

Unpublished manuscript (rejected by J. Biomol. NMR, 1999)

For citation please refer to:

...SPSCAN (<http://www.molebio.uni-jena.de/~rwg/spscan>)

## The program SPSCAN for efficient analysis of macromolecular NMR spectra

Ralf W. Glaser

Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule - Hönggerberg, Zürich, Switzerland, and Institut für Molekularbiologie, Friedrich-Schiller Universität, Jena, Germany

*Keywords:* automated spectral analysis, computer software, NMR structure determination, quantitative spectral evaluation, sequence-specific resonance assignment

*correspondence to:* Institut für Molekularbiologie der FSU, Winzerlaer Str. 10, D - 07745 Jena, Germany. Tel.: +49-3641-657573, Fax: +49-3641-657520, email: [rwg@molebio.uni-jena.de](mailto:rwg@molebio.uni-jena.de)

### Abstract

A new software for semi-automated assignment of NMR spectra and for the extraction of structural constraints has been developed. The program SPSCAN makes extensive use of the information about resonances that is available at a certain stage of the analysis. This information is used in various tools that accelerate the evaluation of further spectra. Automatic functions suggest resonance assignments or other data, but at the same time they display selected areas of the spectra so that the user can decide ambiguous cases interactively. A flexible display organisation allows this combination of automatic functions and user-defined macro programs with interactive work. The approach allows a very fast analysis of NMR spectra, without loss of reliability compared with fully interactive analysis. An object-oriented programming style and comprehensive documentation facilitate changes and additions to the program. In SPSCAN peaks are fitted with Gaussian or Lorentzian lineshapes. The program searches for new peaks in 3D spectra along one-dimensional strips that are obtained by multiplication of the expected peak shape of a 2D spectrum with the spectral intensities of each 3D plane. Small differences in peak position between individual spectra are corrected, and partially overlapping peaks are separated automatically. The algorithm allows an automated determination of relaxation rates, or other data that are obtained from peak volumes in a series of spectra with high accuracy. More information about the program SPSCAN is available under <http://www.molebio.uni-jena.de/~rwg/spscan>.

### Introduction

Sequence-specific resonance assignment and the extraction of distance and dihedral angle constraints from NMR spectra require a large part of the human resources during the determination of macromolecular structures in solution. A wide range of computer programs has been developed to accelerate this process and cope with the large amount of information contained in the spectral data (for a review see Zimmerman & Montelione, 1995). Significant progress toward an automated resonance assignment, in particular for proteins, has been achieved with non-interactive programs. They combine the information obtained from different peaklists and map the resonances to the primary sequence, using algorithms like simulated

annealing (Bartels et al., 1996; Lukin et al., 1997), combinatorial optimisation (Leutner et al., 1998), or constraint propagation methods (Zimmerman et al., 1997). Despite the impressive results of these approaches some limitations are obvious: First, a very high quality of the peaklists is crucial, as the automated program cannot check the spectra in ambiguous situations. Second, the automatic approach differs considerably from an interactive one, so the user cannot interact in critical steps of the process and it is difficult to check the reliability of the final result. Fully automated programs are thus not particularly reliable if possibly the data are ambiguous. Third, even if the sequence-specific resonance assignment can be done automatically, data collection for the calculation of structural and functional properties still requires a large amount of

interactive work. To overcome these limitations by a combination of automated procedures with interactively controlled display tools, a new program, SPSCAN, was developed.

An important design criterion for this program was flexibility. The development of new experiments and new evaluation strategies constantly requires new features of the software. To support user-defined extensions most academic programs like XEASY (Bartels et al., 1995) are available with their source code. Some programs, as the commercial FELIX package (Biosym, San Diego, CA) support a specific command language for implementation of user-developed strategies. Both approaches have their advantages and limitations. Additions to the source code are more flexible than macro commands, but they are mostly limited by lack of documentation and an unsuitable organisation of the program code. SPSCAN has been equipped with a macro language. In addition, an object-oriented programming style and comprehensive documentation facilitate changes and additions to the source code. With this philosophy SPSCAN provides a framework of basic routines to which new functionality can be added at different levels and with minimal effort.

The first part of this paper gives an overview of the functionality of the program. Flexible display functions are a central instrument to combine automatic selection algorithms with an iterative evaluation and comparison of the relevant parts of various spectra. SPSCAN provides a number of semi-automated tools, build on the basis of these display functions. One of these, the tool to evaluate E-COSY spectra, is presented in some more detail as an example. It shows how the user benefits from semi-automated functions that accelerate the time-consuming procedures of an interactive evaluation of spectra without reducing the reliability or accuracy of the results.

A central feature of the new program is its algorithm to search and characterise peaks in a spectrum. Peaks are modelled with Gaussian or Lorentzian lineshapes, and peak-picking in three-dimensional spectra is done along strips, the position of which is defined by a corresponding two-dimensional spectrum. This procedure is similar to strip-oriented work in interactive programs. The implementation and the performance of the peak recognition algorithm are described in the second part.

## Functions and features of the program

### *Spectral display organisation*

SPSCAN can display any number of spectra in any orientation and size. To facilitate a comparison of chemical shifts between these spectral windows the concept of corresponding crosshair coordinates is implemented. A crosshair is a pair of horizontal and vertical lines that can be moved across the window.

It is only connected to the mouse pointer if the left mouse button is pressed, otherwise the pointer can be used independently. The crosshair defines two chemical shift coordinates; in three-dimensional spectra the third coordinate is defined by the selected spectral plane. Correspondence between crosshair coordinates in different windows means the following: If the crosshair is moved or the spectral plane is changed in one window, then the corresponding crosshair coordinates in other windows will change accordingly. Optionally, the spectral plane and/or the displayed part of other windows are adapted to the new coordinates.

This principle is illustrated in Figure 1, which shows schematically how amide peaks in a 2D [ $^{15}\text{N}$ - $^1\text{H}$ ]-COSY spectrum (I) of a protein can be connected interactively to intrasidual alpha peaks in a [ $^{13}\text{C}$ - $^1\text{H}$ ]-COSY spectrum (IV). The connection is made via strips of a  $^{15}\text{N}$ -resolved [ $^1\text{H}$ - $^1\text{H}$ ]-TOCSY (II) and a HNNCA spectrum (III). The user has to press the mouse button on an amide peak in display I, so strips II and III change to the marked  $^{15}\text{N}$  and  $^1\text{H}$  chemical shifts. Now the position of the  $\text{H}\alpha$  peak in strip II and of the  $\text{C}\alpha$  peak in strip III are marked. As a result, display IV shows the area around the putative  $\text{H}\alpha$ - $\text{C}\alpha$  crosspeak. If several assignment possibilities exist due to degeneracy of the amide chemical shifts, the presence of  $^{13}\text{C}$ - $^1\text{H}$  crosspeaks can reveal the correct combination. Correspondence between crosshair coordinates is also useful for different display windows of the same spectrum. A second window can be used either to zoom the current peak (V), or to display part of an orthogonal plane of a 3D spectrum (VI).

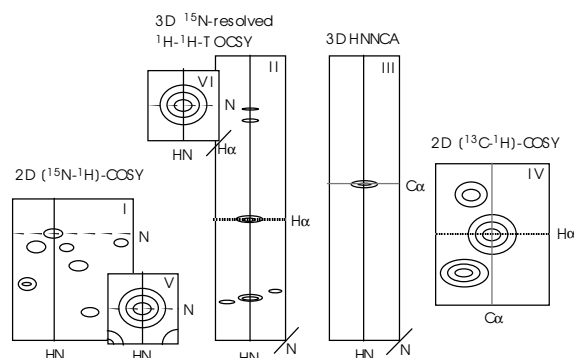


Fig. 1: Use of correlated crosshair coordinates to select relevant parts of different spectra for interactive resonance assignment. For a  $^{15}\text{N}/^{13}\text{C}$  doubly labelled protein the connection between an amide peak in a 2D [ $^{15}\text{N}$ - $^1\text{H}$ ]-COSY spectrum and the intrasidual alpha peak in a 2D [ $^{13}\text{C}$ - $^1\text{H}$ ]-COSY spectrum is established with the two strips from 3D spectra shown in the scheme.

When the dimensions of spectra displayed in two or more windows have the same name, the coordinates are set into correspondence by default. But correspondence can be created, destroyed or exchanged arbitrarily. This is done, for example, to search for transposed peaks in a 3D HC(C)H-COSY spectrum, or to facilitate the assignment of

NOEs between aliphatic and amide protons, using the corresponding crosspeaks in both 3D  $^{15}\text{N}$ -resolved and  $^{13}\text{C}$ -resolved  $[\text{H}-^1\text{H}]\text{-NOESY}$  spectra. The concept of corresponding crosshair dimensions has also been extended to spectra with reduced dimensionality, where it allows the direct visual comparison of zero- and double-quantum coherences in the projected dimension with normal spectra (Szyperski et al., 1998).

### Lineshape comparison

A comparison of lineshapes provides important information about possible resonance assignments of a crosspeak. SPSCAN can automatically provide suitable reference lineshapes from peaks in the relevant range of chemical shifts, using peaks that were previously assigned and have no other peaks nearby which could obscure the shape. The user compares the normalised shapes and selects an assignment possibility with a click of the mouse, which takes only a fraction of the time of an interactive selection of reference shapes. The spectral area from which a shape was taken can be shown and shapes can be extracted interactively.

This routine is used, among other applications, for a utility to check ambiguous resonance assignments made with the NOAH algorithm (Mumenthaler et al., 1997). The NOAH algorithm, which is part of the DYANA package (Güntert et al., 1997), allows an iterative automatic assignment of NOESY peaks based on a list of chemical shifts of the resonances and on a pool of preliminary three-dimensional structures of the molecule. For a number of peaks this algorithm considers several possible assignment combinations. A comparison of their lineshapes with those of unambiguously assigned peaks can resolve the assignment for some of these peaks. Alternatively, the visual check of the

lineshapes may reveal that the ambiguous peak was composed of two overlapping peaks or that the assignment of one of the reference peaks was wrong. Such problems can be corrected immediately, before the program is instructed to present the lineshapes for the next entry in the list of ambiguous peaks.

### Extraction of coupling constants from ECOSY spectra

The correlation of lineshapes is also used for the determination of coupling constants. In E-COSY spectra (Griesinger et al., 1985) the crosspeak between two resonances is split by indirect

couplings to a third resonance into a pattern  $\begin{vmatrix} 1 & 0 \\ 0 & -1 \end{vmatrix}$ .

If the indirect coupling to one of the resonances is large enough so that the two observable components of this pattern do not completely overlap, a small indirect coupling to the other resonance can easily be measured as the displacement between the two components of the crosspeak. The semi-automatic tool to measure the coupling constant is presented here in more detail as a simple demonstration of a close connection of interactive and automated steps (Fig. 2).

This “E-COSY tool” uses a list of previously assigned COSY crosspeaks, or a resonance list from which a list of expected crosspeaks is generated. A library selects the crosspeaks that can be evaluated, and defines their indirect coupling partner. First, the spectral region around a peak is displayed, and the user selects two rectangular areas from the non-overlapping parts of the peak pattern (Fig. 2, left panel). From these areas the

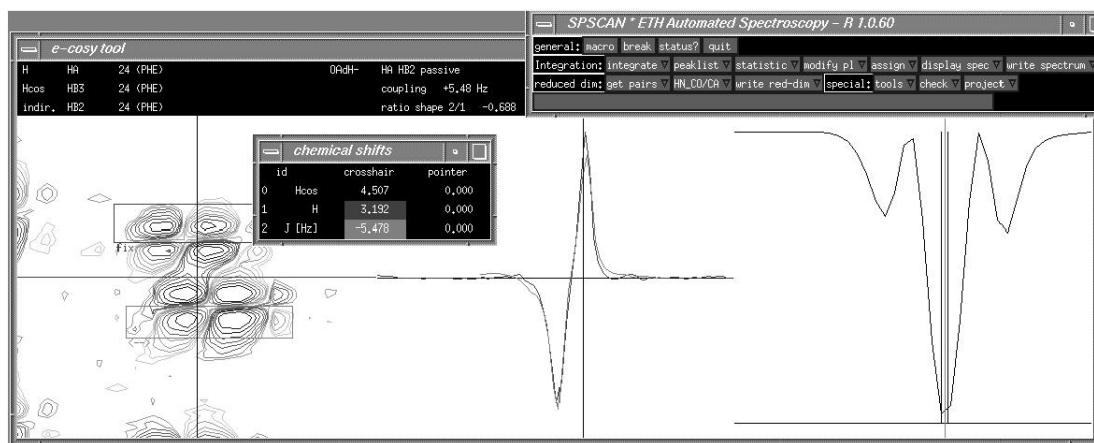


Fig. 2: Tool for evaluation of E-COSY spectra. The main control window of SPSCAN, from which the tool was started, is shown in the upper right. The three display areas of the “e-cosy tool” show the crosspeak pattern with the two user-defined areas to extract the lineshapes (left); the best match between the two lineshapes (centre); and the RMSD between the two scaled lineshapes as a function of displacement (right). The “chemical shifts” window displays the position and colour code of the crosshairs in the various displays.

program extracts the lineshape, i.e. the intensity of all vectors in an area is added or subtracted to the vector with the highest intensity, when the correlation with this vector is positive or negative, respectively. The shapes obtained from the two areas are correlated, and the best correlation is displayed in the central panel (Fig. 2). In the right panel, the sum of squared deviations between the two normalised lineshapes is plotted as a function of their displacement. The best fitting displacement is suggested as the value of the coupling constant, and the displacements with less than twice the minimum RMSD are considered to be an estimate of the error range.

The user can either accept these semi-automatically generated values, or he can displace and scale of the lineshapes with the mouse pointer to find a better fit. The displacement between the lineshapes can also be changed until an obvious mismatch appears, and this position can be used to re-define the error range. However, only in spectral areas with very strong artefacts a better solution than the automatically determined one was found interactively. Normally, the automatically suggested data are accepted, and it takes less than 5 seconds per entry to generate a list of coupling constants that is ready to be used by a structure calculation program. The user has seen all peaks and is sure that the data in the list are reliable and have not been obscured by peak overlap or artefacts.

In a similar way, SPSCAN provides a convenient interface to the program INFIT, which determines scalar coupling constants by inverse Fourier transformation of in-phase doublet peaks (Szyperki et al., 1992).

### **Resonance assignments**

SPSCAN supports resonance assignments in two ways: First, simple automatic routines can identify the assignment of peaks on the basis of chemical shifts in dimensions where resonances have previously been assigned, the type of spectrum, and expected chemical shifts according to a general library of average data. Second, sequential connections between fragments are established with a semi-automatic tool. Information from available peaklists is combined using criteria from a user-defined library. Although the currently implemented approach to determine the probability of a sequential connection is less complex than other recently developed approaches (Bartels et al., 1996, Leutner et al., 1998, Zimmermann et al., 1997), it provides an adequate pre-selection step. All relevant spectral details are presented to the user to make a final decision interactively. The display allows to compare different chemical shifts in parallel. It also allows to reverse the direction of search for sequential neighbours, either to extend pieces of sequentially connected fragments in both directions, or to check other possible neighbours of

a fragment, before a sequential connection is established.

This tool follows the concept of strip-sorting functions in XEASY (Bartels et al., 1995), but SPSCAN provides more flexibility. The probability functions can combine information from an unlimited number of peaklists, and additional information can be considered like the relative volume of a peak compared with other peaks in the strip. The probability matrix of sequential connections can be normalised, so that fragments in which some of the resonances were not observed are adequately represented.

### **Project database**

A database registers the spectra and corresponding peaklists that are used for a particular macromolecule. It contains information about the type, standard orientation and other features of the spectra and keeps track of all resonances assigned to one of the peaks. Possible sequential connections between fragments and types of fragments are stored in probability matrices. This information is provided, modified and used by both automatic and interactive functions of the program. Many functions are available to check the consistency of the entries in the database and present inconsistent data to the user for an interactive correction. Obvious corrections are even made automatically, e.g. changing an atom name from 'QB' or 'HB2' into 'HB' when the fragment type has been changed to valine.

### **General**

SPSCAN is written in C++ and runs on UNIX computers with X-Windows display. The individual tools and routines of the program can be selected from pulldown menus, entered on a command line, or read from macro files. The program manual is written in HTML format, and so are tutorials for the most common applications. A comprehensive documentation of specific classes and methods used in the program is available together with the source code. This facilitates the development of new tools on the source code level and the incorporation of software routines from other, non-interactive programs. Conversely, the object-oriented programming style of SPCAN makes it easy to use selected classes and functions for other software. So a program to evaluate coupling fine structure patterns for metabolic flux analysis by biosynthetic fractional  $^{13}\text{C}$  labelling (Szyperki et al., 1999) was build on the basis of modules from this program.

SPSCAN reads spectra in the formats of the programs XEASY (Bartels et al., 1995), XWinNMR (Bruker Analytische Messtechnik GmbH, Germany), and NMRPipe (Delaglio et al., 1995). Spectra are written in XEASY 16-bit format; so the program can be used to convert spectra into this format. Peaklists, atom lists, sequence files, residue

libraries and lists of coupling constants are read and written in the format used by XEASY and DYANA (Güntert et al., 1997). I recommend to use SPSCAN in combination with these two programs.

In addition to the functionality outlined above, SPSCAN includes many other utilities to extract information from NMR spectra, to manipulate and combine information from different peaklists and atom lists, and to convert data between different formats. A full account of the functions of SPSCAN is available at <http://www.molebio.uni-jena.de/~rwg/spscan>.

## Recognition and Characterization of Peaks

### Integration of peak volumes

Peaks are modelled with a Lorentzian or Gaussian lineshape. They are characterised by the chemical shift,  $\delta_c$ , and linewidth,  $lw$ , in all dimensions of the spectrum, and by their volume,  $V$ . The shape of the peaks is derived from a normalised shape,  $L(x)$  with  $x=(\delta-\delta_c)/lw$ , for which

$$\int_{-\infty}^{\infty} L(p) = 1$$

and

$$L(0.5)=L(-0.5)=0.5L(0) \quad [1]$$

The modelled intensity,  $I_M$ , at a particular pixel,  $p$ , of the discrete spectrum is then

$$I_M = V * I_N \text{ with } I_N = \prod_{d=1}^n (\int_{x1_d}^{x2_d} L(x_d)) \quad [2]$$

where  $x1_d$  and  $x2_d$  are the edges of the spectral pixel in the normalised coordinate system in the spectral dimension  $d$ . With this normalisation a modelled peak with the intensities  $V * I_N(p)$  has the volume  $V$ , even if the linewidth is smaller than the spectral resolution. The volume of an experimental peak is determined from

$$V = \sum_p (I_N * I_E) / \sum_p I_N^2 \quad [3]$$

This sum is calculated only from pixels  $p$  in an area of 3 to 4 linewidths around the peak centre, since for all other pixels  $I_N$  is negligible. The sum of squared deviations between the spectral intensity,  $I_E$ , and the fitted intensity,  $V * I_N$ , is also obtained in the course of this calculation. Contributions to the residual deviation are spectral noise, artefacts, unrecognised peaks, and deviations between real and modelled lineshape. After suitable normalisation it provides a "quality" parameter for the peak, which is used to distinguish peaks from noise.

### Adaptation of chemical shift and linewidth

The algorithm described above requires the knowledge of chemical shifts and linewidths of the

peaks. Therefore, at the start of the assignment process a peaklist has to be produced with another algorithm, or interactively. The linewidth can be estimated. These initial values have to be optimised and adapted to the individual spectra. Therefore, the parameters of the peaks are fitted within certain boundary conditions such that deviations between experimental and modelled peaks are minimised. In an iterative fashion this optimisation is interleaved with the deconvolution of partially overlapping peaks.

### Peak volumes in a series of spectra

The full potential of this algorithm becomes obvious when peaks are compared that differ in intensity, while the other parameters are nearly constant. This is the case in series of spectra recorded for the determination of relaxation times or hydrogen exchange rates. The evaluation of such data is done in a fully automated fashion: The chemical shift and linewidth of the peaks are determined in the first spectrum of the series, or in the spectrum with the best signal/noise ratio. Peak volumes are then determined for each spectrum, while the linewidth for each peak is fixed. The chemical shift can be optimised within narrow boundaries, or it can also be fixed. The volumes from the individual peaklists are combined with the assignment and delay time in a suitable format, and time constants for the individual peaks are determined with an external non-linear fitting program. The individual steps are controlled by a macro program to which the user provides the filenames and the respective delay times of the spectra.

The advantage of this procedure is not only a very rapid evaluation, it also provides highly accurate results that are superior to most alternative evaluation techniques as discussed above. Similar tools exist for related tasks. For example,  $J_3(^{13}\text{C}-^{31}\text{P})$  coupling constants in nucleic acids are calculated from the relative peak volumes with and without  $^{31}\text{P}$  decoupling in constant-time  $^{13}\text{C}-^1\text{H}$  COSY experiments (Legault et al., 1995). Peaks not coupled to  $^{31}\text{P}$  are used to determine systematic and stochastic deviations between the two spectra, which are translated into  $3\sigma$  error ranges for the individual coupling constants.

### Peak search along strips in 3D spectra

An *interactive* search for peaks in a three-dimensional (3D) spectrum is usually guided by knowledge about the peaks in a two-dimensional projection of the experiment. The heteronuclear 3D experiments commonly used to study  $^{15}\text{N}$  labelled or  $^{15}\text{N}/^{13}\text{C}$  doubly labelled proteins are extensions of 2D  $^{15}\text{N}-^1\text{H}$  or  $^{13}\text{C}-^1\text{H}$  one-bond correlation experiments, which can be measured with high sensitivity. Apart from overlaps, peaks that give rise to 3D crosspeaks can easily be identified in the corresponding 2D "base" experiment. At the position of these "base" peaks, "strips" along the

third dimension are extracted from the spectrum, and on these strips the 3D crosspeaks are identified and characterised.

A very similar approach is used by SPSCAN. The 3D spectrum is treated like a series of 2D spectral planes. For each "base" peak a one-dimensional profile along the third dimension is extracted, which consists of the volumes and quality parameters of the respective peak in all spectral planes, calculated as described above. On the basis of these one-dimensional profiles, which correspond to the strips in the interactive approach, crosspeaks are identified. Once crosspeaks have been found, they are used to optimise the chemical shift and linewidth parameters of the 2D base peak for the particular spectrum. With the adapted parameters it may then be possible to identify weaker peaks on the same strip or to separate overlapping 3D peaks.

This approach has a high sensitivity because it uses the intensity across the full area of the expected peak, and not just a 1D intensity vector at the expected peak centre. The intensity of peaks picked on one strip is subtracted from the spectrum to recalculate the profiles of partially overlapping strips, and after 3 to 5 iterations the procedure provides an excellent separation of peaks from neighbouring strips. Based on the knowledge of chemical shift and linewidth of the contributing resonances, even overlapping peaks with a single intensity maximum at an intermediate chemical shift can be correctly separated into their components. The algorithm can be used to determine the volume of peaks with a known three-dimensional position, or to create a new 3D peaklist from a 2D peaklist. Flexibility is achieved with options that limit the range of linewidths and the adaptation of peak positions, the number of peaks per strip and their minimum quality parameters, or search for distinct peak patterns along the third dimension.  $^{15}\text{N}$  and  $^{13}\text{C}$ -resolved NOESY spectra processed in this way provided peaklists for structure calculation of high quality, that needed only a minimum of interactive editing. The program was successfully applied to a large number of spectra, e.g. in the structure determination of the mouse prion protein domain PrP 121-231 (Riek et al., 1996).

In this strip-oriented peak-picking procedure each crosspeak inherits the assignment of its base peak in two dimensions. For some types of spectra this allows to provide the crosspeaks automatically with a complete assignment, and to recognise strips with an unexpected peak pattern, which are labelled for being checked interactively.

### ***Comparison with other peak integration algorithms***

The qualitative and quantitative characterisation of peaks is a critical step in the evaluation of NMR spectra, because peaklists reduce the amount of data by three orders of magnitude, but they should contain all relevant information of the spectra. The

algorithms to recognise peaks in the presence of spectral noise and to determine their parameters differ considerably. To select the best procedure for a particular application, one has to know the features of the various algorithms. The program SPARKY (<http://www.cgl.ucsf.edu/home/sparky>) also uses a peak fitting algorithm, while most other NMR evaluation programs are based on spectral partitioning algorithms. Therefore the systematic and random errors in the determination of peak volumes are discussed below for the peak fitting algorithm of SPSCAN. A series of 11 2D [ $^{15}\text{N}$ - $^1\text{H}$ ]-COSY spectra of the protein P14a with T1 relaxation delays between 15 and 2000 ms (Fernández et al., 1997) were used as example data. Peak volumes were measured in different ways, and T1 relaxation times of backbone  $^{15}\text{N}$  resonances were determined by non-linear least square fitting. Originally, the volumes were obtained by adding the intensity of interactively selected rectangular areas around the peaks with the program XEASY (Bartels et al., 1995). Although this iterative procedure is slow and cannot separate overlapping peaks, it was used by Fernández et al. because it is highly reliability and the systematic errors are small. These data were taken as a reference. To make a comparison between the SPSCAN data and data from a spectral partitioning algorithm, volumes were also determined with the PEAKINT program, that is part of the XEASY package.

The volumes determined with SPSCAN showed systematic deviations from the volumes determined by interactive integration. Using a Gaussian lineshape the volumes were underestimated on average by 6%, and using a Lorentzian shape they were overestimated by 14%. These significant effects ( $p < 0.001$ ) show that the lineshapes that were used to model the peak did not exactly fit the real shape. The deviations are a consequence of describing a resonance only with chemical shift and linewidth, and ignoring coupling finestructure, non-exponential relaxation and processing artefacts.

However, when the lineshape of a peak does not depend on the relaxation delay, all peak volumes in the series of T1 measurements should have the same relative systematic error. Indeed the relaxation times obtained with SPSCAN are very close to the interactively determined data (Fig. 3A). It was necessary, however, to optimise the chemical shift of the peaks for the individual spectra. When the peaks drift away from the position that was determined at the spectrum with the smallest relaxation delay, volumes at longer relaxation delays are underestimated, and consequently the relaxation time is also underestimated (Fig. 3A, SPSCAN data without adaptation of centre). When spectra with longer relaxation delays were used to define the chemical shift of the peaks for the series of spectra, the underestimation of relaxation times vanished, and eventually relaxation times were overestimated (data not shown).

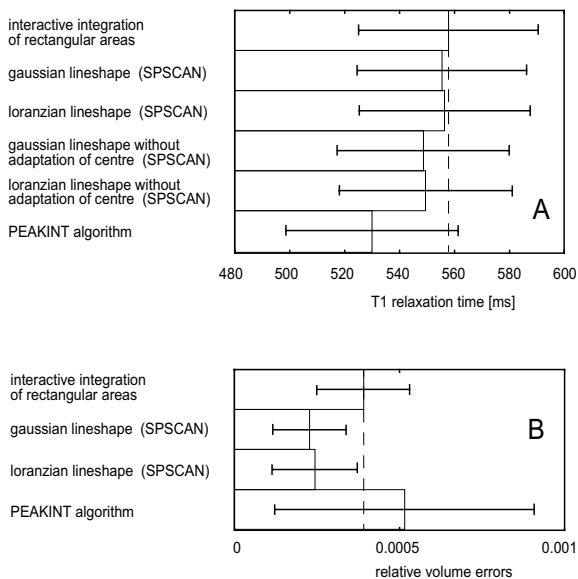


Fig. 3: Comparison of algorithms to extract peak volumes from a series of  $[^{15}\text{N}-^1\text{H}]\text{-COSY}$  spectra of P14a. Volumes  $V$  were determined from spectra with relaxation delays  $t$  between 15 and 2000 ms, and data were fitted to the function  $V(t) = V_{\infty} + V_0 e^{-t/T_1}$ . Average T1 relaxation times and standard deviations for the individual backbone  $^{15}\text{N}$  nuclei are compared between the different algorithms that were used to obtain peak volumes (A). The deviations between measured and fitted volumes,  $\Delta V = V(t) - V_{\infty} - V_0 e^{-t/T_1}$ , are determined to characterise stochastic errors resulting from spectral noise. Squared deviations for all 11 spectra are related to the initial volume. Average and standard deviation of this relative volume error,  $\sqrt{(\sum \Delta V^2 / V_0^2)}$ , is compared between the different algorithms (B).

Determination of volumes with the PEAKINT algorithm even led to a more severe underestimation of relaxation times (Fig. 3A). This algorithm always performs an adaptation of the chemical shifts of the peaks, and the result did not change significantly when different spectra from the series were used to define the starting value. The peak volumes themselves were underdetermined on average by 1.7% compared with interactive integration. In contrast to the SPSCAN algorithm, the relative systematic error was not the same for all spectra, however. The underestimation of peak volumes with PEAKINT results from neglecting points below the noise level and becomes more severe when peak volumes decrease, i.e. in spectra with long relaxation delays. Although different implementations of spectral partitioning algorithms exist, some of which may have smaller errors than PEAKINT, it remains a general feature of this type of algorithms that their systematic errors depend on the relative size of peaks and noise. In contrast, the relaxation times determined with SPSCAN (with automatic adaptation of peak centre) are virtually free of systematic deviations, despite the systematic errors in peak volumes.

Signal theory predicts that the optimal procedure to parametrise a signal after it has been corrupted by the addition of white noise is the correlation of the

expected shape of the signal with the experimental signal (Wozencraft & Jacobs, 1967). The shape fitting algorithm of SPSCAN is very close to this theoretical optimum. The stochastic error that results from spectral noise should therefore be smaller than in other methods. As the true volume of the peaks without noise is not known, the individual volumes were compared with the data from the exponential fit through all 11 data points in the relaxation series, in which the stochastic error is reduced by averaging. The relative deviation between the two values (Fig. 3B) measures the stochastic error, apart from a small contribution that can be attributed to deviations from exponential decay. For the shape fitting algorithm of SPSCAN it was about 40% smaller than for the interactive integration of rectangular areas. The performance of the PEAKINT algorithm was worse than interactive integration. It can be concluded that the peak fitting algorithm of SPSCAN provides a better suppression of spectral noise than interactive integration.

As a matter of practical importance, however, the most striking advantage of using SPSCAN was the time saved. The evaluation of all 11 spectra took only a few minutes of interactive work (to check the correct adaptation of the peaklist and to enter the relaxation delays together with the names of the spectra into a template macro program), and 2 seconds of CPU time on an SGI Indigo-2 workstation.

The systematic errors in the peak volumes discussed above are not really important for many evaluations. For example, in the calculation of distance constraints with  $r \sim V^{-1/6}$  the resulting distance error is less than 3%. Nevertheless, the peak fitting algorithm is not advantageous for all applications. The strength of the algorithm is the use of previously obtained information about the peaks that are expected in the spectrum. If no such information is available or if the peak shape cannot be described as a Gaussian or Lorentzian line, I recommend to use the program AUTOPSY (Koradi et al., 1998) for the recognition of peaks. This program makes no assumptions about the lineshape apart from symmetry properties, and it uses global features of the spectrum to distinguish peaks from artefacts and noise. However, AUTOPSY cannot use peaklists or resonances as an input, so it is not very suitable to collect information about particular, previously assigned peaks.

## Conclusions

SPSCAN is a new addition to the ETH suite of academic programs that cover the complete evaluation of NMR spectra from transformation with PROSA (Güntert et al., 1992) to the calculation of three-dimensional macromolecular structures with DYANA (Güntert et al., 1997). In particular, it complements the display and assignment program

XEASY (Bartels et al., 1995) with a number of new automatic routines and with convenient, highly flexible display tools. The new program does not aim at a fully automatic evaluation. Advantages of automated and interactive evaluation are combined in such a way, that the speed of the evaluation process is increased without reducing reliability. Ambiguous data are presented for an interactive control and for possible user intervention.

A very noise-resistant algorithm for peak recognition and for the integration of peak volumes is implemented in SPSCAN. The algorithm makes optimal use of the information that is available at a certain stage in the evaluation of the spectra. The automatic evaluation of relaxation experiments or similar series of 2D spectra with this algorithm save not only time, but also provide a higher accuracy than interactive evaluation.

The program provides a very flexible framework to evaluate new types of NMR spectra and implement novel semi-automatic evaluation strategies. Flexibility is achieved at several levels. A macro language and user-defined libraries adapt the program to the tasks and conditions of a specific project. Object-oriented programming in C++ facilitates adaptations on the source code level. Comprehensive documentation is available to allow the introduction of new tools and features with minimal effort in any laboratory.

## Acknowledgements

This work was supported with grant GI-237/1-1 of the Deutsche Forschungsgemeinschaft. I thank Prof. K. Wüthrich for his continuous support, and all members of the NMR group at the Institute of Molecular Biology and Biophysics of the ETH Zürich as well as many other users for testing development versions of the program and contributing spectra and other unpublished raw data. Many functions and features of the program were developed as the result of stimulating discussions with M. Billeter, F. Damberger, C. Fernández, P. Güntert, R. Koradi, R. Riek and T. Szyperski.

## References

- Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR* **7**, 207-213.
- Bartels, C., Xia, T.H., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomolecular NMR* **5**, 1-10.
- Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR* **6**, 277-293.
- Fernandez, C., Szyperski, T., Bruyere, T., Ramage, P., Mösinger, E. and Wüthrich, K. (1997) *J. Mol. Biol.* **266**, 576-593.
- Griesinger, C., Sorensen, O. W., and Ernst, R. R. (1985) *J. Am. Chem. Soc.* **107**, 6394-6396
- Güntert, P., Dötsch, V., Wider, G. and Wüthrich, K. (1992) *J. Biomolec. NMR* **2**, 619-629.
- Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.* **273**, 283-298.
- Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Resonance* **135**, 288-297

- Legault, P., Jucker, F. M. and Pardi, A. (1995) *FEBS Lett.* **362**, 156-160
- Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR* **11**, 31-43.
- Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR* **9**, 151-166.
- Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR* **10**, 351-362.
- Riek R., Hornemann, S., Wider, G., Billeter, M., Glockshuber, R. and Wüthrich, K. (1996) *Nature* **382** (1996) 180-182.
- Szyperski, T., Güntert, P., Otting, G. and Wüthrich, K. (1992) *J. Magn. Resonance* **99**, 552-560.
- Szyperski, T., Banecki B., Braun, D. and Glaser, R.W. (1998) *J. Biomolecular NMR* **11**, 387-405.
- Szyperski, T., Glaser, R. W., Hochuli, M., Fiaux, J., Sauer, U., Bailey, J. E., and Wüthrich, K. (1999) *Metabolic Engineering* (in press)
- Wozencraft, J. M. and Jacobs, I. W. (1967) *Principles of Communication Engineering*, Wiley, N.Y.
- Zimmermann, D.E. and Montelione, G.T. (1995) *Current Opinion Struct. Biol.* **5**, 664-673.
- Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.* **269**, 592-610.