*Journal of Biomolecular NMR*, 10 (1997) 207–219
*KLUWER/ESCOM*




J-Bio NMR 461

# Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques

David Croft[a], Johan Kemmink[a], Klaus-Peter Neidig[b] and Hartmut Oschkinat[a,c,*]

[a]*EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany*
[b]*Bruker Analytik GmbH, Silberstreifen, D-76287 Rheinstetten, Germany*
[c]*Forschungsinstitut für Molekulare Pharmakologie, Alfred Kowalke Strasse 4, D-10315 Berlin, Germany*

Received 18 December 1996
Accepted 28 April 1997

*Keywords:* Multidimensional NMR spectroscopy; Automated assignment; Protein NMR; Pattern recognition

**Summary**

One of the major bottlenecks in the determination of protein structures by NMR is in the evaluation of the data produced by the experiments. An important step in this process is *assignment*, where the peaks in the spectra are assigned to specific spins within specific residues. In this paper, we discuss a spin system assignment tool based on pattern recognition techniques. This tool employs user-specified 'templates' to search for patterns of peaks in the original spectra; these patterns may correspond to side-chain or backbone fragments. Multiple spectra will normally be searched simultaneously to reduce the impact of noise. The search generates a preliminary list of putative assignments, which are filtered by a set of heuristic algorithms to produce the final results list. Each result contains a set of chemical shift values plus information about the peaks found. The results may be used as input for combinatorial routines, such as sequential assignment procedures, in place of peak lists. Two examples are presented, in which (i) HCCH-COSY and -TOCSY spectra are scanned for side-chain spin systems; and (ii) backbone spin systems are detected in a set of spectra comprising HNCA, HN(CO)CA, HNCO, HN(CA)CO, CBCANH and CBCA(CO)NH.

## Introduction

Since protein structure determination by NMR (Wüthrich et al., 1982) has become an indispensable tool in biological laboratories, automation of the structure determination process has become an urgent necessity. In the context of genome projects, for example, structure determination on a large scale is required, which would not be practicable without a high degree of automation.

One of the most time-consuming steps is the spectral assignment procedure. This involves sequence-specific resonance assignment of the NMR signals, and the assignment of NOESY spectra, as a prerequisite for structure calculations. To accomplish this task, a set of multidimensional NMR spectra needs to be evaluated, viz. the so-called backbone experiments (see, for example, Kay et al. (1990b)), HCCH-TOCSY (Bax et al., 1990) and HCCH-COSY (Kay et al., 1990a; Ikura et al., 1991) experiments, and various versions of the X-filtered NOESY experiment (Fesik and Zuiderweg, 1988; Marion et al., 1989). Many attempts have been made to computerise this process, either by introducing 'electronic drawing boards' (see Kraulis (1989), Eccles et al. (1991) and Bartels et al. (1995)) or by applying combinatorial approaches to peak lists (see, for example, Cieslar et al. (1988), Eads and Kuntz (1989), Van de Ven (1990), Vuister et al. (1990), Billeter (1991), Kleywegt et al. (1991), Nelson et al. (1991), Oschkinat et al. (1991), Meadows et al. (1994) and Xu et al. (1994)).

The long-term aim of such developments is the construction of a system which can perform the complete assignment process automatically, with little or no manual interference. We assume that there are a number of fundamental criteria which must be taken into consider-

*To whom correspondence should be addressed.
*Abbreviations:* NOE, nuclear Overhauser effect; COSY, correlation spectroscopy; TOCSY, total correlation spectroscopy; NOESY, nuclear Overhauser enhancement spectroscopy; S/N, signal-to-noise ratio.

ation when designing the architecture of such a system. A major design criterion is that the assignment process would be supervised by a high-level system, able to make decisions based on the current state of the assignment. A set of modular tools should perform the basic assignment tasks. Each tool should deliver results weighted according to quality factors, and the supervisory algorithms should, in cases where there was uncertainty, take recourse to the original data in an iterative manner. These tools should be able to recognise baseline offsets, $t_1$-noise and small positional differences between resonances in different spectra; they should also be able to learn about the spectral resolution, the line widths of the peaks, S/N, and amplitude proportions within a data set *on the fly*. Most importantly, the system should be capable of working with spectra having a low S/N.

These demands on the system would require that the maximum possible amount of information, as contained in the spectra themselves, be retained at all stages. Traditionally, assignment procedures rely on peak lists extracted from the multidimensional spectra. This approach has a number of problems. Firstly, in spectra with low S/N or strong $t_1$-noise, a considerable amount of spurious data may be generated, which would make the successful application of any combinatorial approach difficult. Secondly, in crowded spectra containing significant spectral overlap, peaks tend to merge together, and it is difficult to interpolate correct cross-peak positions in areas where this had occurred. Hence a peak list obtained from these spectra can be incomplete, and very likely imprecise with regard to picked frequencies. For this reason, we assume that highly automated procedures should work iteratively with the original spectra by applying, for example, pattern recognition algorithms and line shape fitting techniques in the manner of Denk et al. (1985).

Recent experience shows that there is a basic set of spectra for the sequence-specific assignment of protein $^{1}H$, $^{13}C$ and $^{15}N$ signals from which a minimum combination can be chosen:

**a**: HCCH-COSY
HCCH-TOCSY
**b**: CBCANH
CBCA(CO)NH
**c**: HBHA(CBCA)NH
HBHA(CBCACO)NH
**d**: HNCA
HN(CO)CA
**e**: HNCO
HN(CA)CO

These spectra may all be recorded using the same ($^{13}C$, $^{15}N$-labelled) sample, or perhaps multiple samples, employing random fractional deuteration of nonexchangeable sites (Nietlispach et al., 1996) as appropriate. In all cases, the cross peaks indicate the presence of scalar couplings, and each individual combination or set of spectra reflects a particular pattern of couplings (with corresponding patterns of peaks) that can be found. One may search in all spectra simultaneously or in various subsets. These subsets correspond to different starting points for the assignment procedure. We may start by assigning as many spin systems as possible in set **a**, and then use these to direct our search in the backbone sets **d** and/or **e** using a knowledge of the sequence to guide us. Then, in the next cycle of assignment, we can use this information to constrain a new search in the side-chain experiments **a**, with relaxed searching conditions, e.g. lower threshold values. One can also use spectra from sets **b** and **c** to obtain a more reliable connection between side chain and backbone. This cyclic assignment procedure can be repeated as many times as required to iteratively improve the global assignment.

The choice of strategy for a given protein will depend on the complexity of the spectra and, to some extent, on the relaxation properties of the protein signals. Most effective are combinations which allow the correlation of three like chemical shifts to obtain sequential assignments and which provide, in addition, two or more chemical shifts which provide handles on the side chain. The combination of sets **a**, **b** and **e**, for example, fulfils these criteria, as backbone assignments can be obtained by correlating CO, $C^{\alpha}$ and $C^{\beta}$ resonances of neighbouring residues, and $C^{\alpha}$ and $C^{\beta}$ frequencies can be used to attach side-chain spin systems. For smaller proteins, sets **a** and **d** may be appropriate; in other cases, an individual combination may be chosen.

Extensions using multiplicity-edited experiments (Shaw et al., 1997) and amino acid specific backbone experiments (Dötsch et al., 1996a,b) (the two-dimensional form may suffice) may be incorporated as appropriate.

In this paper, we present one of the basic tools that could be incorporated into an automated assignment strategy fulfilling the above-cited demands for complete automation. It is a program which looks for *patterns* of peaks in multidimensional NMR spectra. The search invokes all possible arguments about line shapes, chemical shift ranges and amplitude ratios. This pattern search extends over a set of spectra, which is chosen according to the criteria discussed above. The results produced by the pattern search are intended to be used as the starting points for further steps in the assignment procedure, serving as a substitute for peak lists.

In the Results section, applications involving side-chain topologies and backbone connectivities will be discussed.

## Pattern search strategy

### *Overview*

The program searches for patterns of peaks and yields results (putative assignments), weighted by probability-

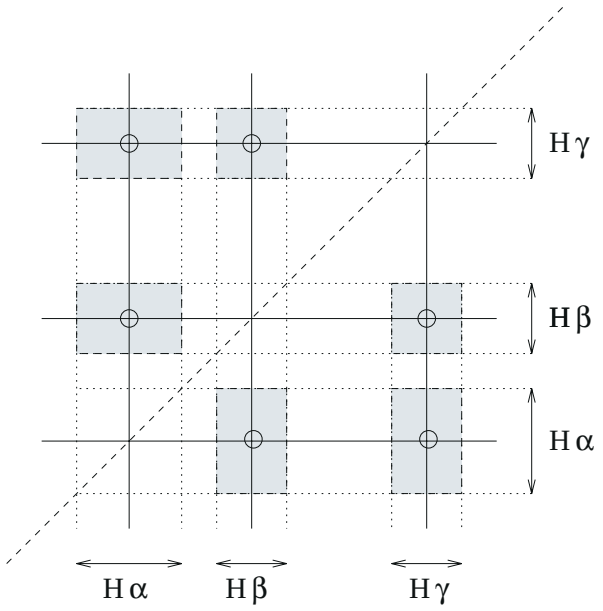



Fig. 1. Idealised peak pattern for threonine in a 2D TOCSY spectrum. The small circles indicate *which* peaks are expected, and the grey patches show the *areas* in which the program will search for them. These areas are called regions of interest. The chemical shift ranges for each spin type determine their boundaries. In this example, the regions of interest in a 2D spectrum for peaks between $H^{\alpha}$, $H^{\beta}$ and $H^{\gamma}$ are shown. For clarity, these regions have been shown as distinct patches, but in reality there will often be overlap.

like factors, which give estimates of correctness for the results. As input, the spectra and the pattern topology need to be fed into the program. The pattern topology describes which peaks are to be expected in the spectra; this information is obtained from a *pattern file*. Figure 1 shows an example of the peaks that one might expect to find for threonine in a 2D TOCSY experiment. For each expected peak in a pattern, a *region of interest* is determined by the pattern file. This is a rectangular or cuboid portion of a spectrum, whose size is bounded by the chemical shift ranges of the spins contributing to that peak.

For each pattern topology, the program steps through three principal operations. These are (i) peak emphasis filtering; (ii) pattern search; and (iii) heuristic filtering. This is shown schematically in Fig. 2. The three stages of execution will be described in more detail in the following sections. A separate Appendix, which can be obtained from the authors upon request, contains a mathematical treatment of the algorithms employed.

*Peak emphasis filtering*

The program takes appropriately processed data points directly from the spectra. These data are first filtered to emphasise genuine peaks, to distinguish different kinds of peaks and to de-emphasise features such as broad peaks, noise or variations in baseline. The filtering procedure works by convolving mask functions with the regions of interest in all spectra (see Schalkoff (1989)). The *mask response* is the output of the convolution algorithm for a single position within a region of interest. The mask functions are designed to fit ideal peaks and may be different for different spectra or even for different cross peaks within a pattern. The latter is necessary in cases where peaks with very different multiplet structures are present. This process generates a set of *convolved regions of interest*. These form the input for the pattern search algorithm. An example of how a 1D mask is employed in a real spectrum is shown in Fig. 3.

*Pattern search*

The program searches for patterns of peaks in the convolved regions of interest in a systematic way. It does this by incrementing the chemical shift values of each spin one data point at a time, and for each increment, examining the mask responses in each region of interest.

At all points where chemical shift coordinates intersect within regions of interest, the mask response values are taken, and added together. The resulting value is called a *score*. Each combination of chemical shifts obtained by these means represents a possible assignment; the calculated score value gives these 'assignments' a quantitative figure of merit, based on the combined mask response values. This is illustrated graphically in Fig. 4.

Such a pattern finding algorithm could potentially lead to a combinatorial explosion. The number of chemical shift combinations goes up approximately to the power of the number of distinguishable spins, which for a spin system such as lysine is large.

By prudent use of *thresholding*, however, the branches of the search 'tree' can be severely pruned, with a corresponding increase in execution speed. When an assignment is first generated, the program compares each mask

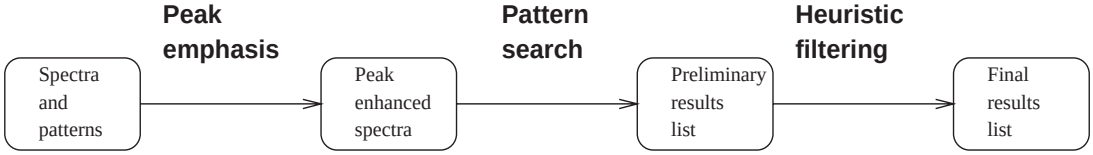



Fig. 2. Program flow diagram. The input required by the program consists of *spectra* and *pattern files*. *Peak emphasis* is a filtering operation which emphasises genuine peaks in selected portions of the spectra (regions of interest). *Pattern search* exhaustively explores the chemical shift space for peak patterns within the regions of interest (see also Fig. 4). A *heuristic filter* applies rules typically used by spectroscopists in order to penalise or remove unlikely results.

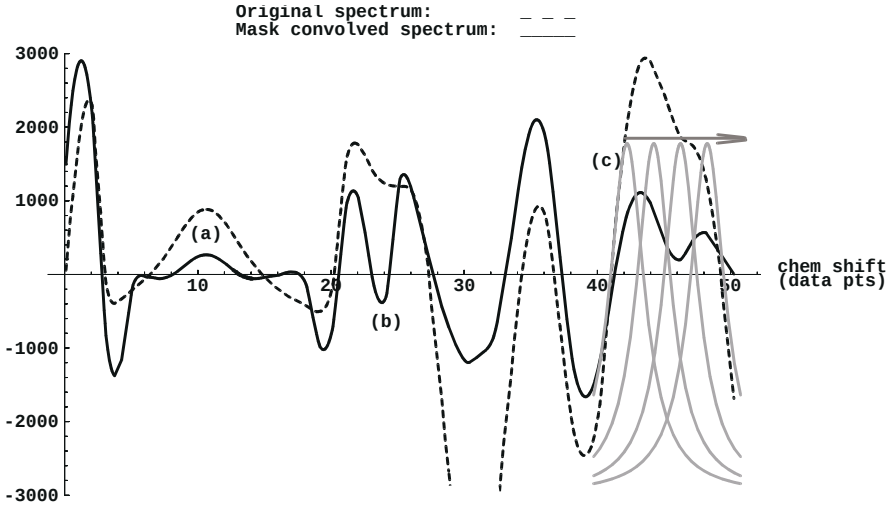



Fig. 3. Convolution of a spectrum with peak detecting mask. Two properties of this convolution are worth noting. At position (a), an artificial broad peak has been inserted into the spectrum. After convolution with a narrow mask, this peak has been significantly flattened. At position (b), two peaks in the original spectrum are so close together that they have merged. Using the convolution approach, this clearly resolves into two peaks. The operation of the convolution algorithm is shown in detail at position (c). The horizontal arrow shows the direction of scan for mask convolution; a mask is shown in four consecutive positions.

response value with a threshold value. Only if all mask response values exceed their corresponding thresholds will a result be retained. Threshold values may be selected either manually or automatically.

Another strategy used to reduce the search space explored by the program is to introduce *search levels*. This breaks up the search into a number of steps or levels. Within a given search level, only a limited number of

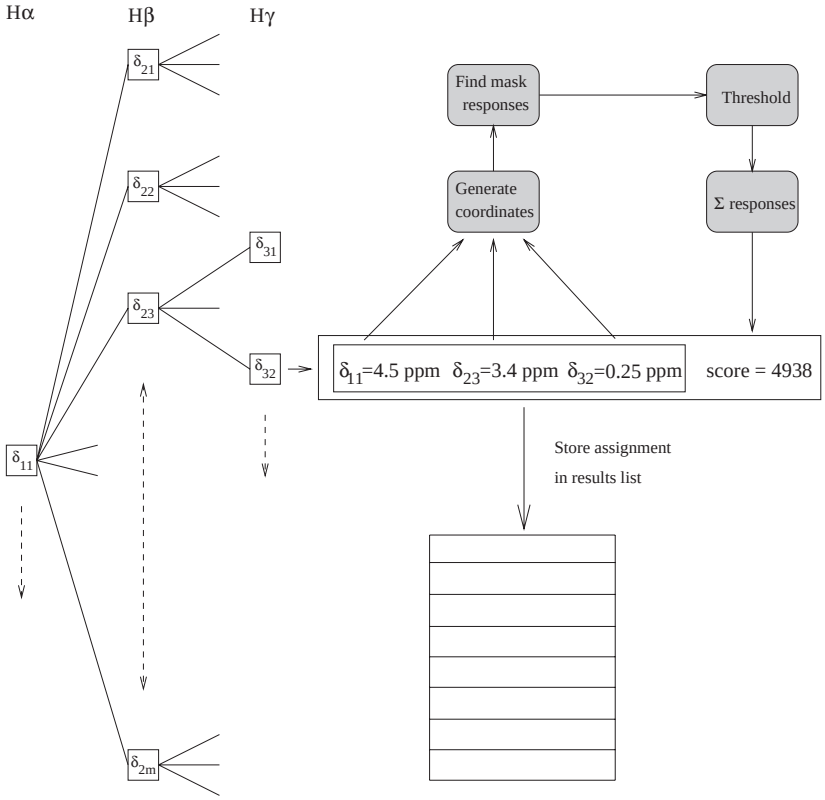



Fig. 4. Pattern searching. The exhaustive generation of chemical shift combinations is shown here as a tree. Only one $H^{\alpha}$ value is illustrated; of course, there will actually be many $H^{\alpha}$ values, one for each data point increment along the $H^{\alpha}$ chemical shift range. Similarly for the $H^{\beta}$'s and the $H^{\gamma}$'s. Each combination of chemical shifts defines sets of coordinates in the regions of interest. The mask response values for each putative peak (taken from the convolved regions of interest) are summed to give a score for this putative assignment. The assignment is stored in a list, ordered by score value.

peaks are searched for. Which peaks are searched for at which search level is determined by the pattern file. At each search level, a set of intermediate results will be generated. These results are used to constrain the search at the next search level, instead of using the full chemical shift ranges, with a significant time saving.

The assignments produced by the search algorithm will be stored in a *results list*, ordered according to score. To avoid storing results with only marginally different chemical shift values, the incoming assignment is checked against existing ones to see if there are any from the same spin system, but with a higher score. If not, the new assignment is stored, otherwise it is rejected. The results list forms the input for the next step in the program.

*Result list filtering*

Typically, the pattern search process will generate between 50 and 2000 results, depending on the pattern being searched for and the exact search control parameters being used. Although the correct results would be expected at the top of the results list, manual inspection would be a daunting task. Hence, a suite of heuristic results list filtering algorithms has been written to bring the list down to a manageable size and to make sure that the correct results are more likely to be the highest scoring ones. The algorithms embody criteria similar to those used during manual assignment of spin systems. There are two classes of filtering algorithm: *penalising* and *deleting*. The penalising type multiplies the score for a given result by a value which lies between zero and one, called a penalisation factor. The penalisation factor is a measure of how 'sensible' a result looks according to a particular criterion. The deletion type is mainly used for removing redundant results. The parameters required by the filtering algorithms may be supplied to the program via the pattern file.

The following penalisation algorithms are available:

*Too-far* Penalise in cases where two chemical shifts are more than a given distance apart. For example, if the pattern file defines the search ranges for $H^{\beta}/H^{\beta'}$ in an AMX spin system as being from 0.5 ppm to 3.0 ppm, the program might find candidate patterns with $H^{\beta} = 0.53$ ppm and $H^{\beta'} = 2.94$ ppm. Such wide separations of chemical shift values seldom arise in real spectra, and should therefore be penalised.

*Nucleus-order* Penalise results which contain badly ordered chemical shifts. For instance, in the case of serine, the ranges for $H^{\alpha}$ and $H^{\beta}$ are heavily overlapping, but the $H^{\beta}$ chemical shifts tend to be lower than the $H^{\alpha}$ chemical shifts.

*Peaks* Penalise results whose chemical shift values do not lie exactly at the maximum points of spectrum peaks. This will tend to promote 'model' results, those where all peaks are distinct and precisely aligned.

*Excess-peaks* This function penalises results where, at a given chemical shift, there are more peaks in the spectrum than would be expected for the spin system currently being searched for. For example, in the case of glycine, the peak pattern which we are searching for in a 2D COSY or 2D TOCSY spectrum may also match peaks from AMX spin systems.

*Uninst-resp* In order to allow for missing peaks or even missing groups of peaks, search levels may be 'skipped' by the program. This will lead to some mask responses having no value – they are uninstantiated. Results with many uninstantiated mask responses are considered to be 'poor' and can be penalised with this algorithm.

*Uninst-chm-shft* In the extreme case, skipping search levels in the manner described for *Uninst-resp* can result in chemical shift values not being found either. Results with many missing (or uninstantiated) chemical shifts are undesirable, and this algorithm can be used to penalise them.

*Thresh-count* Penalise if any mask responses are below absolute threshold values. This feature allows results to be accepted even if one or more of the mask response values is below threshold, a situation which can arise if the program performs dynamic threshold adjustment during search.

*Deg-pks-pen* Geminal protons may sometimes have similar or identical chemical shifts (degeneracy); deleting such results is undesirable. However, due to the way in which the search algorithm is implemented, the program has a strong tendency to find degenerate chemical shifts. Degenerate peaks penalisation can be used to counteract this tendency.

The following deleting algorithms are available:

*Ovrlp-clus* This will look for triplets of results for which, in a given pair of spins, one has different chemical shifts and the other two identical ones. For example,

| # | $C^{\alpha}$ | $H^{\alpha}$ | $C^{\beta}$ | $H^{\beta}$ | $H^{\gamma}$ | $H^{\gamma'}$ | Score |
|---|---|---|---|---|---|---|---|
| 1 | 56.9 | 4.36 | 36.2 | 3.45 | 2.18 | 2.18 | 20 813 |
| 2 | 56.9 | 4.36 | 36.2 | 3.45 | 1.56 | 2.18 | 17 993 |
| 3 | 56.9 | 4.36 | 36.2 | 3.45 | 1.56 | 1.56 | 13 329 |

In this case, it is highly likely that result 2 is correct, but the other two are not. Hence, we retain result 2 and discard the others.

*Exs-penal* A result containing more than two or three chemical shift degeneracies has a rather low probability of being correct. This algorithm removes such overdegenerate results.

*Sym* If two spins have overlapping chemical shift ranges (for instance, the $H^{\beta}/H^{\beta'}$ spins), then the program can produce two results with identical (or very similar) score values, but with transposed chemical shifts between the two spins. One of these, the one with the lowest score, is deleted.

*Sec-clus* The results lists produced by the pattern search algorithm tend to contain many results with chemical shift sets which differ only slightly from one another. Few of these results will correspond to genuine assignments – most will be 'near misses'. Secondary clustering is designed to remove the near misses and retain the genuine assignments. The algorithm provides two chemical shift difference bands. These specify the maximum allowable differences in chemical shifts between corresponding spins in two results. For any two given results to be clusterable, a given fraction of the differences in chemical shifts must lie within one band, whilst the rest must lie within the other band. If two results are considered clusterable, then the lowest scoring result will simply be deleted.

*Lo-score* Results in a results list are ordered according to their score. As one descends the results list, the scores gradually decline. If there is a sudden decline in score, it indicates that multiple penalisations have cut in simultaneously, and it is likely that subsequent results will be of poor quality. These are deleted.

## Experimental background

Program development and testing was carried out on a Silicon Graphics 4D/480VGX 8 processor machine. The basic pattern recognition algorithms described in the previous section were tested using a suitable data set (HCCH-COSY and -TOCSY, HNCA, HN(CO)CA, CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO) of the N-terminal domain of human protein disulphide isomerase (Kemmink et al., 1995,1996). The protein shows multiple signals, most probably due to proline cis–trans isomers. It is known that $Pro^{83}$ occurs predominantly in a cis conformation, and the last three residues, $Pro^{118}$, $Ala^{119}$ and $Ala^{120}$, are part of a flexible region at the C-terminus of the protein.

The spectra were converted from Snarf (Frans van Hoesel, University of Groningen, Groningen, The Netherlands) to Bruker format, using in-house software (see 'UXNMR Operator's Manual: Data Processing', Bruker Analytik GmbH, for full details of the Bruker spectrum format). A modified version of the Bruker program AURELIA (Neidig, 1992) has been produced to view the results generated by CATCH23. This allows the user to view the putative assignments superimposed upon the original spectra. It works by reading in the results file for a given pattern search run, and using the information in this file to select which spectra to display. With three-dimensional spectra, a single slice is displayed for each spectrum. A list of the results is also displayed, and if the user clicks on one of them, the assigned chemical shifts will appear as lines within the selected spectra. This allows the user to assess whether or not the search strategy has worked successfully for each result.

## Results

Two sets of pattern search experiments are presented in this section: the first concentrates on side chains and the second on the backbone. In designing a search pattern, one may aim for different kinds of results. A small number of results, all of which are definitely correct, can be obtained by performing a pattern search with relatively strict conditions. A larger number of results, amongst which there may be some incorrect ones, but which will contain all possible correct instances of the pattern being searched for, can be obtained by using more relaxed constraints. In this latter case, additional selection would then take place during the sequential assignment. The patterns used to generate the results presented below employed fairly strict conditions.

*Side-chain pattern*

In the first set of experiments, the patterns for glycine, alanine, AMX, serine, threonine, valine, GLX, isoleucine and leucine were tested on the HCCH-COSY and HCCH-TOCSY spectra.

The data set used for searching side-chain spin systems should fulfil some basic technical requirements, but is allowed to be noisy and may contain artefacts. As one of the technical requirements, the mixing time for the HCCH-TOCSY spectrum should be chosen such that correlations between protons separated by four or more bonds are visible, e.g. from the α-carbon to the γ- or δ-carbons. In our case, a mixing time of 24 ms was used. Furthermore, HCCH-COSY and -TOCSY spectra should be recorded in an orthogonal manner, i.e. in one case all cross peaks should occur along F1 and in the other along F2. In this way, some effects of $t_1$-noise can be compensated for. Also, we would expect the sample to be dissolved in $D_2O$ for the HCCH-type experiments, with only mild water suppression by irradiation applied. The resolution for both COSY and TOCSY should be at least 13.4 Hz per data point on the proton axes and 39.1 Hz per data point on the carbon axis.

As a key feature of the side-chain pattern search, limits are imposed on the chemical shift ranges scanned by the program in order to make the search more specific and to speed up operation. The limits used in the tests described below are based on chemical shift values published in the literature (e.g. Groß and Kalbitzer (1988) or Wishart et al. (1991)).

A summary of the full results set is shown in Fig. 5. The results of four selected pattern searches are discussed in detail below.

*Alanine*

Alanine constitutes one of the simplest patterns, but the β-carbon has a very characteristic chemical shift. The correct results are concentrated at the top of the results list, indicating that the ranking introduced by the results

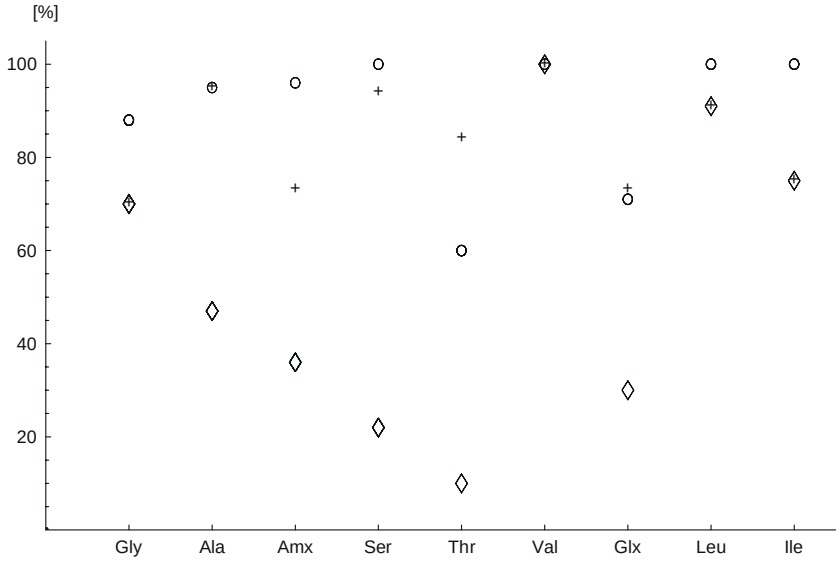



Fig. 5. Summary of side-chain results. Three measures are used to characterise the performance of the program with a given pattern, all normalised and expressed in percent, and indicated by three different symbols. The '○' symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The '◇' symbol shows the number of correct assignments compared to the total number in the results list. The '+' symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the top of the list.

list filtering heuristics is correct. Nonetheless, the list also contains many non-alanines, for various reasons. First, the alanine pattern is *so* simple that peaks from other spin systems are also found. For example, the $H^{\gamma}/H^{\delta}/C^{\delta}$ peaks of proline or the $H^{\beta}/H^{\alpha}/C^{\alpha}$ peaks of leucine fall within the search ranges for the $H^{\beta}/H^{\alpha}/C^{\alpha}$ peaks of alanine. In some of these cases, partially correct alanine spin systems are found, with a valid peak in the $C^{\beta}$ plane but a false peak in the $C^{\alpha}$ plane. Such results *could*, in principle, be filtered out by the secondary clustering mechanism, but then alanines with overlapping chemical shifts would not be distinguished by the program. Finally, some of the results *look* like alanines when manually inspected, but do not fit into the results from the backbone spectra. It is possible that these are due to minor conformations, caused by proline cis–trans isomerisation.

*Valine*

The uniqueness of this pattern, plus the good dispersion of valine peaks, plus the fact that these peaks do not overlap significantly with peaks from other spin systems combined to produce an exceptionally clean results list. The full valine results set is given in Table 1.

*Isoleucine*

Isoleucine presents a rather unique pattern of peaks. However, identifying these patterns was not a trivial task, due to very weak peaks between $H^{\beta}$ and $H^{\gamma 1}$ in the TOCSY spectrum. The pattern used to tackle these problems is shown in Fig. 6, along with the associated chemical shift ranges and results list filtering procedures employed. In Table 2 the full list of isoleucine results is shown. Although the sequence contains only three isoleucines, the program has found a fourth 'ghost' spin system with no corresponding $C^{\alpha}$ or $C^{\beta}$ chemical shifts in the CBCANH or CBCA(CO)NH spectra. Manual inspection indicates that this is indeed a genuine isoleucine spin system. Careful inspection of the $^{13}C$ 3D NOESY confirmed the pres-

TABLE 1
VALINE RESULTS

| Resl | $H^{\beta}$ | $H^{\alpha}$ | $C^{\alpha}$ | $C^{\beta}$ | $H^{\gamma 3}$ | $C^{\gamma 3}$ | $H^{\gamma 3'}$ | $C^{\gamma 3'}$ | Resp | Assign |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | **2.29** | **4.08** | **61.1** | **32.3** | **0.88** | **20.4** | **0.91** | **21.4** | 7 940 | $Val^{9}$ |
| 1 | **1.90** | **4.25** | **62.6** | **31.4** | **0.91** | **19.8** | **0.88** | **20.8** | 10 195 | $Val^{11}$ |
| 2 | **2.18** | **4.36** | **61.4** | **34.5** | **0.79** | **21.4** | **0.25** | **22.3** | 94 778 | $Val^{29}$ |
| 3 | **1.30** | **4.32** | **60.1** | **36.4** | **0.49** | **21.4** | **0.71** | **20.4** | 123 100 | $Val^{65}$ |
| 4 | **1.75** | **3.58** | **63.9** | **30.4** | **0.79** | **21.4** | **0.16** | **20.4** | 124 827 | $Val^{79}$ |
| 5 | **2.11** | **3.41** | **67.6** | **31.4** | **0.91** | **21.4** | **1.05** | **22.9** | 141 060 | $Val^{109}$ |

Chemical shift values in bold are those that have been correctly identified by the program. The 'Resp' column gives the score for each result. The 'Assign' column shows the position of the result in the sequence, if appropriate.

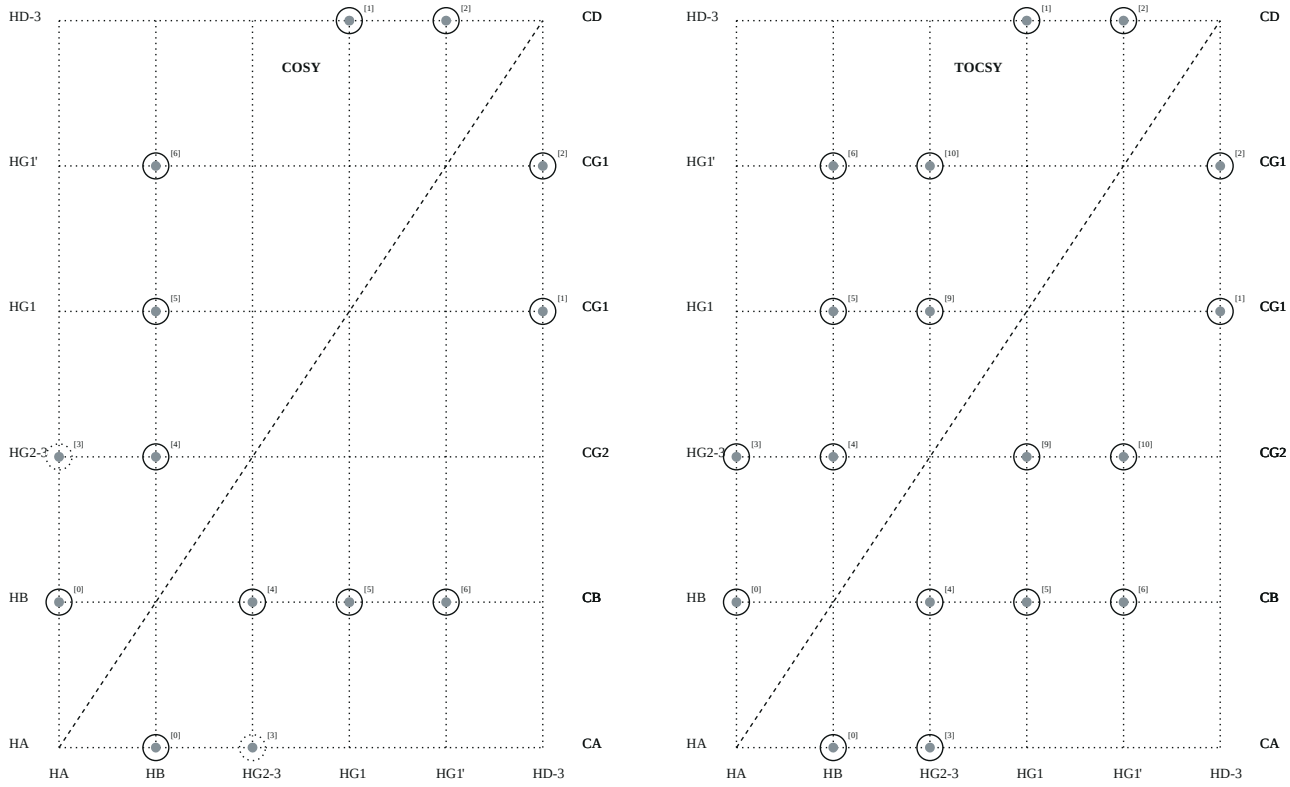



Fig. 6. Isoleucine pattern. Full circles represent peaks which are *expected*; dashed circles represent peaks which should *not* be present. The numbers in square brackets adjacent to each peak show the search level at which the peak is searched for. The columns of carbon spin names on the right-hand side of the figures indicate the carbon planes within which the corresponding row of peaks will be found. Results list processing procedures employed: *Ovrlp-clus*, *Sym*, *Sec-clus*, *Fract-score*, *Exs-penal*, *Too-far*, *Std-dev*, *Thresh-count*, *Excess-peaks*, *Uninst-chm-shft*, *Uninst-resp*. Chemical shift ranges: N (115.5→133.0); H (6.91→9.88); CO (170.6→179.0); $C^{\alpha}$ (49.0→66.6); $H^{\alpha}$ (3.06→6.00); $C^{\beta}$ (30.1→42.6); $H^{\beta}$ (0.73→2.70).

ence of this additional spin system, which is probably also due to the existence of cis–trans isomeric proteins. The pattern search also yielded a second proton at each $C^{\gamma 1}$, which were not found in the original manual assignment.

In Fig. 7, one of the isoleucine results is shown graphically. This is interesting for a number of reasons. It displays near-degeneracy for both $H^{\beta}/H^{\gamma 1'}$ and $H^{\gamma 1}/H_3^{\gamma 2}$. The $C^{\alpha}$ plane in the HCCH-COSY spectrum contains significant artefacts. In spite of this, the program has managed to locate the correct peaks. Finally, this spin system shows extensive overlap with the ghost isoleucine, demonstrating the program's ability to distinguish between similar spin systems.

*Leucine*

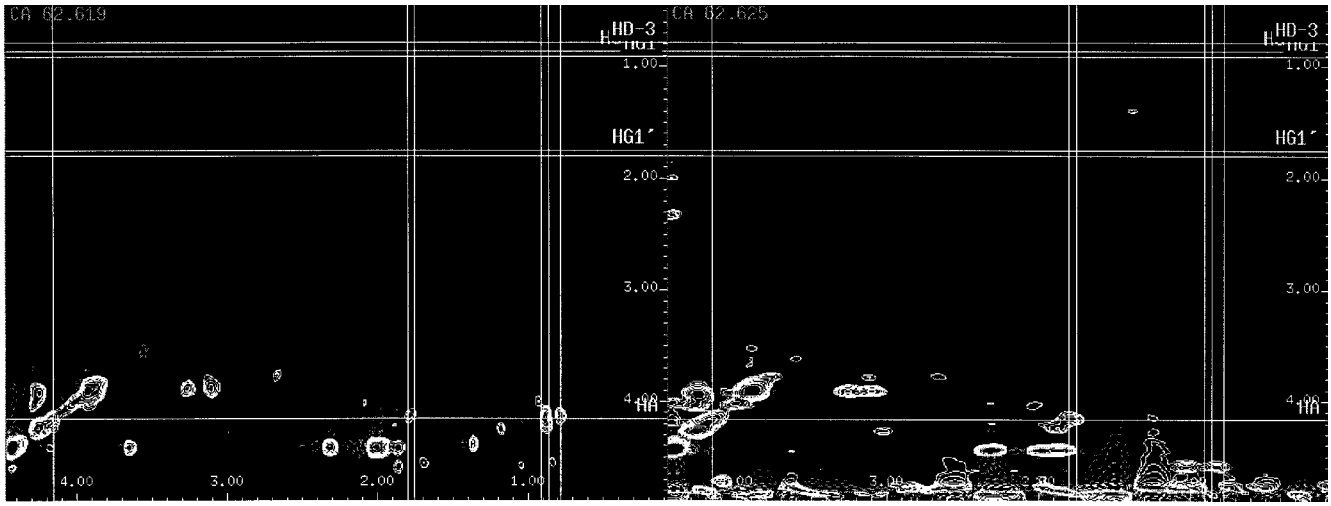
Leucine also produces a unique pattern of peaks, and the β-carbon chemical shift is unusually large. As with the isoleucines however, the peaks between $H^{\beta}$ and $H^{\gamma}$ were small, making this a challenging problem for the program. All 10 leucine patterns were found, although in one case an $H^{\gamma}$ chemical shift did not agree with the value reported by Kemmink et al. (assignments are available as supplementary information to Kemmink et al. (1995)), and in another case the $H^{\beta}$ was incorrectly assigned by the program as degenerate with the $H^{\beta'}$. Also, one $C^{\delta}$ chemical shift found by the program was incorrect. The 'extra' result found was composed almost entirely of peaks from a valine spin system.

TABLE 2
ISOLEUCINE RESULTS

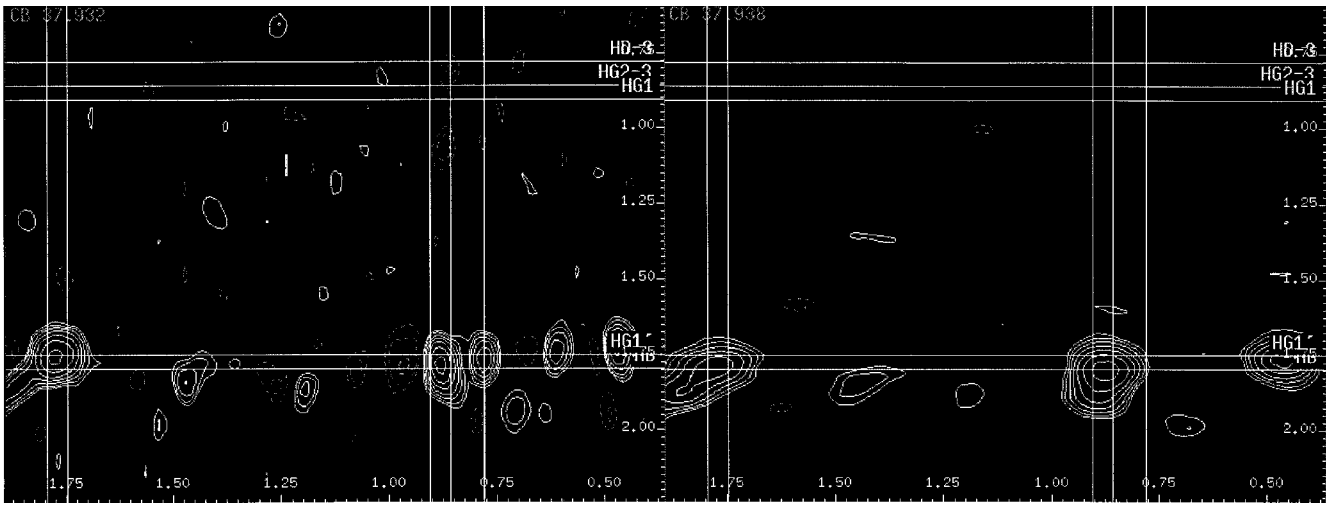
| Resl | $H^{\beta}$ | $H^{\alpha}$ | $C^{\alpha}$ | $C^{\beta}$ | $H_3^{\gamma 2}$ | $C^{\gamma 2}$ | $H^{\gamma 1}$ | $C^{\gamma 1}$ | $H^{\gamma 1'}$ | $H^{\delta 3}$ | $C^{\delta}$ | Resp | Assign |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.86 | 4.14 | 61.1 | 37.9 | 0.91 | 16.7 | 1.19 | 27.0 | 1.47 | 0.86 | 12.3 | 4 611 | |
| 1 | **1.82** | **4.18** | **62.9** | **38.3** | **0.88** | **17.9** | **0.93** | **27.9** | **1.77** | **0.80** | **13.9** | 12 007 | $Ile^{60}$ |
| 2 | **1.77** | **3.85** | **64.8** | **38.6** | **0.47** | **16.4** | **0.58** | **27.3** | **1.85** | **0.63** | **12.9** | 39 044 | $Ile^{108}$ |
| 3 | **1.99** | **5.55** | **60.1** | **38.9** | **0.72** | **17.9** | **0.66** | **27.6** | **1.61** | **0.77** | **12.9** | 95 047 | $Ile^{85}$ |

Chemical shift values in bold are those that have been correctly identified by the program; chemical shift values in plain text are erroneous. The 'Resp' column gives the score for each result. The 'Assign' column shows the position of the result in the sequence, if appropriate. No entry at all in this column means that something is visible in the spectra, but this does not correspond to a known spin system.

a




b




c

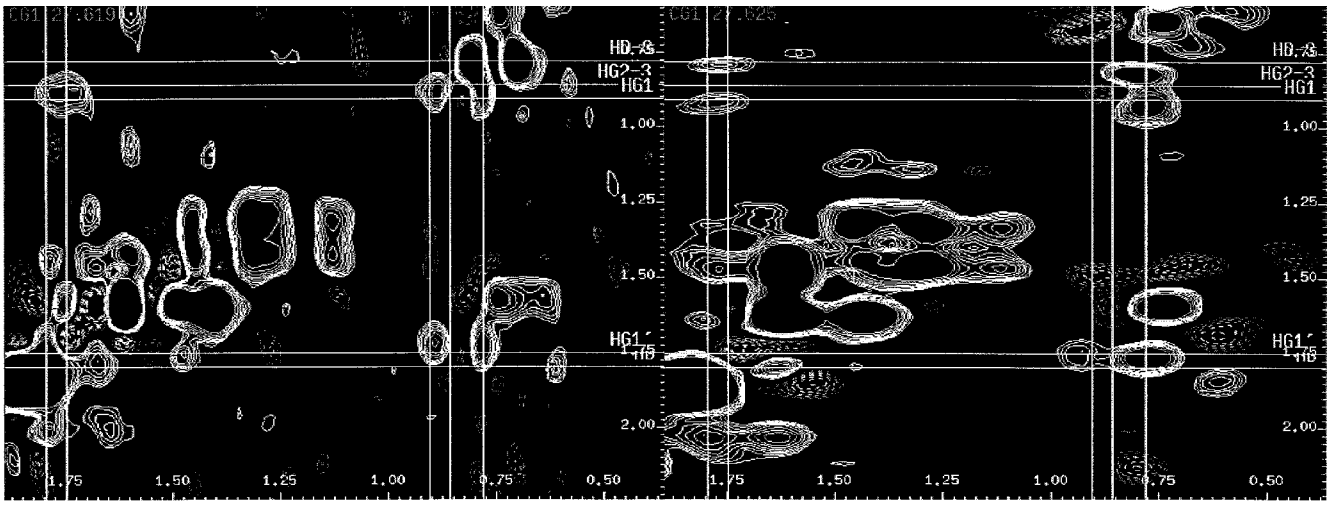
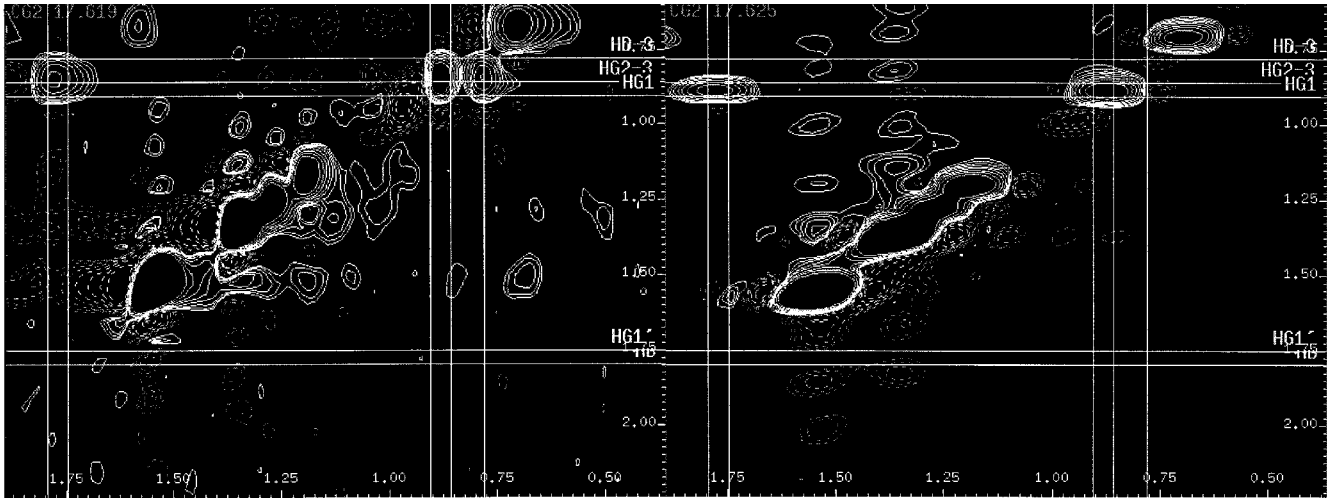



Fig. 7. Isoleucine result, #1 from the list given in Table 2, showing the (a) $C^{\alpha}$ plane, (b) $C^{\beta}$ plane, (c) $C^{\gamma 1}$ plane, (d) $C^{\gamma 2}$ plane, and (e) $C^{\delta}$ plane of the HCCH-TOCSY (left) and -COSY (right). The lines indicate the F1 and F3 frequencies of the chemical shift positions of the individual spins, i.e. $H^{\alpha} = 4.18$, $H^{\beta} = 1.82$, $H^{\gamma 1} = 0.93$, $H^{\gamma 1'} = 1.77$, $H^{\gamma 2} = 0.88$, $H^{\delta} = 0.80$. The carbon (F2) frequencies are given in the upper left-hand corner of each plot. In the HCCH-TOCSY, F3 is the horizontal axis and F1 the vertical; in the HCCH-COSY, F3 is the vertical axis and F1 the horizontal.

d




e

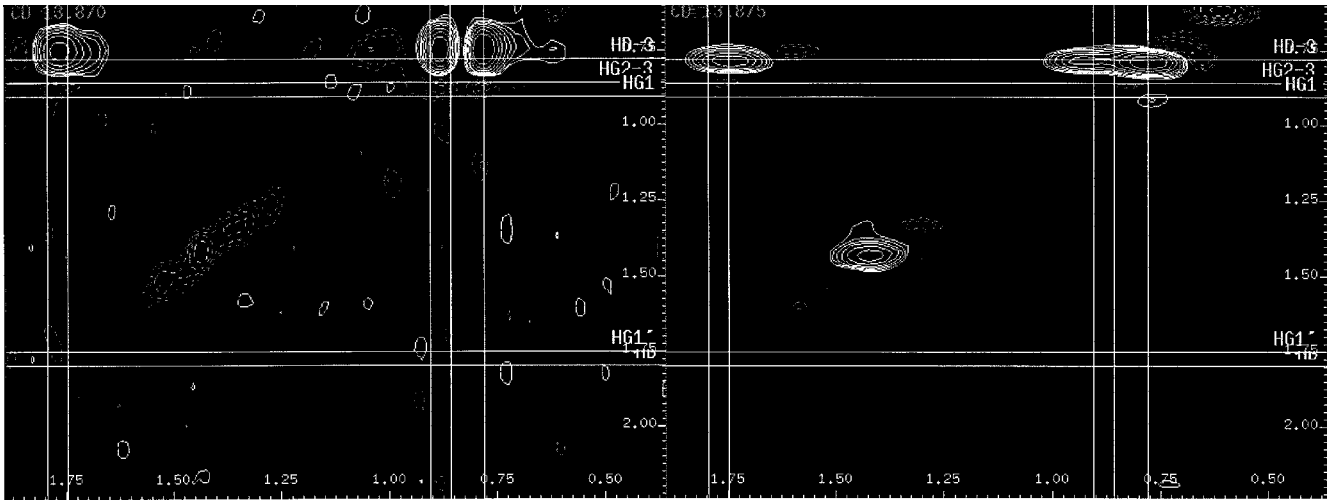



Fig. 7. (continued).

*Evaluation of backbone spectra*

The feasibility of pattern searches in backbone spectra was tested with a set of HNCO, HN(CA)CO, CBCANH, CBCA(CO)NH, HNCA and HN(CO)CA spectra. They are intended to constitute a minimal data set for reliable assignment; they may be augmented by other types of spectra if required. The spectra were not recorded under uniform conditions. Different NMR spectrometers were used, with different field strengths, slightly different temperatures, and gaps of several months between some of the measurements. For these reasons, the chemical shifts of the amide protons varied substantially in a nonsystematic manner.

The emphasis in the searches presented in this section was therefore put on compensating for the large chemical shift differences between the spectra, with the unavoidable consequence that some of the results contained incorrect chemical shifts.

Search patterns containing the following chemical shifts were constructed:

$N_i$, $NH_i$
$C_i^{\alpha}$, $C_i^{\beta}$, $CO_i$
$C_{i-1}^{\alpha}$, $C_{i-1}^{\beta}$, $CO_{i-1}$

where i is the current residue in the sequence and i–1 is its predecessor. This pattern allows connected pairs of residues in the protein sequence to be recognised. The results from the search for this pattern may be fed into a sequential assignment tool, which is capable of chaining them together.

A special feature of glycine is that it shows very characteristic $C^{\alpha}$ chemical shifts. Also, $C^{\beta}$ peaks do *not* show up in the HNCA and HN(CO)CA spectra. Hence, glycine $C^{\alpha}$ peaks can be identified uniquely in these spectra. This fact was used in constructing two patterns: one which

TABLE 3
XX PAIR RESULTS

| Resl | H#1 | N#1 | CO | CO#1 | $C^{\alpha}$#1 | $C^{\alpha}$ | $C^{\beta}$#1 | $C^{\beta}$ | Resp | Orig | Assign |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | **8.43** | **128.3** | – | – | 55.6 | **63.2** | 42.4 | 55.0 | 8605 | 24217 | 12 |
| 1 | **7.11** | **118.8** | – | – | **60.7** | **58.8** | **70.2** | **39.2** | 9553 | 40140 | 100 |
| 2 | 8.19 | 127.7 | – | – | **53.1** | 62.0 | **19.0** | 70.8 | 12682 | 52724 | |
| 3 | **8.03** | 118.2 | – | – | 62.0 | **58.8** | 70.8 | 39.2 | 13092 | 53879 | |
| 4 | **8.78** | **119.1** | – | – | 64.5 | **63.8** | **72.7** | **34.2** | 20540 | 81138 | 84 |
| 5 | **8.93** | **128.3** | – | – | **52.5** | **60.7** | **38.6** | **36.7** | 22831 | 59419 | 66 |
| 6 | **6.27** | **118.5** | – | – | **48.7** | **55.6** | **20.3** | **42.4** | 24712 | 64383 | 33 |
| 7 | **9.80** | **128.0** | – | – | **62.6** | **54.4** | **34.8** | **44.9** | 25686 | 65846 | 29 |
| 8 | **8.37** | 122.8 | – | – | 56.9 | **63.8** | 33.5 | **32.9** | 26400 | 68628 | |
| 9 | 8.15 | 115.4 | – | – | 62.6 | **55.0** | **53.7** | **41.8** | 26523 | 67377 | |
| 10 | **9.25** | **127.1** | – | – | **57.5** | **59.4** | **31.0** | **64.5** | 27457 | 105676 | 59 |
| 11 | **9.84** | **126.5** | – | – | **56.9** | **55.6** | **19.7** | **31.7** | 28367 | 72045 | 105 |
| 12 | 8.29 | 133.3 | – | – | 55.0 | **55.0** | **19.7** | 43.0 | 28435 | 73099 | |
| 13 | **8.76** | **125.6** | – | – | **63.2** | **58.8** | **39.2** | **30.4** | 29014 | 74183 | 46 |
| 14 | **8.14** | **125.3** | – | – | **60.1** | **64.5** | **32.3** | **34.2** | 29490 | 76220 | 40 |
| 15 | 8.54 | 111.7 | – | – | **57.5** | **57.5** | 45.5 | 30.4 | 31941 | 81952 | |
| 16 | **8.74** | **117.9** | – | – | **64.5** | **59.4** | **34.2** | **31.0** | 33541 | 84724 | 39 |
| 17 | 8.42 | 109.8 | – | – | 46.8 | **53.1** | 45.5 | **19.0** | 35366 | 90042 | |
| 18 | **8.78** | **131.1** | – | – | **53.7** | **52.5** | **20.9** | **38.6** | 36834 | 92038 | 67 |
| 19 | **8.86** | **126.2** | – | – | **55.0** | 55.6 | **41.8** | 32.9 | 36868 | 92137 | 1 |
| 20 | **9.50** | **121.9** | – | – | **50.0** | **55.0** | **26.0** | **43.6** | 39269 | 97839 | 63 |
| 21 | **8.62** | **124.0** | – | – | **63.8** | **57.5** | **38.6** | **31.0** | 41415 | 103767 | 60 |
| 22 | **9.18** | **132.7** | – | – | **54.4** | **54.4** | **44.9** | **46.8** | 44937 | 111426 | 28 |
| 23 | **8.86** | **129.9** | – | – | **55.6** | **64.5** | **31.7** | **31.0** | 49016 | 119375 | 80 |
| 24 | **7.26** | **116.9** | – | – | **55.0** | **53.1** | **30.4** | **19.0** | 58113 | 138454 | 24 |
| 25 | **9.82** | **132.3** | **172.7** | **174.7** | **60.7** | 64.5 | **39.9** | **72.7** | 64743 | 125855 | 85 |
| 26 | **8.10** | **118.5** | **178.1** | **177.8** | 65.7 | **65.7** | **32.3** | **39.2** | 102231 | 107022 | 109 |
| 27 | 8.53 | **127.1** | **176.4** | **176.3** | 63.2 | 63.2 | 19.0 | **34.8** | 107732 | 123390 | |
| 28 | **9.25** | **125.9** | **173.2** | **175.8** | **55.6** | **52.5** | **43.6** | **32.3** | 116676 | 124838 | 62 |
| 29 | **8.34** | **124.6** | **172.2** | 178.3 | **60.1** | **55.0** | 52.5 | **30.4** | 118183 | 126868 | 25 |
| 30 | **7.46** | **118.8** | **179.1** | **178.1** | **58.8** | **60.7** | **32.3** | 53.7 | 123579 | 137738 | 114 |
| 31 | **9.74** | **124.9** | **173.8** | **173.5** | **56.3** | **54.4** | **42.4** | **38.6** | 129173 | 157976 | 87 |
| 32 | **9.40** | **126.8** | **174.7** | **173.8** | **54.4** | **60.7** | **38.6** | **39.9** | 132435 | 142155 | 86 |
| 33 | **7.20** | **117.6** | **177.7** | **179.1** | **53.7** | **57.5** | **18.4** | **42.4** | 143988 | 150035 | 22 |
| 34 | **9.40** | **130.5** | **173.5** | **173.9** | **57.5** | **55.6** | **41.8** | **42.4** | 146274 | 163026 | 88 |
| 35 | **7.08** | **119.7** | **180.6** | **177.7** | **58.8** | **66.4** | **30.4** | **31.0** | 157069 | 173426 | 45 |
| 36 | **7.52** | **122.5** | **178.6** | **178.1** | **65.7** | **57.5** | **39.2** | **41.1** | 185736 | 271141 | 108 |
| 37 | **7.35** | **120.6** | 182.8 | **178.6** | **58.2** | **58.2** | **41.1** | **41.1** | 194921 | 204572 | 107 |
| 38 | **7.43** | **118.8** | **179.1** | **178.1** | **58.2** | **60.7** | 56.9 | **31.7** | 198028 | 233222 | 114 |
| 39 | **9.08** | **126.5** | **176.4** | **175.6** | **56.9** | **55.0** | **31.0** | **34.8** | 199412 | 253019 | 98 |
| 40 | **8.37** | **117.9** | **177.2** | **175.2** | **63.8** | **54.4** | **70.2** | **41.1** | 200218 | 406223 | 93 |
| 41 | **9.12** | **129.3** | **175.9** | **175.8** | **56.3** | **62.0** | **43.6** | **32.9** | 207811 | 254292 | 10 |
| 42 | **8.39** | **119.7** | **179.2** | **180.0** | **58.2** | **62.6** | **43.0** | **30.4** | 211763 | 257968 | 112 |
| 43 | **8.35** | **121.6** | 177.3 | **175.8** | **55.6** | **57.5** | **33.5** | **31.0** | 220956 | 295113 | 5 |
| 44 | **9.96** | **124.3** | **176.7** | **176.6** | **55.0** | **53.7** | **38.0** | **33.5** | 234546 | 330988 | 90 |
| 45 | **6.92** | **123.4** | **175.8** | **177.3** | **57.5** | **63.8** | **31.7** | **68.9** | 239923 | 543642 | 69 |
| 46 | **8.66** | **108.6** | **181.5** | 176.1 | 48.7 | **55.6** | 47.4 | **19.0** | 242511 | 290281 | G51 |
| 47 | **8.29** | **123.1** | **177.8** | **179.4** | **62.6** | **56.9** | **30.4** | **38.6** | 251932 | 318695 | 111 |
| 48 | **8.63** | **110.5** | **177.0** | **175.8** | **64.5** | **53.7** | 63.2 | **20.3** | 253246 | 488422 | 68 |
| 49 | **8.22** | **117.9** | 183.1 | **180.0** | **59.4** | **55.6** | **28.5** | **18.4** | 253668 | 272056 | 75 |
| 50 | **9.04** | **122.5** | **175.5** | **169.9** | **55.6** | **57.5** | **42.4** | **40.5** | 255781 | 361790 | 32 |
| 51 | **7.99** | **124.9** | **176.9** | **176.3** | **63.8** | **53.1** | 51.9 | **39.9** | 256295 | 374923 | 17 |
| 52 | **8.27** | **116.6** | **175.0** | **173.6** | 62.6 | **59.4** | **42.4** | **32.9** | 263196 | 411262 | 26 |
| 53 | **7.89** | **119.1** | **177.2** | **180.9** | **55.6** | **63.2** | **18.4** | **39.9** | 264047 | 288018 | 47 |
| 54 | **8.47** | **120.3** | **180.0** | **179.1** | **60.7** | **58.2** | **32.3** | **43.0** | 265415 | 337539 | 113 |
| 55 | **8.96** | **123.7** | **173.9** | **176.6** | **53.7** | **57.5** | **33.5** | **41.8** | 266203 | 403783 | 89 |
| 56 | **8.51** | **122.8** | **177.5** | **178.9** | **58.2** | **62.0** | **40.5** | **63.2** | 269578 | 512650 | 72 |

Chemical shift values in bold are those that have been correctly identified by the program; chemical shift values in plain text are erroneous. The 'Resp' column gives the score for each result. The 'Assign' column shows the position of the result in the sequence, if appropriate. If a letter appears beside the number, this gives an indication of which residue was selected by the program in the case of an erroneous assignment. No entry at all in this column means that something is visible in the spectra, but this does not correspond to a known spin system.

TABLE 3
(continued)

| Resl | H#1 | N#1 | CO | CO#1 | $C^{\alpha}$#1 | $C^{\alpha}$ | $C^{\beta}$#1 | $C^{\beta}$ | Resp | Orig | Assign |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 57 | **7.40** | **122.2** | **177.7** | **176.6** | **57.5** | **56.3** | **15.2** | **43.0** | 274509 | 324549 | 43 |
| 58 | **9.09** | **124.3** | **176.3** | **178.4** | **61.3** | **54.4** | **31.7** | **34.8** | 276869 | 321707 | 14 |
| 59 | **8.67** | **122.8** | **177.2** | **176.4** | **55.0** | **63.8** | **34.8** | **31.0** | 288152 | 347166 | 97 |
| 60 | **8.31** | **121.6** | **177.3** | **176.9** | **53.1** | **58.2** | **27.2** | **31.7** | 293059 | 470343 | 70 |
| 61 | **8.13** | **123.4** | **180.0** | **177.3** | **56.3** | 55.0 | **42.4** | **41.1** | 294932 | 509704 | 76 |
| 62 | **8.03** | **124.6** | **180.5** | **179.4** | 64.5 | 53.1 | 40.5 | 39.9 | 299325 | 495336 | 55 |
| 63 | **8.89** | **123.7** | **175.8** | **176.3** | **55.6** | **55.6** | **31.7** | **33.5** | 301393 | 441913 | 6 |
| 64 | **8.50** | **127.7** | 173.5 | **175.3** | **58.8** | **56.9** | **39.2** | **31.0** | 308410 | 459288 | 99 |
| 65 | **9.54** | **122.5** | 175.8 | **172.4** | **55.0** | **50.0** | **36.7** | **26.0** | 312303 | 511128 | 64 |
| 66 | **9.36** | **109.8** | **175.6** | **173.9** | 45.5 | **55.0** | 46.8 | **38.0** | 320815 | 605053 | G91 |
| 67 | **7.47** | **119.7** | **179.1** | **177.3** | **53.7** | **53.7** | **19.0** | **18.4** | 329225 | 413948 | 23 |
| 68 | **7.55** | **117.2** | **176.9** | **177.5** | 53.1 | **53.1** | 27.2 | **27.2** | 345719 | 546748 | 71 |
| 69 | **8.02** | **124.0** | 175.0 | 173.3 | **55.6** | **59.4** | **17.8** | **32.3** | 346587 | 463386 | 55 |
| 70 | **7.42** | **117.9** | **179.2** | **176.4** | **56.9** | **58.2** | **31.0** | 18.4 | 347625 | 445397 | 115 mix |
| 71 | **8.01** | **122.5** | **180.9** | **180.0** | **60.1** | **55.0** | **32.9** | **18.4** | 363844 | 521864 | 48 |
| 72 | **7.95** | **120.9** | **178.3** | **181.5** | 64.5 | **55.6** | **19.0** | 47.4 | 366110 | 469399 | 50 |
| 73 | **8.33** | **124.9** | 172.2 | **178.3** | 53.7 | **60.1** | 20.3 | **32.9** | 396681 | 538397 | 49 mix |
| 74 | **7.54** | **121.2** | **178.7** | **179.2** | **55.0** | **60.1** | **18.4** | **32.3** | 397662 | 618172 | 41 |
| 75 | **7.97** | **120.3** | **178.0** | **179.4** | **55.6** | **58.2** | **18.4** | **43.6** | 405770 | 515512 | 74 |
| 76 | **8.71** | **124.0** | **176.9** | **175.5** | **57.5** | **63.2** | **30.4** | **32.3** | 407801 | 428804 | 4 |
| 77 | **7.94** | **113.9** | **178.6** | **172.2** | **56.3** | **53.7** | 55.0 | **20.3** | 420532 | 854628 | 95 |
| 78 | **8.23** | **120.3** | **181.2** | **178.7** | **60.1** | **55.6** | **29.1** | **17.8** | 430907 | 499378 | 19 |
| 79 | **7.67** | **111.4** | **174.4** | **172.1** | 46.2 | **62.0** | **44.9** | **70.8** | 440311 | 848437 | G117 |
| 80 | **7.85** | **124.9** | **178.7** | **179.7** | **55.6** | **60.1** | **19.0** | **29.1** | 460729 | 756067 | 20 |
| 81 | **7.54** | **115.4** | **179.4** | **177.0** | **56.9** | **55.0** | **30.4** | **17.8** | 475581 | 726196 | 56 |
| 82 | **9.00** | **119.4** | **174.4** | **176.3** | **54.4** | **53.7** | **34.8** | **44.3** | 477934 | 659581 | 13 |
| 83 | **7.62** | **119.1** | 175.3 | 176.9 | 53.1 | **55.6** | **42.4** | **19.0** | 485949 | 740066 | 21 |
| 84 | **8.88** | **120.6** | **176.3** | **181.2** | **55.6** | **63.8** | **17.8** | **40.5** | 486312 | 639348 | 18 |
| 85 | **7.89** | **113.9** | **175.** | **176.** | 56.3 | 53.7 | 46.2 | 19.7 | 507601 | 879585 | 58 G prev |
| 86 | **7.86** | **122.5** | **178.9** | **178.0** | **58.2** | **58.2** | **43.6** | **40.5** | 520259 | 782423 | 73 |
| 87 | **7**.60 | **109.8** | **174.9** | **175.0** | 48.7 | **59.4** | 47.4 | **38.0** | 520760 | 862200 | G78 |
| 88 | **7.97** | **117.9** | **175.0** | **178.0** | **55.6** | **55.6** | **31.7** | **29.8** | 535844 | 770223 | 104 |
| 89 | **8.19** | **110.2** | **178.4** | **175.3** | **60.7** | **61.3** | 59.4 | **31.7** | 545255 | 850729 | 15 |
| 90 | **7.75** | **120.9** | 173.9 | **180.5** | **59.4** | **58.8** | **31.7** | **41.8** | 553288 | 925835 | 54 |
| 91 | **8.92** | **115.4** | **179.2** | **178.3** | **58.2** | **57.5** | **40.5** | **19.7** | 556662 | 915514 | 106 G prev |
| 92 | **8.03** | 110.8 | **177.0** | **175.2** | 47.4 | **56.9** | 46.2 | **30.4** | 637098 | 1020011 | G57 |
| 93 | **7.14** | **114.5** | **177.3** | **175.0** | **59.4** | **59.4** | **38.0** | **27.9** | 650197 | 1024138 | 77 |
| 94 | **8.14** | **123.1** | **179.8** | **177.3** | **59.4** | **59.4** | **27.9** | **28.5** | 653075 | 903416 | 76 |
| 95 | **8.46** | **127.4** | 177.5 | **175.3** | **51.2** | **55.0** | 17.8 | **41.8** | 712969 | 767308 | 2 |
| 96 | **7.23** | **108.6** | **174.9** | **171.6** | **62.0** | **55.6** | **44.3** | **31.7** | 759883 | 1141307 | G81 |
| 97 | **8.41** | **125.9** | **176.9** | 181.1 | **52.5** | **63.8** | **19.0** | **32.3** | 1246263 | 2151414 | 119 |
| 98 | **7.75** | **130.2** | **176.4** | **182.6** | **54.4** | **52.5** | **20.3** | **19.7** | 2090649 | 3027476 | 120 |

searched for XG (non-glycine/glycine) pairs and one which searched for XX (non-glycine/non-glycine) pairs. X is a pseudo-amino acid, with very broad search ranges, but its $C^{\alpha}$ chemical shift range does not extend down into the glycine region. The second residue in these pairs corresponds to residue i in the above chemical shift terminology and the first to residue i–1.

The XX pattern used information from the HNCO, HN(CA)CO, CBCANH and CBCA(CO)NH spectra only. The XG pattern used all these spectra and, in addition, the HNCA and HN(CO)CA spectra. The results list for XG pairs contained seven results, all of which are substantially correct assignments. The expected number of results of this kind is eight. Since the glycine missed by the program was not manually assignable anyway, this can be considered a very satisfactory outcome. The final results list for the XX pairs is shown in Table 3; it contained a total of 99 results, of which 80 were substantially correct assignments. Of the incorrect results, eight involved a glycine, eight were completely unidentifiable, and the remainder contained mixtures of spin systems which shared multiple common chemical shifts. The number of XX results expected from the sequence is 98.

We assume that a sequential assignment tool will be able to make allowances for missing or falsely assigned spins, or disregard those results which contain too many errors. The results lists therefore represent a useful basis for further assignment steps.

## Conclusions

The pattern search program is capable of working on spectra as they are used in everyday work; they do not need to be of especially good quality if more than one spectrum is available to provide some redundancy. Pattern search within the original spectra allows patterns to be found even in cases where there is chemical shift displacement between spectra, or where peaks are weak or nonexistent. The final results list can be subjected to various heuristic filtering operations to remove redundant results and reorder the surviving results according to plausibility. The results presented show that the program can deal with a wide variety of different spectrum types. Output from both side-chain and backbone pattern search experiments are intended to be fed into a sequential assignment tool, instead of peak lists, to reduce the demands on the combinatorial approach. It is expected that this step would provide further filtering of poor results.

## Acknowledgements

This work was supported by Eureka Grant No. BE011/17620A from the Bundesministerium für Forschung und Technologie and by Bruker Analytik GmbH.

## References

Bartels, C., Xia, T., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

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

Billeter, M. (1991) In *Computational Aspects of the Study of Biological Macromolecules by NMR Spectroscopy* (Eds., Hoch, J.C., Redfield, C. and Poulsen, F.M.), NATO ASI Series Vol. A225, Plenum Press, New York, NY, U.S.A., pp. 279–290.

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

Denk, W., Wagner, G., Rance, M. and Wüthrich, K. (1985) *J. Magn. Reson.*, **62**, 350–366.

Dötsch, V., Oswald, R.E. and Wagner, G. (1996a) *J. Magn. Reson.*, **B110**, 107–111.

Dötsch, V., Oswald, R.E. and Wagner, G. (1996b) *J. Magn. Reson.*, **B110**, 304–308.

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

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

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

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

Ikura, M., Kay, L.E. and Bax, A. (1991) *J. Biomol. NMR*, **1**, 299–304.

Kay, L.E., Ikura, M. and Bax, A. (1990a) *J. Am. Chem. Soc.*, **112**, 888–889.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990b) *J. Magn. Reson.*, **89**, 496–514.

Kemmink, J., Darby, N.J., Dijkstra, K., Scheek, R.M. and Creighton, T.E. (1995) *Protein Sci.*, **4**, 2587–2593.

Kemmink, J., Darby, N.J., Dijkstra, K., Nilges, M. and Creighton, T.E. (1996) *Biochemistry*, **35**, 7684–7691.

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

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

Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989) *Biochemistry*, **28**, 6150–6156.

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

Neidig, K.-P. (1992) *AURELIA: Computer aided analysis of 2D and 3D NMR, User's Guide*, Bruker Analytik GmbH, Rheinstetten, Germany.

Nelson, S.J., Schneider, D.M., Di Stefano, D.L. and Wand, A.J. (1991) *Bull. Magn. Reson.*, **13**, 14–21.

Nietlispach, D., Clowes, R.T., Broadhurst, R.W., Yutaka, I., Keeler, J., Kelly, M., Ashurst, J., Oschkinat, O., Domaille, P.J. and Laue, E.D. (1996) *J. Am. Chem. Soc.*, **118**, 407–415.

Oschkinat, H., Holak, T.A. and Cieslar, C. (1991) *Biopolymers*, **31**, 699–712.

Schalkoff, R.J. (1989) *Digital Image Processing and Computer Vision*, Wiley, New York, NY, U.S.A.

Shaw, G.L., Müller, T., Mott, H.R., Oschkinat, H., Campbell, I.D. and Mitschang, L. (1997) *J. Magn. Reson.*, **124**, 479–483.

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

Vuister, G.W., Boelens, R., Padilla, A., Kleywegt, G.J. and Kaptein, R. (1990) *Biochemistry*, **29**, 1829–1839.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.

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

Xu, J., Straus, S.K., Sanctuary, B.C. and Trimbles, L. (1994) *J. Magn. Reson.*, **B103**, 53–58.