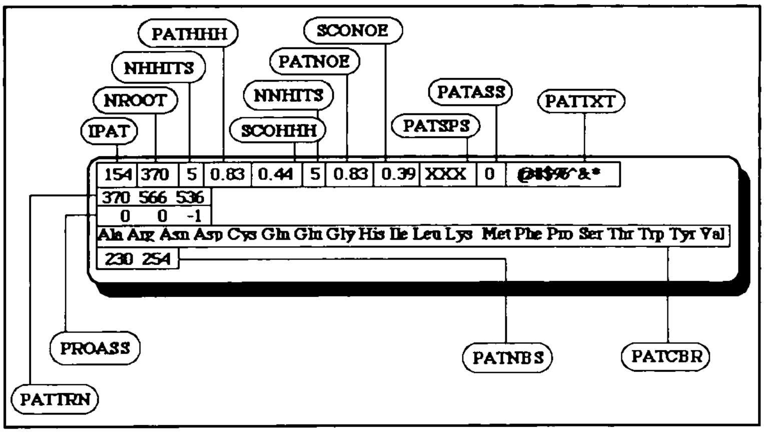
$$\{P_i\}\ ;\ i = 1,\ N_{pat} \tag{7}$$

where N_{pat} is the number of patterns. Each pattern contains a set of channel numbers d_{ij} which are mutually highly correlated through cross peaks:

$$P_i \leftrightarrow \{d_{ij}\}\ ;\ j = 1,\ N_c(i) \tag{8}$$

where $N_c(i)$ is the number of channels in pattern P_i. The objective is to first find the correspondence between residues in the protein and patterns in the experimental data, i.e. to determine the mapping ***B***

$$\boldsymbol{B}\,(R_i) = P_j \tag{9}$$



154 370 5 0.83 0.44 5 0.83 0.39 XXX 0 @#$%^&*
370 566 536
0 0 -1
Ala Arg Asn Asp Cys Gln Gln Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val
230 254

Fig. 1. Example of a pattern as it is used and stored by *CLAIRE*. The names of the attributes are shown; the significance of the.most important ones is explained in the text.

and, subsequently, the correspondence between the protons of the residue and the channel numbers of the pattern:

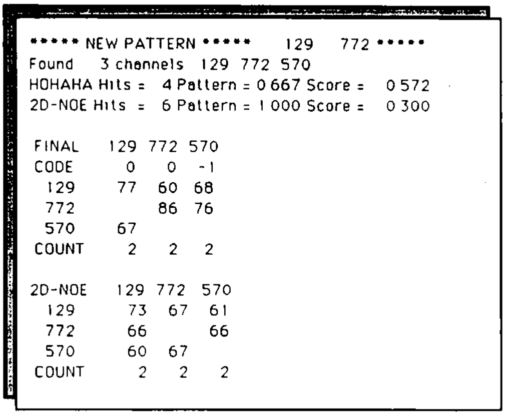
$$A\ (p_{ik}) = d_{jl} \tag{10}$$

Patterns

. Patterns are the basic entities around which all programs in the package revolve. A pattern may be equivalent to a partial or complete spin system or to a combination of two or more spin systems (e.g. in the case of aromatic residues). A pattern's major components are, of course, its channel numbers, but in the actual data structure several other items of information are stored as well. In Fig. 1, one record from a pattern file is displayed along with the names of the various attributes. Some of the more important attributes are discussed in the following; most of them can be edited through the software (*CONPAT*).

NROOT is the pattern's channel in the amide region which is probably (but not necessarily, for instance in the case of an arginine side-chain $N^{\varepsilon}H$) attributable to the backbone-amide proton. It is used in the detection of inter-pattern NOEs.

In order to explain the meaning of the variables NHHITS, PATHHH, SCOHHH, NNHITS, PATNOE and SCONOE, we refer to Fig. 2 and Fig. 3. Figure 2 shows some (edited) output from SPIN2D which is produced whenever a new pattern has been detected. This particular pattern stems from a HOHAHA peak that is located near position (129, 772). It consists of three channels, namely 129 and 772 themselves, as well as channel 570. The two small matrices give an overview of the cross peaks of these channels that are present in the HOHAHA and the 2D NOE peak data set, respectively (see also Fig. 3). The first row in the first matrix (labelled 'CODE') contains the values of PROASS (see below). The next three rows depict the intensities of the relevant cross peaks that were encountered in the HOHAHA peak data set (actually, the numbers are: integer (10 * log10 (intensity) + 0.5); thus, a value of 60 means that the peak had an intensity of about 10^6). The final row contains the cross-peak counts: the number of channels that each channel has cross peaks with, not counting itself. The matrix labelled '2D NOE' contains basically the same



```
***** NEW PATTERN *****     129    772 *****
Found    3 channels  129  772  570
HOHAHA Hits =   4 Pattern = 0 667 Score =   0 572
2D-NOE Hits  =   6 Pattern = 1 000 Score =   0 300

 FINAL    129  772  570
 CODE       0     0   -1
  129      77    60   68
  772            86   76
  570      67
 COUNT      2     2    2

2D-NOE    129  772   570
  129       73   67   61
  772       66         66
  570       60   67
 COUNT       2    2    2
```

Fig. 2. Example of the output produced by *SPIN2D* when it has encountered a new pattern (slightly edited). Shown are the HOHAHA peak from which the pattern originates in this case, a peak at (129,772) -, the final set of channels that constitute the pattern, the various 'scores' and the cross-peak matrices. Refer to the text for a more detailed explanation.

information (mutatis mutandis). The two lines above the cross-peak matrices contain the values for NHHITS, PATHHH, SCOHHH (first line) and NNHITS, PATNOE and SCQNOE (second line), in this order. NHHITS is the total number of cross peaks in the HOHAHA cross-peak matrix, not counting any diagonal peaks. PATHHH is this number, divided by the maximum number of possible cross peaks (for N channels, this maximum is simply: $N^2 - N$, once again disregarding diagonal peaks). The idea is that a high value for PATHHH means that the channels of the pattern have relatively many cross peaks in the HOHAHA spectrum and therefore form a more 'consistent' set than another pattern which closely resembles it but which has a lower value for PATHHH. Precisely this argument is used in *FILPAT*: if two patterns are fuzzy subsets of one another (i.e., they contain approximately the same channel numbers), then the one with the highest value of PATHHH is retained. SCOHHH is a measure for how close the peaks of this pattern lie to the (integer) channels: the lower this value, the closer the peaks are on average. The values of NNHITS, PATNOE and SCONOE should be interpreted in similar terms.

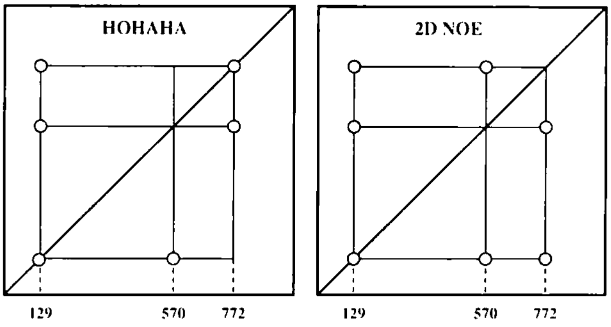


Fig. 3. Schematic overview of the cross peaks encountered in the HOHAHA (left panel) and 2D NOE (right panel) peak data sets for the pattern displayed in Fig. 2. The small circles correspond to non-zero entries in the cross-peak matrices of Fig. 2.

Since these numbers give an indication of the 'reliability' or 'consistency' of a pattern, they can be used in *FILPAT* for disposing of undesired patterns. This can be accomplished by requiring that PATHHH for an acceptable pattern exceeds a certain minimum value, and/or that the ratio of PATNOE and PATHHH exceeds a certain minimum value. The underlying heuristics associated with these options are:

- a pattern is more reliable if its channels have a lot of cross peaks (for example, one wouldn't 'trust' a pattern that consists of five channels which have only four cross peaks);
- a pattern is more reliable if there are more cross peaks in the 2D NOE spectrum than there are in the HOHAHA spectrum.

Once patterns have been altered, the values of all these variables are set to zero, since they are no longer valid or, indeed, needed.

PATSPS and PATASS define the 'assignment state' of the pattern. PATSPS is the three-letter code of the amino acid type to which the pattern has been assigned, or 'XXX' by default as long as it remains unidentified. PATASS is the number of the amino acid residue to which the pattern has been assigned (i.e., the sequence number in the protein), or zero if it is still unassigned. This yields three possible 'assignment states' for each pattern:

- unidentified pattern (PATSPS = 'XXX', PATASS = 0);
- identified but unassigned pattern (e.g. PATSPS = 'Gly', PATASS = 0);
- specifically assigned pattern (e.g. PATSPS = 'Ala', PATASS = 27).

PATTRN is the array that contains the actual channel numbers that make up the pattern.

PROASS initially contains information about the stage of the pattern-detection procedure in which the corresponding channel number was added to the set. The more negative this number is, the later the channel was added to the set (and, hence, the less reliable it is). At a later stage, this array is used to store the assignments of the individual channels.

PATCBR contains the three-letter codes of all amino acid types to which the pattern should be matched by the assignment program *ASSI2D*. For example, if the user is convinced that a pattern represents an AMX-type spin system, then the list may be narrowed down to contain only such residue types. In that case, *ASSI2D* will not consider the omitted residue types (Ala, Gly, etc.) as possible matches. However, whether or not any of the remaining residue types will match with the pattern depends entirely on the values of the parameters for *ASSI2D*.

Finally, PATNBS may contain channel numbers which the user considers to be potentially attributable to the amide proton of the C-terminal neighbour of the residue to which the pattern corresponds, i.e. NH($i+1$). *ASSI2D* may be instructed to discard the inter-pattern-correlation matrix generated by *XREF2D* and to use the information provided in the arrays PATNBS instead. The order of the channel numbers in PATNBS corresponds to decreasing likelihood of representing NH($i+1$).

It should be noted that by using PATCBR and PATNBS and by selecting sufficiently low thresholds, one can in fact use *ASSI2D* as a program that suggests all possible assignments that are consistent with this user-provided information. In such a case, the program acts similarly as Billeter's assignment program (Billeter et al., 1988).

Channels and tolerances

Since there are bound to be slight chemical-shift differences between HOHAHA and 2D NOE spectra, the precise location of a pattern's constituent channel numbers may differ slightly in both spectra. Patterns that are generated by the software contain channel numbers that are based on the HOHAHA peak data set. However, since most programs solely use the 2D NOE spectrum, it is important to adapt these channel numbers so that they comply with the 2D NOE spectrum. This will usually be done in the pattern-sorting operation that follows *FILPAT* and precedes *XREF2D*.

Another practical aspect concerns the use of some tolerance with respect to peak positions. In actual fact, we use two such tolerances, namely an absolute (or intra-spectrum) tolerance τ_A, and a relative (or inter-spectrum) one, τ_R. As explained earlier, two logical spectra S_{HOHAHA} and S_{2DNOE} are constructed from the respective peak data sets. The absolute tolerance τ_A relates to the spectrum from which information is primarily extracted, τ_R relates to the other one. In effect, this means that τ_A is used for the 2D NOE spectrum by all programs except *SPIN2D*, where it is used for the HOHAHA spectrum. Conversely, τ_R relates to the 2D NOE spectrum in *SPIN2D* and, where needed, to the HOHAHA spectrum in other programs. In Fig. 4 the use of a tolerance to build a 'logical spectrum' from peak data is illustrated.

CONPAT - interactive pattern editing

CONPAT was designed as a tool to provide the essential visual feedback to the original data, i.e. the 2D NMR spectra. Although the *CLAIRE* programs only use peak data sets extracted from a HOHAHA and a 2D NOE spectrum, *CONPAT* enables inspection of any spectrum (e.g. COSY and D_2O spectra). The extra information that this provides can be used, for example, to assign $C^{\alpha}H$ and $C^{\beta}H$ protons correctly or to trace aromatic spin systems.

Apart from offering a host of pattern-manipulation options, *CONPAT* contains a number of display options which render it useful as a stand-alone aid in any assignment process. Two of these options particularly merit mentioning, namely the pattern-contour and the scroll-contour options. Both have a common trait in that they facilitate detailed scrutiny targeted to several small areas of a spectrum simultaneously, without any visual interference from the rest of the spectrum. The idea of the pattern-contour option is that the user selects a set of N channel numbers (for

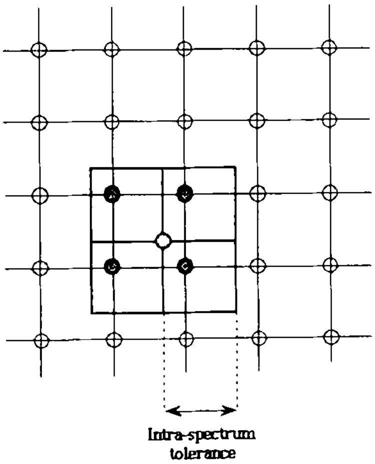


Fig. 4. Demonstration of the use of the intra-spectrum tolerance τ_A when *CLAIRE* creates 'logical spectra' (see text). A peak is in effect 'spread out' over all grid points whose coordinates differ by no more than τ_A of those of the peak.

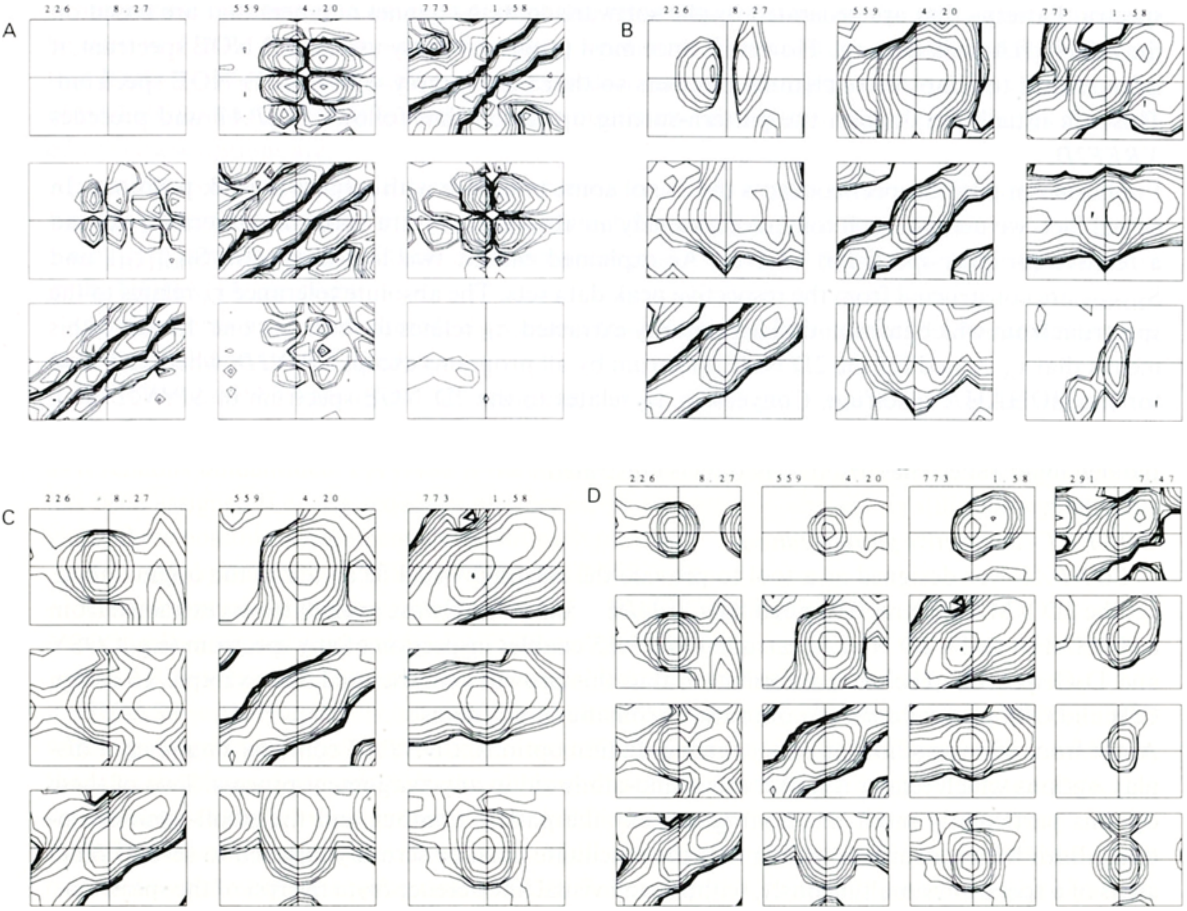


Fig. 5. Examples of the use of *CONPAT*'s pattern-contour option. Panel A shows the pattern that corresponds to the spin system of residue Ala^9 of phoratoxin B in the COSY spectrum. The boxes are twelve channels wide in both directions and they are centred about the diagonal and cross-peak positions for this pattern. Hence, for example, the top-left box extends from channel 220 to 232 in the horizontal direction (ω_2) and from 767 to 779 in the vertical direction (ω_1). It thus corresponds to the NH-C$^\beta$H cross-peak position (since this is a COSY spectrum, obviously no cross peak is observed). Since the channel numbers of the pattern refer to the 2D NOE spectrum, the actual cross peaks in other spectra may be slightly displaced. Panel B shows the same pattern in the HOHAHA spectrum. In panel C the pattern of Ala^9 is shown again in the 2D NOE spectrum. Comparison of panels B and C shows how well both spectra were aligned (which was accomplished at the time of the experiment rather than by mathematical manipulations). Panel D shows the same set of channel numbers in the 2D NOE spectrum, but with the addition of the NH channel of Arg^{10} (top row and right column). One clearly observes an intense NH(i)-NH($i+1$) peak, a slightly weaker C$^\alpha$H(i)-NH($i+1$) peak and a strong C$^\beta$H(i)-NH($i+1$) peak.

example, a pattern or a pattern plus one or two potential NH($i+1$)-channels) and produces N^2 small contour plots of small environments near all the possible diagonal and cross-peak positions (see Fig. 5).

The scroll-contour option can be employed to extend patterns, to find candidate NH($i+1$)-channels, to locate the aliphatic brethren of an aromatic spin system, etc. Again, the user selects a number of channels which are deemed interesting. With this option, it is possible to only look

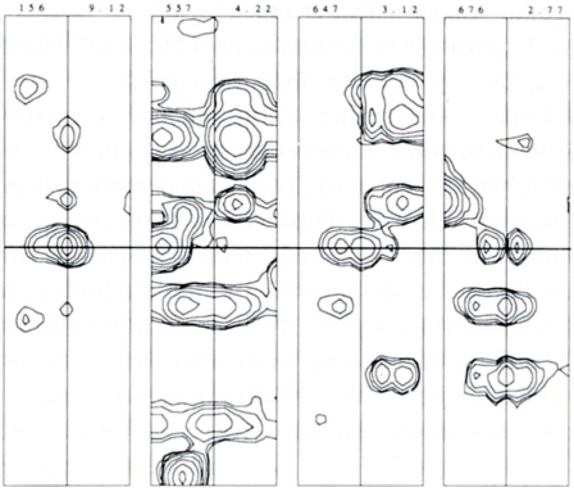


Fig. 6. Demonstration of the scroll-contour option of program *CONPAT*. We have selected the NH, $C^{\alpha}H$ and both $C^{\beta}H$ channels of the pattern corresponding to Asn[14] in phoratoxin B. For each of these channels (plus six channels to the left and right of them) we have plotted small strips in the amide region, in this case from channel 170 to 250. By a mere key-press it is possible to 'scroll' up or down (in steps of, for instance, seventy channels), thus facilitating very close scrutiny of the amide regions of the selected channels. In this case, this results in detection of the NH channel of Thr[15] (bold horizontal line) since we observe strong NH(i)-NH($i+1$) and $C^{\beta}H$(i)-NH($i+1$) contacts.

at those channels (plus a few to the left and right of it) and to 'walk' through the spectrum in small steps (see Fig. 6).

SPIN2D - pattern detection

Our pattern-detection algorithm is a modification of that used by Cieslar et al. (1988). The user has to define three search ranges (illustrated in Fig. 7):

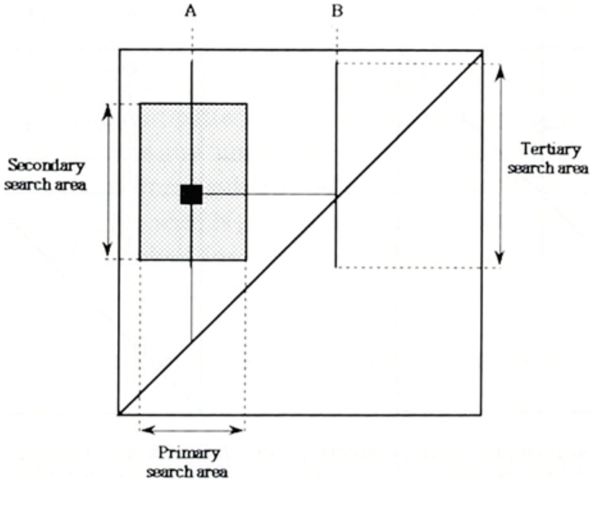


Fig. 7. Definition of the primary, secondary and tertiary search areas as used by program *SPIN2D* in the pattern-detection process (see text). Note that each peak in the dotted area of the HOHAHA spectrum may give rise to a pattern.

– the primary search range L_{1l} - L_{1u} (usually the amide region);
– the secondary range L_{2l} - L_{2u} (usually the region where $C^{\alpha}H$ and $C^{\beta}H$ protons are found);
– the tertiary range L_{3l} - L_{3u} (usually the entire aliphatic region).

Each peak in the area defined by the primary and secondary ranges in the HOHAHA spectrum (S_{HOHAHA}) yields two correlated channel numbers (c_A and c_B) and, thus, is considered to be the beginning of a pattern (see Fig. 7). Subsequently, a two-step procedure begins in which *SPIN2D* attempts to expand the pattern with channels in the tertiary range that are correlated (through cross peaks) with the channel numbers that are already part of the pattern. Initially, the program will look for channels c_C, such that c_C lies in the tertiary range and both c_A and c_B have a cross peak with it in the HOHAHA spectrum (see Fig. 8). If this procedure yields no new channels, then the program will look for channels c_C in the tertiary range that have a cross peak with channel c_A in the HOHAHA and one with c_B in the 2D NOE spectrum, or vice versa.

It is not recommended to use channels that have cross peaks with c_A and c_B only in the 2D NOE spectrum, since in that case inter-residue cross peaks may easily interfere. If this procedure has still not yielded any new channels for the pattern, optionally, all channels c_C that have a cross peak in the HOHAHA spectrum with the amide channel c_A may be collected. However, since amide-resonance overlap is not uncommon, this is a manoeuvre which is not devoid of risk.

If at this stage a pattern still consists of only two channel numbers it is retained only if the user has instructed the program to do so. However, if a pattern does contain more than two channels, a second expansion step commences. In this step, *SPIN2D* compares each of the newly found channels (c_C) with channels c_A and c_B in turn (only in the HOHAHA spectrum this time so as to prevent 'dilution' of the patterns as well as possible). In the example of Fig. 9, channels c_A and c_C have a cross peak with channel c_D in common, whereas c_B and c_C both have one with c_E. Hence, after one iteration of this second step the pattern consists of channels $\{c_A, c_B, c_C, c_D, c_E\}$. If the maximum number of such iterations M_{it} were greater than one, then an additional cycle would be carried out, this time comparing c_D and c_E with c_A and c_B in turn, etc. This process continues until M_{it} iterations have been carried out, or until, at the start of a new iteration, the number of channels

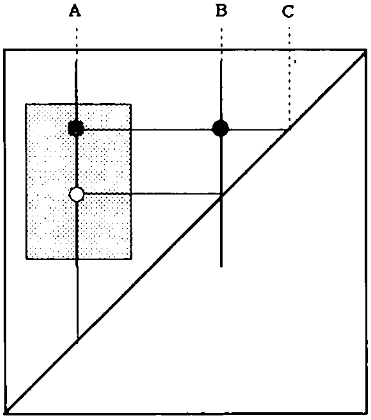


Fig. 8. Example of a successful attempt to extend a pattern in the first expansion step of *SPIN2D*. Channel C has a cross peak with both channel A and channel B and is therefore added to the pattern. Refer to the text for details.

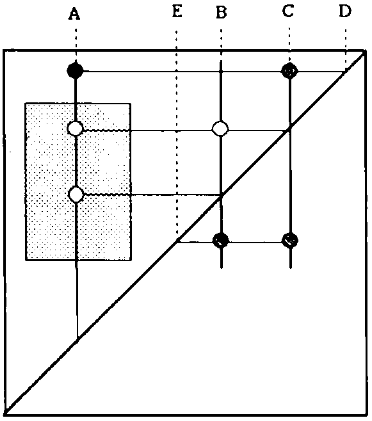


Fig. 9. Illustration of the second step of the pattern-expansion procedure of *SPIN2D* (see text).

in the pattern, $N_c(i)$, exceeds or equals a user-defined maximum value M_c. If, at this stage, the number of channels exceeds M_c, a 'channel shake-out' operation is performed. The program counts for each channel the number of other channels in the pattern that this channel has cross peaks with in the HOHAHA spectrum. If there are any channels for which this number is less than M_{xl} (a user-provided parameter), they are deleted from the pattern. If there are still too many, the value of M_{xl} is incremented by one and the process is repeated. This is continued until either $N_c(i) \leq M_c$ or $M_{xl} = M_{xu}$ (also provided by the user). In the latter case, all channels of the pattern are obviously well correlated through cross peaks and those that were added last are deleted until M_c channels remain. This approach is not useful for patterns that contain only two channel numbers (if they are considered at all). For such patterns, the program simply counts the number of cross peaks in the HOHAHA and 2D NOE spectra (yielding a number between one and four) and retains a pattern only if this number exceeds a user-defined minimum value M_2. Any 'surviving' patterns are printed (see Fig. 2) and stored in a pattern file (see Fig. 1).

FILPAT - filtering patterns

Typically, *SPIN2D* will generate five to ten times as many patterns as there are spin systems in the protein. This has two causes:

Inspection of Fig. 4 reveals that one peak may easily give rise to several patterns which differ only marginally in the exact location of their constituent channels. For instance, if the peak shown in Fig. 4 lies at (128.78, 523.31) then four patterns may arise, namely from (128, 523), (129, 523), (128, 524) and (129, 524). In addition, these patterns need not contain the same set of channels. Using the same example, and looking at Fig. 8, it is obvious that the cross peak at (c_A, c_C) may, for example, occur for channel 128, but not necessarily also for channel 129. If this would indeed be the case, then the patterns originating from channel 128 would be extended with channel c_C, whereas the other two (from channel 129) would not. In the rest of the expansion procedure the same phenomenon may occur several times, so that in the end the four patterns, even though they stem from the same peak, may end up being quite dissimilar.

In addition to all this (caused by the – necessary – use of a 'tolerance'), there is another problem. Consider Fig. 9: on channel c_A, there are cross peaks with both channel c_B and c_C inside the dotted principal-search area. Therefore, both cross peaks will give rise to one or more patterns.

All in all, the set of channels $\{c_A, c_B, c_C\}$, even though they might represent only one spin system, may give rise to up to eight different patterns. It is the purpose of *FILPAT* to remove (most of) such redundancy as well as to delete any 'unreliable' patterns (vide supra). In order to accomplish this, we have used the concept of 'fuzzy subsets' (Kleywegt et al., 1989): given two sets $T \leftrightarrow \{t_i\}$; $i = 1, N_T$ and $U \leftrightarrow \{u_j\}, j = 1, N_U$ with $N_T \leq N_U$ and given some tolerance $D \geq 0$, T is defined to be a fuzzy subset of U if and only if for every $t_i \in T$ there exists a $u_j \in U$, such that $| t_i - u_j | \leq D$. For example, a pattern $\{167, 523, 646\}$ would be a fuzzy subset of a pattern $\{166, 645, 664, 524\}$ if $D \geq 1$.

Initially, *FILPAT* deletes all patterns for which the 'scores' are considered to be too low, namely if:

$$(\text{PATHHH} < M_{pa}) \vee (\text{PATNOE} < M_{pa}) \vee (\text{PATNOE} / \text{PATHHH} < M_{ra}) \qquad (11)$$

where M_{pa} and M_{ra} are user-defined parameters. Subsequently, if the fuzzy-subset tolerance D is greater than zero, *FILPAT* will delete all patterns that are fuzzy subsets of at least one other pattern. However, it only considers pairs of patterns whose number of channels differs by no more than M_N: $|N_c(i) - N_c(j)| \leq M_N$. The rationale for this is that, in general, it is not so easy to extract small spin systems from large patterns. Suppose one has two patterns {229, 544, 568, 601} and {228, 532, 545, 567, 645, 600, 723}. It may well be that the former represents an AMX-type spin system, whereas the latter corresponds to the same spin system, but has been 'polluted' due to the phenomena outlined above. In such cases, one would like to retain both patterns (which would occur if $M_N \leq 2$). It may happen that two patterns contain the same number of channels and are fuzzy subsets of each other. In that case, the pattern with the highest value of (NHHITS + NNHITS) is kept or, in case of a tie, the one with the lowest value of (SCOHHH + SCONOE). Finally, the program will print all pairs of patterns which contain several similar channels without one pattern being a proper fuzzy subset of the other. The output that this yields may be of help to the user when sorting out the patterns (using *CONPAT*). Pairs of such patterns with different values of NROOT are flagged, since they might point to arginyl spin systems (i.e. one pattern stemming from the backbone NH, the other from the side-chain $N^{\varepsilon}H$).

XREF2D - cross-referencing patterns

XREF2D is a program that computes the inter-pattern-correlation matrix and, optionally, reshuffles the patterns in order to enable the gathering of subsequent patterns into uninterrupted stretches by *ASSI2D* (Kleywegt et al., 1989). Before the program constructs S_{2DNOE}, the user may opt to filter out all 2D NOE peaks that can be assigned to intra-residue $C^{\alpha}H$-NH and $C^{\beta}H$-NH contacts as well as to sequential NH(i)-NH($i+1$), $C^{\alpha}H(i)$-NH($i+1$) and $C^{\beta}H(i)$-NH($i+1$) contacts (provided that some assignments have already been made), so as to prevent multiple use of the NOE peaks. The inter-pattern-correlation matrix is called $C_p(i,j)$ and is defined as the number of channels of pattern P_j that are either equal to the value of NROOT of P_j or lie in a user-defined range L_{xl} - L_{xu} (typically, the region for the $C^{\alpha}H$ and the $C^{\beta}H$ protons), that have a cross peak in the 2D NOE spectrum with the root channel of pattern P_i. Thus, the higher the value of $C_p(i,j)$, the more likely it is that P_j corresponds to a residue that is the N-terminal ('$i-1$') neighbour of the residue that corresponds to P_i. In other words, if $\boldsymbol{B}(R_k) = P_i$, then it is likely that $\boldsymbol{B}(R_{k-1}) = P_j$ if $C_p(i,j)$ is 'large'. Optionally, the user may require the values of $C_p(i,j)$ to be set to zero unless the amide channels (NROOT) of patterns P_i and P_j have a cross peak in the 2D NOE spectrum. This may be useful in the first few assignment cycles in order to get assignments for well-defined α-helical stretches in the protein (which are characterised by strong sequential NH-NH contacts).

Once the matrix C_p has been computed, the patterns may be shuffled (Kleywegt et al., 1989). This entails altering the order of the patterns so as to bring patterns that have many NOEs next to each other. More precisely, if $C_p(i,j)$ is large, and $C_p(j,i)$ is not, then the order is changed such that if P_j becomes P_k in the new order, then P_i will become P_{k+1}. The program starts shuffling at the first unassigned pattern (to ensure that any assigned patterns that precede it remain unaffected). The program will look for the pair of patterns P_k, P_l which has the highest value of $C_p(l,k)$ of all remaining patterns. This pair will then 'bubble up'. Subsequently, the program will search for a pattern P_m that has the highest correlation with P_l, in other words: $C_p(m,l)$ exceeds some user-defined threshold and $C_p(l,m)$ is not larger than one (the implied rationale is that one tends to find at most one ($i,i-1$)-contact, namely the NH-NH). If there exists such a pattern P_m, it will in turn bubble up, and the program starts looking for a pattern P_n that is sufficiently correlated with P_m.

If the search is unsuccessful, the program will seek a new pair of highly correlated patterns, etc. After the shuffling process, the reordered patterns and the correspondingly reordered matrix C_p are stored.

ASSI2D - generating assignments

ASSI2D is the program that integrates all information (the patterns, their correlation matrix and the amino acid sequence) and uses this to suggest assignments that are consistent with the available data and satisfy all constraints imposed by the user.

ASSI2D will first retrieve the statistical information regarding the expected values for the chemical shifts of the various types of protons (Gross and Kalbitzer, 1988). The user may choose to use either the median or the average values of those chemical shifts. In addition, given a small positive value for the parameter Δ_s, some proton pairs which, in practice, are often indistinguishable will be 'contracted' into a single proton type. For instance, the C^{δ}H-protons of arginine have median shifts of 3.17 ppm (σ = 0.19) and 3.20 ppm (σ = 0.19) (Gross and Kalbitzer, 1988). If $\Delta_s \geq 0.03$ ppm then these protons would be replaced by one generic C^{δ}H-proton. Its expected chemical shift would be the average of the two individual values and the standard deviation therein (σ) would be set to the greater of the two individual standard deviations. This option is intended to help reduce N_{exp}, the number of discernible protons in an amino acid's basic spin system, where a basic spin system (Kleywegt et al., 1989) contains all protons which, in principle, can be connected to the C^{α}H-proton(s) through J-couplings. This is useful in the matching process of patterns versus amino acid types, since it increases the likelihood that small patterns (i.e., patterns containing few channel numbers) could correspond to larger spin systems.

Subsequently, the program may either read the inter-pattern-correlation matrix as created by *XREF2D* or generate one itself through use of the user-supplied PATNBS arrays. In the latter case, it creates a matrix in which a pattern is most correlated with all (if any) patterns that have the same NROOT (amide channel) as the first element of PATNBS, a bit less correlated to those patterns that have their NROOT equal to the second element of PATNBS, etc.

Then the program will match patterns to amino acid residue types, resulting in a matrix Q, where Q (i, j) is a measure for the likelihood that pattern P_i corresponds to a residue of type j (i = 1, N_p; $j \in$ {Ala, Arg, ..., Val}; $0 \leq Q(i,j) \leq 1$). If a pattern has already been assigned, then the corresponding entry in the matrix Q will be set to one and all the others (for this particular pattern) to zero. As yet unassigned patterns will be subjected to the matching procedure with every amino acid type which:

- is not a proline;
- occurs in the pattern's PATCBR array;
- still occurs in the set of unassigned residues of the protein;
- has a suitable expectation value for its amide proton, in other words: channel NROOT of the pattern under scrutiny, when converted to the ppm-scale, lies within σ_{max} standard deviations of the expectation value of the chemical shift for this NH.

If an amino acid type satisfies all these criteria, then the actual matching of channel numbers (of the pattern) and protons (of the particular amino acid type) will commence. A channel d_i is matchable to a proton p_j, if the ppm-value δ_i of this channel lies within σ_{max} standard deviations σ_j of the expected chemical shift ε_j of the proton:

$$\Delta_{ij} = |\delta_i - \varepsilon_j|/\sigma_j \leq \sigma_{max} \tag{12}$$

When all unused channels have been matched against a proton, the one (if any) that has the lowest deviation will be tentatively 'assigned' to it (which means that this particular channel cannot be matched to any subsequently considered protons). Once this has been done for all protons of a specific amino acid type, the total number of 'assignable' protons is N_{ass}. This number must be at least equal to M_d: when matching a pattern P_i with an amino acid type j, M_d is taken as the lower value of M_p, a user-defined number, $N_c(i)$, the number of channels in the pattern, and $N_{exp}(j)$, the number of discernible protons in the amino acid's basic spin system. If this condition is not met, $Q\ (i,j)$ is set to zero; otherwise the average chemical-shift deviation is computed and only if this is not larger than σ_{ave} will the pattern be considered 'matchable' to this particular amino acid type. If this is the case, a matching score will be computed, the precise value of which depends on the selected score option. We have tested several scoring formulas, but the ones that tend to give the best results are:

$$Q\ (i,j) = N_{ass} \ / \ M_e \tag{13}$$

and

$$Q\ (i,j) = N_{ass} \ / \ M_f \tag{14}$$

where M_e equals the lower and M_f equals the higher value of $N_c(i)$ and $N_{exp}(j)$.

If the user has requested this, the next step entails grouping highly correlated neighbouring patterns into so-called fragments (the alternative is to consider each pattern as a separate fragment). If a pattern P_i is the first of a new fragment, then its neighbour P_{i+1} will be added to the fragment if the following conditions are met:

- if P_i has been assigned to a residue R_k, then P_{i+1} must have been assigned to R_{k+1}. In other words, uninterrupted stretches of assigned patterns/residues are collected into one fragment;
- if P_i has not yet been assigned, then NROOT of P_{i+1} must be unique (given a certain tolerance). If this is not the case (i.e. if there are other patterns with virtually the same amide channel), then none of these patterns can unambiguously be assumed to be the correct $(i+1)$-neighbour of P_i;
- if P_i has not yet been assigned, then the following requirements must be met as well:

$$C_p\ (i+1,i) \geq M_{c1} \tag{15}$$

$$C_p\ (i,i+1) \geq M_{c2} \tag{16}$$

$$C_p\ (i,i+1) \leq M_{c3} \tag{17}$$

This implies that there must be at least M_{c1} $(i,i+1)$-contacts and at least M_{c2} $(i,i-1)$-contacts, but no more than M_{c3}. This procedure yields a set of N_{frag} fragments F_i $(1 \leq N_{frag} \leq N_{res})$:

$$F_i \leftrightarrow \{P_{1+j}\}\ ; j = 0,\ N_f(i)-1 \tag{18}$$

where $N_f(i)$ is the number of patterns contained in fragment F_i.

Subsequently, for each fragment a list is generated that contains all possible C-terminal neighbouring fragments. A fragment is a potential C-terminal neighbour of another fragment, if the first pattern of the former and the last pattern of the latter satisfy the requirements stated above with respect to their correlations C_p.

The next step aims to find all possible matchings of each fragment onto the amino acid sequence of the protein. A fragment $F_i \leftrightarrow \{P_{l+j}\}$ ($j = 0, N_f(i)-1$) can be matched to a stretch of residues $\{R_{k+j}\}$ ($j = 0, N_f(i)-1$) if the following conditions are met:

- $k > 1$ (the rationale for this is that the amino protons of the first residue are usually not detected);
- $k + N_f(i) - 1 \leq N_{res}$;
- if any pattern P_{l+j} in the fragment has already been assigned, it must be to residue R_{k+j} (or non-specifically to the same residue type as R_{k+j});
- the assignment probability for each pattern and the corresponding residue type must exceed a certain threshold: $Q(P_{l+j},R_{k+j}) \geq M_Q$ ($j = 0, N_f(i)-1$).

The matching score $Q_F(i,k)$ is set to zero if these conditions are not met and to $\sum_j Q(P_{l+j},R_{k+j})$ ($j = 0, N_f(i)-1$) otherwise.

Finally, the program will generate all assignments that are consistent with all the data and that satisfy all the constraints imposed by the user (through the various parameters). This proceeds in a recursive depth-first search procedure which exhaustively yields all solutions. *ASSI2D* will use all fragments that contain at least M_{len} patterns and will consider all their possible matchings onto the protein. For each successful match it will check all possible neighbouring fragments to see if they actually match onto the C-terminal end of the former fragment. In other words, if fragment F_i fits at position R_k in the protein, *i.e.* $Q_F(i,k) > 0$, then it will find all potential neighbour fragments F_j for which $Q_F(i,k+N_f(i)) > 0$. With each such fragment (if any) this process is repeated for all its potential neighbours, etc. This will continue until:

- a fragment has no potential neighbours;
- none of the possible neighbours fits next to a fragment onto the protein;
- a proline residue is encountered;
- the end of the amino acid sequence has been reached;
- a suitable fragment is encountered which is already a part of the current stretch.

When one or more of these conditions are met, the program will check whether the assignment is a subset of one that was generated earlier. If this is not the case and if the assignment comprises at least M_{ass} residues it will be written to a file.

It should once again be stressed that *ASSI2D* does not actually assign any patterns. It is up to the user to judge the merits of the plausible assignments and to effectuate them (both can be accomplished through use of *CONPAT*).

CHECKA - the finishing touch

CHECKA is a program that executes a variety of tasks that help in applying the finishing touch to the assignment process. The various tasks are:

Assigning individual resonances. For each assigned residue, the program will compute and print the matrix Δ_{ij} according to Eq. (12). In addition, for each of the as yet unassigned protons *CHECKA* will find the best-matching unused channel number in the pattern and assign the proton to it. This is a purely chemical-shift-based assignment which usually gives fairly reasonable results nevertheless. Naturally, the user can override these assignments easily (through *CONPAT*) on the basis of information that is not available to the program (for instance, a COSY spectrum).

Assigning individual HOHAHA and 2D NOE peaks. For each of the HOHAHA and 2D NOE peaks, *CHECKA* may print all possible assignments. Optionally, all peaks for which no assignments can be generated (yet) may be written to new files.

Extending some patterns. CHECKA may be instructed to try and find candidate channels for protons that are not part of the basic spin system of assigned residues of the types Tyr, Phe, Trp, His, Arg, Lys, Asn and Gln. For each unassigned proton the program defines a target-channel range that corresponds to the ppm-range from $(\varepsilon_j - \sigma_{max} * \sigma_j)$ to $(\varepsilon_j + \sigma_{max} * \sigma_j)$. Subsequently, the program will simply count the number of cross peaks in the 2D NOE data set that each of the channels in this range has with the channels that are already part of the pattern. The best-scoring channels are sorted and printed. It is left to the user to assess and implement the results of this operation.

Searching for prolyl spin systems. Another option involves the quest for resonances that belong to prolyl spin systems. In order for this to work, at least one neighbouring residue (preferably the N-terminal one) of a proline residue must have been assigned. *CHECKA* will once again generate candidate channels, first for the $C^{\delta}H$-protons by looking for cross peaks with the NH, $C^{\alpha}H$, $C^{\beta}H$ and $C^{\gamma}H$-protons of the previous residue (if they exist and have been assigned) and, secondly, for the $C^{\alpha}H$, $C^{\beta}H$ and $C^{\gamma}H$ prolyl protons, by looking for cross peaks with the amide proton of the C-terminal residue (if it exists and has been assigned). Candidate channels are printed; it is up to the user to assess this information (*CONPAT*).

APPLICATION TO PHORATOXIN B

Introduction

Phoratoxin is a member of a family of homologous toxic proteins that occur in European and American mistletoes which also comprises ligatoxin and several viscotoxins. Furthermore, all these proteins display homology to crambin and the thionins. There are two forms of phoratoxin, called A and B. The amino acid sequence of phoratoxin B is shown in Fig. 10 (Mellstrand and Samuelsson, 1974; Thunberg, 1974). In phoratoxin A, Ile^{25} is replaced by a Val residue (Mellstrand and Samuelsson, 1974; Thunberg, 1974). The Asp at position 45 in phoratoxin B, which is

```
  1               5                  10
Lys Ser Cys Cys Pro Thr Thr Thr Ala Arg

                 15                  20
Asn Ile Tyr Asn Tyr Cys Arg Phe Gly Gly

                 25                  30
Gly Ser Arg Pro Ile Cys Ala Lys Leu Ser

                 35                  40
Gly Cys Lys Ile Ile Ser Gly Thr Lys Cys

                 45
Asp Ser Gly Trp Asp His
```

Fig. 10. The amino acid sequence of phoratoxin B.

substituted by an Asn in phoratoxin A, is probably an artifact due to deamination during the isolation of the protein (Thunberg, 1974). Phoratoxin is well amenable to NMR studies (Lecomte et al., 1987; Clore et al., 1987), complete resonance assignments as well as a solution structure are available (Clore et al., 1987). Phoratoxin was therefore deemed an interesting object for testing our assignment software.

Experimental procedures

Phoratoxin B belonged to a batch previously described (Lecomte et al., 1987). For the NMR measurements 25 mg was dissolved in 450 μl D_2O/H_2O (1:19) at pH 5.4, the same sample was lyophilized and dissolved in 450 μl D_2O. All 2D NMR spectra were recorded on a Bruker AM500 spectrometer operating at 500 MHz, interfaced with an Aspect 3000 computer. The data were processed on a μVAX II computer with the *TRITON* software package (written in FORTRAN-77) developed in our laboratory.

The 2D HOHAHA spectrum in H_2O was recorded at 304 K using an MLEV-17 pulse sequence of 34.5 ms with a 60° 17th pulse and sandwiched between two trim pulses of 3 ms (Bax and Davis, 1985). The H_2O signal was suppressed by irradiation during the relaxation delay (1.4 s). TPPI was used for the t_1-amplitude modulation (Marion and Wüthrich, 1983). A total of 400 free-induction decays (FIDs) of 2K data points, 128 scans each, were collected. The t_1 period was incremented from 40 μs to 32 ms. The time-domain data in both dimensions were weighted with a sine-bell function shifted by $\pi/3$. The data were processed to yield a phase-sensitive spectrum of 1K × 1K data points with a digital resolution of 6.1 Hz/point and were baseline corrected in the ω_1 dimension.

The 2D NOE spectra in H_2O were recorded at 304 K with a 32-step phase cycle (States et al., 1982). The H_2O signal was suppressed by irradiation in the relaxation delay followed by a SCUBA pulse sequence (Brown et al., 1988). The NOE mixing period contained a 180° pulse in the centre, followed by a homospoil pulse. In order to obtain a good match between the 2D NOE and HOHAHA spectra, the sum of the NOE mixing time (100, 200 or 300 ms) and the relaxation delay was kept the same as the relaxation delay of the 2D HOHAHA experiment (1.4 s). Furthermore, the relaxation delay of the 2D NOE experiments was preceded by an (x y −x −y) train of 2 ms pulses and of 40 ms duration at the same power level as the HOHAHA sequence. TPPI was used for the t_1 amplitude modulation. A total of 400 FIDs of 2K data points, 64 or 96 scans (100 ms) each, were collected in the same session as the 2D HOHAHA experiment. The data were processed in a similar fashion as the 2D HOHAHA spectrum.

The DQF-COSY spectra were recorded using the pulse sequence and phase cycling devised by Rance et al. (1983). The suppression of the H_2O signal in the H_2O spectrum was done by irradiation followed by the SCUBA pulse sequence. In order to suppress the observation of double-quantum coherences, which were present in the t_1 evolution period due to rapid pulsing, a delay was inserted prior to the relaxation delay which was varied randomly from one t_1 increment to another between 0 and 100 ms (Vermeulen et al., 1987). TPPI was used for the t_1 dimension and 650 FIDs of 2K data points, 64 scans each, were acquired. The t_1 period was incremented from 40 μs to 52 ms. The data were processed in a similar fashion as described above.

The 2D NOE spectrum in D_2O was recorded in a traditional fashion with a 32-step phase cycle, HOD irradiation in both the relaxation delay and the mixing period, and TPPI. The NOE mixing time was 200 ms, 400 FIDs were recorded, 64 scans each. The data were processed as described above.

The NMR spectra that were used in the assignment process are listed in Table 2. Peakpicking was carried out with our software package *STELLA* (Kleywegt et al., 1990). For the HOHAHA spectrum this resulted in a set of 667 peaks, the 2D NOE spectrum yielded 1730 peaks.

Pattern detection, filtering and selection

SPIN2D was used to detect patterns starting from the amide-aliphatic region. We used the following set of parameters: $\tau_A = 1.01$ channels, $\tau_R = 2.01$, $M_{it} = 3$, $M_c = 8$, $M_{xl} = 2$, $M_{xu} = 4$ and $M_2 = 2$. This operation yielded a total of 508 patterns. *FILPAT* was run with values for $M_{pa} = 0.1$, $M_{ra} = 1.0$, $D = 1$ and $M_N = 3$. After this shake-out, only 92 patterns remained, 88 having been deleted due to insufficient 'pattern scores' and 328 due to them being a fuzzy subset of one or more other patterns. The surviving patterns contained two to eight channels (5.1 on average). The next task involved editing these patterns. This was accomplished with *CONPAT* using a variety of spectra (see Table 2). First, patterns were merged, extricated, added and deleted as necessary to make them correspond to spin systems (in the view of the user). After this operation, 41 patterns remained which contained 3.8 channels on average. Secondly, the channels of the patterns were made to refer to the appropriate 2D NOE spectrum ($\tau_m = 200$ ms). Thirdly, a contour plot of the NH-C$^\alpha$H-region of the COSY spectrum was produced (on-screen), which enabled the identification of three cross peaks which were not included in the set of 41 patterns (by having *CONPAT* plot boxes at all cross-peak positions in the appropriate area that can be explained). Use of the scroll-contour option with various spectra enabled us to derive a pattern from each of these peaks (one containing five channels, the other two containing four channels).

Phoratoxin has the favourable property that it contains one specimen of each of the four types qf aromatic residue. It was relatively straightforward to characterise the aromatic resonances of the tyrosine (Tyr^{13}) and phenylalanine (Phe^{18}), as well as some resonances belonging to the tryptophan (Trp^{44}). Using the scroll-contour and pattern-contour options, we could identify the corresponding aliphatic resonances of each of these residues as well as the patterns to which they belonged. No cross peaks were observed for the ring protons of His^{46}; one was found during the assignment process and another was located afterwards in a 1D NMR spectrum recorded in D_2O.

By comparing contour plots of the aromatic region in the H_2O COSY and the D_2O COSY, we

TABLE 2
OVERVIEW OF NMR SPECTRA USED FOR ASSIGNING PHORATOXIN B

Experiment	Solvent	Matched [a]	Used by *CLAIRE* [b]
COSY	H_2O		
COSY	D_2O		
HOHAHA	H_2O	*	*
2D NOE	D_2O		
2D NOE (τ_m = 100 ms)	H_2O	*	
2D NOE (τ_m = 200 ms)	H_2O	*	*
2D NOE (τ_m = 300 ms)	H_2O	*	

[a] Marked spectra were recorded in such a manner that they are mutually well-aligned (see text for details).

[b] Marked spectra were peakpicked and used by all *CLAIRE* programs; the unmarked spectra were used only through program *CONPAT*.

could identify the cross peaks of the side-chain amino protons of the two asparagyl residues (phoratoxin contains no glutamines). Each of these could once again be linked to one of the existing patterns. In one case, this even yielded an additional C$^{\beta}$H-channel for a pattern which had previously consisted of only an NH, a C$^{\alpha}$H and one C$^{\beta}$H-channel.

Subsequently, we used *XPAT2D* (a program that attempts to extend patterns) to find possible additional channels for each of the patterns. This was successful for ten patterns yielding one, two or, in one case, three new channel numbers for each of them. *FILPAT* was rerun and identified two patterns which had no less than six 'fuzzy similarities'. These patterns were merged and identified as corresponding to an arginine. Since both C$^{\delta}$H-protons were contained in the pattern, it was possible to decide which channel number corresponded to the side-chain N$^{\varepsilon}$H-proton and which to the backbone NH-proton. At this stage, we had 43 patterns left which, on average, contained 4.5 channel numbers.

Before proceeding with the assignment, we input some extra information with respect to the identity of some of the patterns. Naturally, three out of four aromatic residues as well as both asparagines had already been characterised. In addition, we could identify six glycines, three threonines and one arginine. For most other patterns, the contents of their PATCBR arrays could be reduced to include only a few possible amino acid residue types (from one to eleven, averaging about four).

Assignment

The flow of the iterative assignment process is shown in Table 3. The first two assignment cycles will be described with a fair amount of detail in the following in order to exemplify the general modus operandi.

The first run of *XREF2D* and *ASSI2D* (in which the obligatory NH-NH-contact option was used) yielded two assignments, one stretching from Thr^{8} to Phe^{18} (8 fragments), the other from Thr^{8} to Cys^{16} (6 fragments). These two propositions were identical for Thr^{8} to Thr^{15}, an assignment that looked convincing both in terms of patterns versus amino acid-type matches and in terms of inter-pattern NOE contacts. Both suggestions contained a different pattern to account

TABLE 3
OVERVIEW OF THE ASSIGNMENT OF PHORATOXIN B SPECTRA USING THE *CLAIRE* SOFTWARE PACKAGE

Cycle	Assigned residues				Remarks
1	Thr^{8} → Phe^{18}				-
2	Gly^{19} → Gly^{21}	Ser^{42} → Asp^{45}	Gly^{31} → Lys^{33}	Gly^{37}	-
3	Ile^{34} → Ile^{35}	Thr^{38} → Lys^{39}			-
4	Cys^{40} → Asp^{41}				-
5	His^{46}	Ser^{2} → Cys^{4}			a
6	Ser^{22} → Arg^{23}				b
7	Ile^{25} → Ser^{30}				c

[a] Pro^{5} assigned using *CHECKA* and *CONPAT*.

[b] Pro^{24} assigned using *CHECKA* and *CONPAT*.

[c] Ile^{25} through Ser^{30} assigned manually (using *CONPAT*).

for Cys^{16}, but the first, in addition, comprised an assignment for a long side-chain type of pattern to Arg^{17} and the preassigned pattern corresponding to Phe^{18}, as well as convincing sequential NOE contacts. This assignment was therefore accepted and effectuated.

A second run of *XREF2D* (with omission of all used intra-residue and sequential NH-NH, C$^{\alpha}$H-NH and C$^{\beta}$H-NH peaks) and *ASSI2D* yielded no new assignments. The obligatory NH-NH-contact option was therefore switched off and both programs were run again, yielding three new assignment suggestions. Interestingly, the first of these, involving Gly^{19} and Gly^{20}, was actually wrong but, nevertheless, put us onto the right track. When the sequential contacts between Phe^{18} and the proposed Gly^{19} were investigated (using *CONPAT*), it was observed that the actual NH-resonance of Gly^{19} had to be three channels away from that of the proposed pattern. In the set of Gly patterns, there were two which had their NH-resonances close together, at channel 267 and 270, respectively. The one with its NH at channel 270 (which had intense NOE contacts with resonances of Phe^{18}), however, was 'tied up' as the second pattern in a fragment and was therefore unavailable as a candidate for assignment to Gly^{19}. This 'narrow escape' demonstrates that there are risks involved in dividing the set of patterns into fragments, since the underlying heuristic ('patterns with many mutual cross peaks probably correspond to neighbouring residues') may not always be valid. This may be a result of overlap, of non-sequential NOE contacts (for example, long-range contacts with aromatic ring protons), of medium-range contacts (e.g. in α-helices), of long-range contacts such as occur in β-sheets, and/or of inaccuracies in peak positions and the use of tolerances. In this case, the Gly pattern with its NH at channel 270 was assigned to Gly^{19}. When this pattern was inspected (*CONPAT*), we noticed that both intra-residue C$^{\alpha}$H-NH cross peaks had shoulders at channel 267. Since, in addition, the two suggested patterns for Gly^{19} and Gly^{20} were well connected through sequential NOE contacts, and since the only place left in the sequence with two adjacent glycyl residues was at 20-21, it was concluded that these two patterns corresponded to Gly^{20} and Gly^{21}.

The second assignment covered Ser^{42} to Asp^{45}; the pattern suggested for Gly^{43} had previously been identified as a glycine and Trp^{44} had already been assigned. The other two patterns looked like proper AMX spin systems and the entire stretch was characterised by convincing sequential contacts. The suggested assignment was therefore implemented. The third assignment comprised Gly^{31} to Lys^{33}; the pattern suggested for Gly^{31} was indeed a glycine and the one for Lys^{33} had been identified as a long side-chain spin system. Since the sequential contacts were in order, this assignment was also implemented. At this stage, there was only one glycine left to assign (Gly^{37}) and only one pattern left which had been identified as a glycine; it was therefore assigned to Gly^{37}.

XREF2D and *ASSI2D* were rerun several times in order to obtain assignments for the rest of the protein (see Table 3). *CHECKA* was used twice in order to detect resonances belonging to the two prolyl residues.

In the end, residues 25 to 30 had to be assigned manually (using *CONPAT*). The fact that this particular stretch of the protein was difficult to assign is due to the fact that the sample also contained some phoratoxin A. The difference between both proteins lies in residue 25 (Ile versus Val), but this substitution also affects the chemical shifts of the protons (particularly, the amide protons) of residues 26 to 29 (Clore et al., 1987).

The obtained assignments (the mapping *A*) are listed in Table 4. When comparing these assignments with those published by Clore et al., one observes that the major differences lie in the extent to which long side-chain type spin systems have been detected. For ten residues, Clore et al. as-

TABLE 4
ASSIGNMENTS OBTAINED WITH *CLAIRE* FOR PHORATOXIN B

	Chemical shift (ppm)				
Residue	NH	$C^{\alpha}H$	$C^{\beta}H$	$C^{\gamma}H$	Others
Ser[2]	8.93	5.18	3.73,3.09		
Cys[3]	9.35	5.24	4.34,2.67		
Cys[4]	8.35	5.42	2.84		
Pro[5]		4.37	1.57	1.84	$C^{\delta}H$ 3.56,3.90
Thr[6]	6.51	4.12	4.48	1.08	
Thr[7]	8.82	3.99	4.20	1.28	
Thr[8]	7.68	3.97	4.05	1.27	
Ala[9]	8.27	4.20	1.58		
Arg[10]	7.47	4.53	2.22,1.94	2.05,1.72	$C^{\delta}H$3.49,3.39 $N^{\varepsilon}H$ 9.33
Asn[11]	8.18	4.63	3.07,3.01		$N^{\delta}H_2$ 7.64,6.93
Ile[12]	8.36	3.71	1.93	$C^{\gamma1}H$ 1.14,0.91 $C^{\gamma2}H$ 0.88	
Tyr[13]	9.01	3.72	3.67,3.22		$C^{\delta}H$ 6.98 $C^{\varepsilon}H$ 6.81
Asn[14]	9.12	4.22	3.12,2.77		$N^{\delta}H_2$ 7.81,7.00
Thr[15]	8.45	3.98	4.28	1.29	
Cys[16]	8.23	4.20	3.36,3.01		
Arg[17]	8.58	4.09	1.84,1.69	1.24,1.01	$C^{\delta}H$ 3.15,2.56
Phe[18]	8.82	4.45	3.31		$C^{\delta}H$ 7.31 $C^{\varepsilon}H$ 7.40 $C^{\zeta}H$ 7.39
Gly[19]	7.73	4.27,3.86			
Gly[20]	7.77	4.44,3.72			
Gly[21]	8.23	4.11,3.43			
Ser[22]	8.51	4.41	4.27,4.11		
Arg[23]	9.27	4.10	2.04,1.90	1.69	$C^{\delta}H$ 3.26
Pro[24]		4.37	2.39,1.89	2.17,2.04	$C^{\delta}H$ 3.92,3.81
Ile[25]	7.14	3.92	2.04	$C^{\gamma2}H$ 0.90	
Cys[26]	8.72	4.76	2.76,2.40		
Ala[27]	9.45	4.05	1.60		
Lys[28]	7.69	4.19	1.60		
Leu[29]	8.13	4.20	1.78,1.67	0.91	
Ser[30]	7.73				
Gly[31]	8.22	4.44,3.98			
Cys[32]	8.12	5.19	3.09,2.52		
Lys[33]	9.17	4.69	1.55,1.49	1.22,1.11	
Ile[34]	8.62	4.70	1.91	$C^{\gamma2}H$ 0.78	
Ile[35]	8.68	4.87	2.05	$C^{\gamma1}H$ 0.97 $C^{\gamma2}H$ 0.78 $C^{\delta}H$ 0.53	
Ser[36]	8.77	4.58	3.92,3.88		
Gly[37]	7.64	4.36,4.00			
Thr[38]	8.14	4.38	4.38	1.22	
Lys[39]	7.81	4.64	1.75,1.63	1.39	
Cys[40]	9.00	4.52	3.09,2.65		
Asp[41]	8.72	4.52	2.78,2.34		
Ser[42]	8.57	4.26	3.93		
Gly[43]	8.94	4.41,3.93			
Trp[44]	8.39	4.56	3.56,2.92		$C^{\delta1}H$ 7.24 $N^{\varepsilon1}H$ 10.21 $C^{\varepsilon3}H$ 7.45 $C^{\zeta2}H$ 7.61 $C^{\zeta3}H$ 7.55 $C^{\eta2}H$ 7.35
Asp[45]	8.23	4.87	2.56,2.48		
His[46]	7.90	4.42	2.92,2.50		$C^{\delta2}H$ 7.00 $C^{\varepsilon1}H$ 8.51

signed more protons than we did (not counting degenerate assignments): Lys^{1} (3 extra), Arg^{23} (2), Ile^{25} (2), Lys^{28} (6), Leu^{29} (1), Ser^{30} (3), Lys^{33} (4), Ile^{34} (3), Thr^{38} (1) and Lys^{39} (2). In five cases, we found one assignment that Clore et al. did not list: Ile^{12}, Tyr^{13}, Phe^{18}, Trp^{44} and Asp^{45}. It is worth noting, that apart from these relatively minor differences there are no major discrepancies between the two assignments.

In conclusion, we wish to stress again that the present method is not a fully automatic one. Rather, it provides the user with a number of useful tools that assist in the assignment process. The end result is promising – considering that it was basically achieved by a worker without much previous exposure to protein NMR and without the aid of any paper plots. Hence, in the hands of an experienced spectroscopist, who may derive additional information, for instance, from spectra at different conditions etc., *CLAIRE* should be a powerful tool. Program *CONPAT* by itself is of more general use, e.g. to workers who do not wish to use any of the other *CLAIRE* programs and to those who work on other types of biopolymers than proteins. By organising spin systems in terms of patterns, bookkeeping is facilitated. In addition, *CONPAT* offers a number of very useful 'windows' on the original data. At present, further testing of the software on a 'real unknown' is underway, which might lead to minor adjustments and/or extensions. Simultaneously, we are attempting to extend the approach to the interpretation of 3D NMR spectra of proteins.

ACKNOWLEDGEMENT

The authors thank Prof. G. Samuelsson for the sample of phoratoxin B.

REFERENCES

Bax, A. and Davis, D.G. (1985) *J. Magn. Reson.*, **65**, 355-360.

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

Brown, S.C., Weber, P.L. and Mueller, L. (1988) *J. Magn. Reson.*, **77**, 166-169.

Catasti, P., Carrara, E. and Nicolini, C. (1990) *J. Comput. Chem.*, **11**, 805-818.

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

Clore, G.M., Sukumaran, D.K., Nilges, M. and Gronenborn, A.M. (1987) *Biochemistry*, **26**, 1732-1745.

DiStefano, D.L. and Wand, A.J. (1987) *Biochemistry*, **26**, 7272-7281.

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

Englander, S.W. and Wand, A.J. (1987) *Biochemistry*, **26**, 5953-5958.

Feng, Y., Roder, H., Englander, S.W., Wand, A.J. and DiStefano, D.L. (1989) *Biochemistry*, **28**, 195-203.

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

Kaptein, R., Boelens, R., Scheek, R.M. and Van Gunsteren, W.F. (1988) *Biochemistry*, **27**, 5389-5395.

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. and Kaptein, R. (1990) *J. Magn. Reson.*, **88**, 601-608.

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

Lecomte, J.T.J., Kaplan, D., Llinás, M., Thunberg, E. and Samuelsson, G. (1987) *Biochemistry*, **26**, 1187-1194.

Marion, D. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Comm.*, **113**, 967-974.

Meier, B.U., Bodenhausen, G. and Ernst, R.R. (1984) *J. Magn. Reson.*, **60**, 161-163.

Meier, B.U., Mádi, Z.L. and Ernst, R.R. (1987) *J. Magn. Reson.*, **74**, 565-573.

Mellstrand, S.T. and Samuelsson, G. (1974) *Acta Pharm. Suec.*, **11**, 347-360.

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

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

Rance, M., Sørensen, O.W., Bodenhausen, G., Wagner, G., Ernst, R.R. and Wüthrich, K. (1983) *Biochem. Biophys. Res. Comm.*, **117**, 479-485.

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

Thunberg, E. (1974) *Acta Pharm. Suec.*, **20**, 115-122.

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

Van de Ven, F.J.M., Lycksell, P.O., Van Kammen, A. and Hilbers, C.W. (1990) *Eur. J. Biochem.*, **190**, 583-591.

Vermeulen, J.A.W.H., Lamerichs, R.M.J.N., Berliner, L.J., DeMarco, A., Llinás, M., Boelens, R., Alleman, J. and Kaptein, R. (1987) *FEBS Lett.*, **219**, 426-430.

Wand, A.J. and Nelson, S.J. (1988) *Trans. Amer. Cryst. Ass.*, **24**, 131-144.

Wand, A.J. , DiStefano, D.L., Feng, Y., Roder, H. and Englander, S.W. (1989) *Biochemistry*, **28**, 186-194.

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

Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311-319.

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

Wüthrich, K. (1989) *Science*, **243**, 45-50.

APPENDIX

List of symbols

For ease of reference, some logical symbols that are used in the text and some of the most often used symbols that pertain to the assignment of protein spectra have been collected here:

Symbol	Meaning
$\{\ \}$	denotes a set
$\in$	'is an element of'
$\leftrightarrow$	'corresponds to'
$\exists$	'there exists'
$\mid$	'such that'
$\wedge$	logical 'and'
$\vee$	logical 'or'
$\{R_i\}\ ;\ i = 1, N_{res}\ ;\ R_i \in \{\mathrm{Ala}, ..., \mathrm{Val}\}$	symbolic representation of a protein as an ordered set of N_{res} amino acid residues R_i
$R_i \leftrightarrow \{p_{ij}\}\ ;\ j = 1, N_p(i)$	symbolic representation of a residue Ri as a set of $N_p(i)$ protons p_{ij}
$\{X_{yi}\}\ ;\ i = 1, N_{pks}(y)\ ;\ y \in \{\mathrm{HOHAHA, 2D\ NOE}\}$	symbolic representation of a data set consisting of $N_{pks}(y)$ peaks from a spectrum of type y
$X_{yi} \leftrightarrow (c_{1yi}, c_{2yi}, I_{yi})$	symbolic representation of a peak in a spectrum of type y, characterised by two coordinates (c_{1yi}, c_{2yi}) and an intensity I_{yi}
$\{P_i\}\ ;\ i = 1, N_{pat}$	symbolic representation of a set consisting of N_{pat} patterns
$P_i \leftrightarrow \{d_{ij}\}\ ;\ j = 1, N_c(i)$	symbolic representation of a pattern P_i containing $N_c(i)$ channel numbers d_{ij}
$\boldsymbol{B}(R_i) = P_j$	the mapping of residues R_i onto patterns P_j
$\boldsymbol{A}(p_{ik}) = d_{jl}$	the mapping of protons p_{ik} onto channels d_{jl}