Journal of Biomolecular NMR, **27:** 69–79, 2003.
KLUWER/ESCOM
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Fully automated sequence-specific resonance assignments of heteronuclear protein spectra

Daniel Malmodin[a], Christina H.M. Papavoine[b] & Martin Billeter[a,*]
[a]*Biochemistry and Biophysics, Göteborg University, Box 462, 405 30 Göteborg, Sweden*
[b]*Medicinal Chemistry, AstraZeneca R&D Mölndal, 431 83 Mölndal, Sweden*

Received 25 February 2003; Accepted 18 April 2003

Key words: automation, AUTOPSY, azurin, GARANT, peak picking, resonance assignments, structure genomics

Abstract

Full automation of the analysis of spectra is a prerequisite for high-throughput NMR studies in structural or functional genomics. Sequence-specific assignments often form the major bottleneck. Here, we present a procedure that yields nearly complete backbone and side chain resonance assignments starting from a set of heteronuclear three-dimensional spectra. Neither manual intervention, e.g., to correct lists obtained from peak picking before feeding these to an assignment program, nor protein-specific information, e.g., structures of homologous proteins, were required. By combining two earlier published procedures, AUTOPSY [Koradi et al. (1998) *J. Magn. Reson.*, **135**, 288–297] and GARANT [Bartels et al. (1996) *J. Biomol. NMR*, **7**, 207–213], with a new program, PICS, all necessary steps from spectra analyses to sequence-specific assignments were performed fully automatically. Characteristic features of the present approach are a flexible design allowing as input almost any combination of NMR spectra, applicability to side chains, robustness with respect to parameter choices (such as noise levels) and reproducibility. In this study, automated resonance assignments were obtained for the 14 kD blue copper protein azurin from *P. aeruginosa* using five spectra: HNCACB, HNHA, HCCH-TOCSY, ^{15}N-NOESY-HSQC and ^{13}C-NOESY-HSQC. Peaks from these three-dimensional spectra were filtered and calibrated with the help of two two-dimensional spectra: ^{15}N-HSQC and ^{13}C-HSQC. The rate of incorrect assignments is less than 1.5% for backbone nuclei and about 3.5% when side chain protons are also considered.

Abbreviations: AUTOPSY – a program for automated peak picking; GARANT – a program for resonance assignments; PICS – a program for calibration and filtering of peak lists

Introduction

High-resolution NMR is a very versatile tool for studies of proteins, capable of providing a wealth of different information including knowledge on three-dimensional (3D) structures (reviewed in Güntert, 1998), on internal dynamics (reviewed in Korzhnev et al., 2001) or on ligand binding (reviewed in Hajduk et al., 1999; Härd, 1999). This diversity in applications is reflected in a wide variety of different NMR experiments. Thus, techniques have been devised for either homonuclear or heteronulcear NMR, for studies of ligand binding or relaxation, for characterizations of the backbone only or for complete structure determinations, and many more. Common to most applications is the need for sequence-specific assignments. These usually form the most time-consuming step (Wüthrich, 1986). Especially in view of high-throughput NMR studies on proteins there is a clear need for automated tools. The ultimate goal is a fully automated procedure that can be applied to a variety of combinations of spectra. The latter is important when considering practical aspects such as the lack of labeling. Other examples are backbone relaxation or ligand-binding studies where the assignments of side chain nuclei would be wasted efforts, in con-

*To whom correspondence should be addressed. E-mail: martin.billeter@bcbp.gu.se

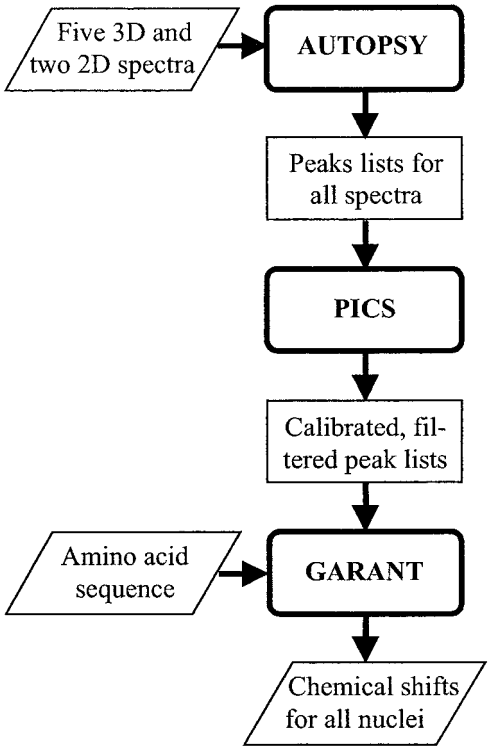

Figure 1. Flow-chart summarizing the automated assignment with the three programs AUTOPSY (Koradi et al., 1998), PICS and GARANT (Bartels et al., 1996, 1997). Input and output files are indicated by tilted rectangles, intermediate data lists by normal rectangles. The latter were passed unmodified from one program to the next. Input spectra were provided to AUTOPSY in the XEASY-format (Bartels et al., 1995).

trast to determinations of 3D structures. Furthermore, the method should be robust with respect to spectral noise, artifacts, the interpretation of weak signals and the choice of parameters. Advantages inherent to automation are simultaneous consideration of large quantities of information, systematic and iterative data evaluation and the use of libraries and databases. Disadvantages are often the absence of possibilities to reconsider earlier obtained data, which ideally could be avoided by feedback options.

Triggered by the considerable potential gains in time and other investments, many efforts have been made in the past to achieve the above goal (Moseley and Montelione, 1999; Moseley et al., 2001). Often however, partial problems were addressed. Examples are the identification of cross peaks (Koradi et al., 1998), the identification of spin systems (Li and Sanctuary, 1997a, 1997b), their sequential assignment (Billeter et al., 1988), or on the connection of short fragments of spin systems (Güntert et al., 2000). The input was mostly based on measurements of scalar couplings and NOEs, but also of chemical shifts (Gronwald et al., 1998; Atreya et al., 2000) or dipolar couplings in weakly aligned systems (Tian et al., 2001). Very often, specific sets of experiments are required (Buchler et al., 1997; Lukin et al., 1997; Zimmerman et al., 1997; Leutner et al., 1998; Atreya et al., 2000). The choice of algorithms was extensive, including statistical approaches (Lukin et al., 1997; Zimmerman et al., 1997), genetic or evolutionary algorithms (Bartels et al., 1996; 1997), exhaustive searches (Güntert et al., 2000), and also more specific and genuine ideas (Billeter, 1991).

The approach presented here combines two published algorithms, AUTOPSY (Koradi et al., 1998) and GARANT (Bartels et al., 1996; 1997), with a new program called PICS to achieve most of the goals set forth above. Complete automation is assured from the input consisting of various spectra to the output, a list with sequence-specific assignments. The approach is general in the sense that it relies neither on a particular set of spectra nor on protein-specific data. Tests were performed on real spectra. It is shown that the choices of run-time parameters (e.g., defining noise levels) are not critical.

AUTOPSY (Koradi et al., 1998) combines novel noise detection, symmetry considerations and lineshape comparisons for optimal picking of peaks in multidimensional spectra. The results are presented as lists of peaks with chemical shifts, intensities and a reliability measure for each peak. To our knowledge, the present report includes the first applications of AUTOPSY to three-dimensional spectra. GARANT (Bartels et al., 1996, 1997) is a program in which sequence-specific assignments are the result of optimally matching a list with peaks expected from the amino acid sequence to a list with experimental peaks (Billeter, 1991). It furthermore relies on a scoring scheme and optimization based on an evolutionary algorithm. The major result is a list of nuclei, each with possibly several assignment possibilities and their respective probabilities.

The goal of the present approach is a flexible tool that yields exactly those sequence-specific assignments that are required in the context of a given project. As part of a study of backbone relaxation, the resulting assignments could consist of resonances of the backbone nuclei from an input of triple-resonance spectra. As part of a structure determination, a wider range of input spectra would be required in order to achieve assignments for side chain nuclei. In the latter

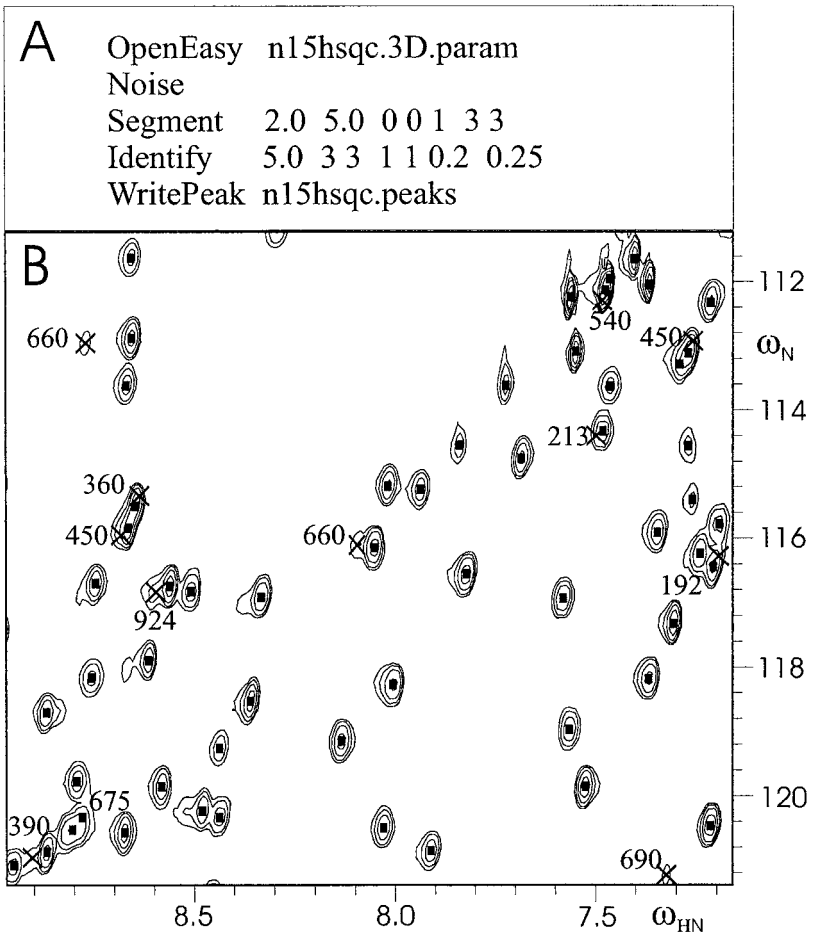

Figure 2. Illustration of peak picking by AUTOPSY using the 2D ^{15}N-HSQC. (A) AUTOPSY macro used for the peak picking (see Koradi et al., 1998). The commands and their parameters perform (i) reading of the XEASY-formatted spectrum with name 'n15hsqc'; (ii) determination of local noise levels; (iii) segmentation into connected regions with intensity exceeding the noise level; (iv) identification of peaks; (v) output of the resulting peak list 'n15hsqc.peaks'. The parameters 2.0 and 5.0 are factors of the noise level, '3 3' indicates the minimal size of a region or a peak, and 0.25 is a measure of the allowed symmetry deviation within each peak. The remaining parameters refer to aspects that are not relevant for this spectrum, e.g. the handling of diagonals. (B) Central region of the ^{15}N-HSQC with peaks picked by AUTOPSY using the macro and parameters given in part A. Dots identify picked peaks that can be assigned to backbone or side chain (upper left corner) amides of azurin. Crosses mark additionally picked peaks by AUTOPSY. Numbers besides the crosses and next to one of the dots ('675' in the lower left corner) indicate in how many of the 1125 parameter sets of a systematic (see text) the corresponding peak was picked. All peaks marked by dots with no number were picked by all parameter sets. Only positive contour levels are drawn.

case, the next step would probably consist of combined NOESY assignment and structure calculations, which has recently received a lot of attention (Güntert, 1998; Linge et al., 2001; Gronwald et al., 2002; Herrmann et al., 2002). The use of the results from the presently discussed automation of sequence-specific resonance assignments in conjunction with tools for characterization of backbone dynamics, structure determination or others will be discussed elsewhere.

Methods

A summary of the spectra that were recorded for the 14 kD blue copper protein azurin from *P. aeruginosa* is given in Table 1. All spectra were obtained at pH 5.5 and 30 °C for the same ^{13}C and ^{15}N doubly labeled sample with 1 mM concentration. Spectra were processed with the NMRPipe software (Delaglio et al., 1995) using twofold zero-filling along all dimensions. Spectral files in the format of XEASY (Bartels et al., 1995) were produced. Figure 1 provides an overview

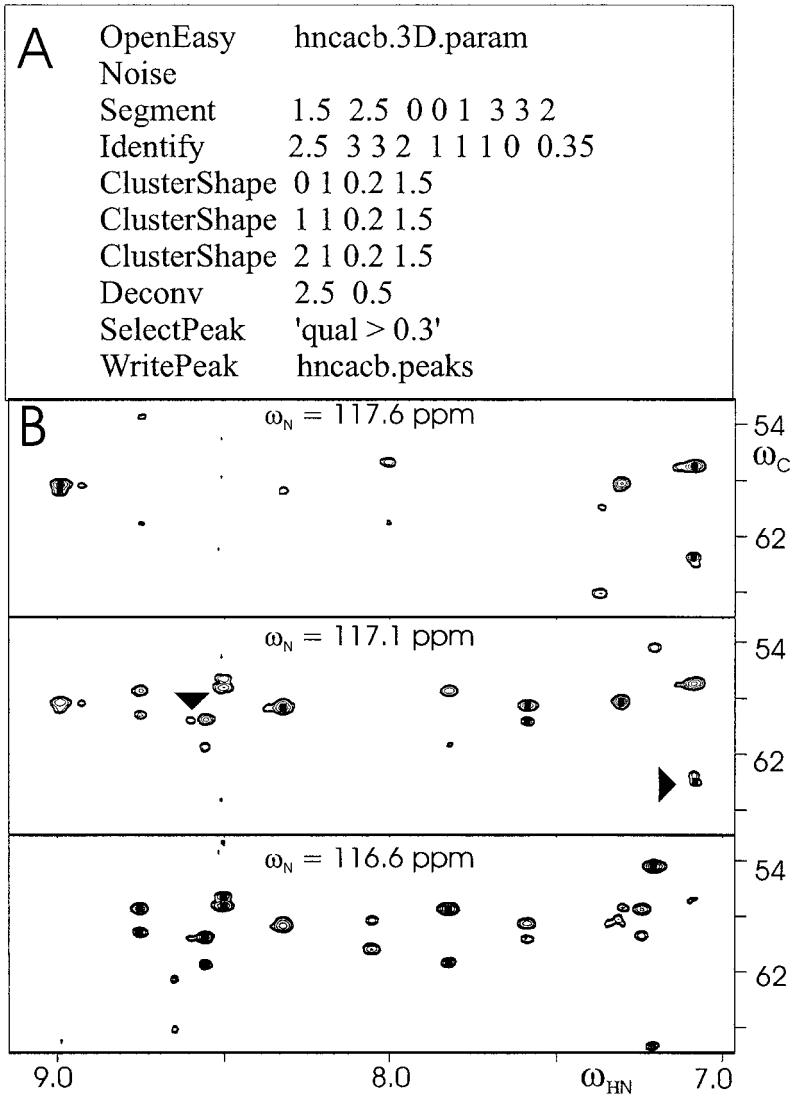

Figure 3. Illustration of the peak picking by AUTOPSY applied to the 3D HNCACB. (A) AUTOPSY macro used for the peak picking (see Koradi et al., 1998). In addition to the commands explained in the caption of Figure 2, three additional types of commands are applied: a first one to cluster 1D cross-sections of peaks for the use in the second command, where deconvolution is applied to overlapped regions in order to resolve small peaks. A third new command removes peaks with low quality factors. (B) Selected region of the HNCACB spectrum with three adjacent planes identified by their ω_N shifts. Dots indicate picked peaks that can be assigned to expected peaks. In the middle plane, the arrow to the left identifies a peak that was picked by AUTOPSY but misses an assignment in the reference, while the arrow to the right points to a peak that was missed by AUTOPSY.

of the individual processing steps, the programs involved and the data flow. The programs AUTOPSY (Koradi et al., 1998) and GARANT (Bartels et al., 1996, 1997) were applied as described earlier with the exception of the extension of AUTOPSY to three-dimensional spectra and the introduction of spectra of type HNHA (Vuister and Bax, 1993) to GARANT.

A new program, PICS (Peak Improvement by Calibration and Selection), was implemented for the improvement of the peak lists obtained from AUTOPSY prior to their use within GARANT. It performs three tasks: (a) It controls the proper position of the diagonal in each peak list; (b) it calibrates pairs of peak lists relative to each other; (c) it filters the lists from the 3D spectra by selecting only peaks for which a corresponding signal in a 2D spectrum is present. Relative calibration of two spectra with two dimensions in common was achieved by tabulating shift differ-

Table 1. Overview of spectra used and result of peak picking

Spectrum	Size of acquired data[a]	Number of picked peaks[b]	
^{15}N-HSQC	2048*128		169 [150]
^{13}C-HSQC	512*256		667
HNCACB	896*220*40	Positive	246 [246]
		Negative	223 [224]
HNHA	512*64*32	Positive	128
		Negative	176
HCCH-TOCSY	1024*200*40		1461
^{15}N-NOESY-HSQC	512*200*40		1773
^{13}C-NOESY-HSQC	512*200*40		3141

[a]Complex data points in time domain.
[b]Number of peaks picked by AUTOPSY after filtering with PICS. Numbers in brackets indicate the number of expected peaks based on the sequence (see text). For the HNCACB and the HNHA, separate entries are given for positive (HN-N-Cα and HN-N-Hα) and negative (HN-N-Cβ and HN-N-HN) peaks.

ences between peak pairs with one peak taken from each spectrum. All pairs of peaks with similar shifts in each of the two dimensions were considered. A constant correction that centers these distributions of shift differences on zero is then added to all shifts of one spectrum. For these relative calibrations, an order of spectra is defined. In the present application, the ^{15}N-HSQC was chosen as starting point. For a next group of three spectra, HNCACB, HNHA and ^{15}N-NOESY-HSQC 3D, the ω_{HN} and ω_N axes were calibrated to the initial spectrum, while preserving the ^{1}H-^{1}H diagonal in the latter two spectra. In order to proceed with the third group of spectra, ^{13}C-NOESY-HSQC and HCCH-TOCSY, an artificial peak list with ω_{HC} and ω_C entries was created by combining the HNCACB and the HNHA peak lists. This allowed calibrations of both the ^{1}H and the ^{13}C dimension in the third group of spectra. Finally, the ^{13}C-HSQC was calibrated with respect to the ^{13}C-NOESY-HSQC.

For the evaluation of the result of the automated assignment procedures, independent resonance assignments were used. Due to the change of pH and temperature with respect to published resonance assignments (Leckner, 2001), azurin was reassigned manually using the same spectra as for the automated procedure. This reference assignment is complete for the backbone and almost complete for the side chains, but no stereospecific assignments are included. For a more thorough analysis of the peak picking results from AUTOPSY, the reference assignment was used to construct lists of expected peaks for the 2D ^{15}N-HSQC and the 3D HNCACB spectra. The reference list of expected peaks for the 2D ^{15}N-HSQC contains 123 backbone peaks and 27 side-chain peaks for six Gln, seven Asn and one Trp; ^{15}N-^{1}H peaks from Arg, Lys and His side chains were not available. For the 3D HNCACB, the expected peak list consists of 246 backbone peaks involving α-carbons and 224 backbone peaks involving β-carbons.

Results and discussion

Overview of the assignment approach

The present approach for fully automated sequence-specific assignments consists of three steps (Figure 1). First, the program AUTOPSY (Koradi et al., 1998) is applied to a number of mainly three-dimensional spectra (Table 1); this is to our knowledge the first successful application of AUTOPSY to 3D spectra. Next, a new program, PICS, performs a series of intermediate processing steps to increase consistency between the peak lists obtained from AUTOPSY. These improved peak lists then form the input to the program GARANT (Bartels et al., 1996, 1997), which yields the final result: Sequence-specific resonance assignments. The entire assignment relies exclusively on automated procedures. Apart from the spectra and the amino acid sequence no additional information is used. This procedure is presented below for the 14 kD protein azurin, and discussed in view of the choice of run-time parameters for AUTOPSY and GARANT and of the quality of the assignments obtained.

Peak picking with AUTOPSY of the ^{15}N-HSQC

The 2D ^{15}N-HSQC recorded for azurin provides a reliable list of valid shift combinations of ^{1}HN and ^{15}N for all ^{15}N-edited 3D spectra, and it was therefore chosen as the starting spectrum for calibrations of the 3D spectra (see below). This spectrum also serves as a short illustration of AUTOPSY and a systematic analysis of the program's performance when changing run-time parameters in the macro given in Figure 2A. The first and the last commands provide AUTOPSY with filenames of the spectrum (in XEASY-format; Bartels et al., 1995) and the resulting peak list, respectively. *Noise* level determination in AUTOPSY requires no parameters. The *Segment* command identifies regions of interest. In the macro of Figure 2A these regions consist of neighboring spectral data points with intensities exceeding twice the local noise level and a minimal size of three-by-three points. At least one of these points must exceed the noise by a factor of five. The *Identify* command picks peaks in the above-determined regions. A peak maximum must exceed the noise five times and the minimal size of the peak is again three-by-three data points. The last parameter defines a symmetry requirement of the line shape of the peak. (All other parameters are not of relevance here; see Koradi et al., 1998.)

Part of the result from the AUTOPSY peak picking with this macro is shown in Figure 2B. Compared to a manually defined reference peak list with 123 backbone and 27 side-chain peaks (see Methods), AUTOPSY missed one side-chain peak in a heavily overlapped region and picked 20 additional peaks. Eleven of the latter are indicated by crosses in Figure 2B. Apart from small but real peaks, AUTOPSY sometimes picked more than the true number of peaks in overlapped regions. Typically one of the extra peaks had significantly smaller amplitude, which is indicative for line shape distortions of the true peaks. These additional peaks closely coincided with true peaks and had therefore little effect on the further assignment process. The parameters of Figure 2A may be derived by consideration of spectral characteristics such as intensity distribution or spectral resolution, but the easiest and fastest method is visual inspection of a few peaks in the spectrum. The robustness of the peak picking with respect to the parameter choices was demonstrated by systematic variation of all parameters described above. Thus, the two noise-related factors in Figure 2A were varied in the interval 1.4–2.6 with a step of 0.3, and 4.0–6.0 with a step of 0.5, respectively.

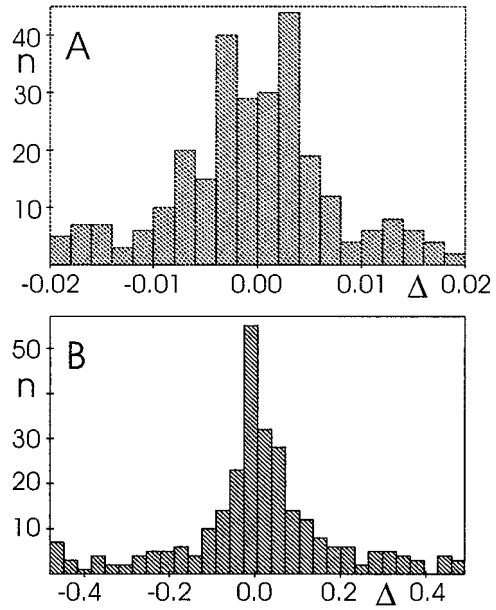

Figure 4. Distribution of shift differences, Δ, between peaks from a peak list obtained by combining data from HNCACB and HNHA, yielding an artificial HACACB peak list (see text), and the peak list from the ^{13}C-HSQC-NOESY. Shown are the distributions after calibration for the proton (A) and the carbon (B) dimensions.

The region and peak sizes were changed independently in each dimension from 2–4 data points, and the symmetry parameter was varied in the interval 0.15–0.35 with a step of 0.05. The resulting 1125 different parameter sets yielded peak lists in which only three backbone peaks were not detected by all parameter sets. One weak peak was missed when both noise-related factors were simultaneously increased (one of the two strongly overlapping peaks in the lower left of Figure 2B), and two strongly overlapped peaks were missed when the symmetry criterion was tightened in combination with an increase of some of the other parameters.

Peak-picking with AUTOPSY of the HNCACB

The macro used for the HNCACB spectrum is given in Figure 3A. From the total of 246 expected Cα peaks and 224 expected Cβ peaks (see Methods), AUTOPSY missed 7, respectively 15, peaks. The number of additionally picked peaks after filtering with PICS (see below) was of the same order. Figure 3B displays for a selected spectral region three neighboring planes with different nitrogen shifts. Besides a number of picked peaks that can be assigned to expected peaks, this region contains one example each for a peak that is

missed by AUTOPSY and a peak that is picked in addition, possibly originating from a side chain (see arrows in Figure 3B).

A systematic analysis of the parameters in the macro was also performed for the HNCACB spectrum in order to check the robustness of the peak picking. AUTOPSY was run 108 times testing all combinations of the following parameter choices. The minimum intensity of a data point in a segment was given values of 1.2, 1.5 and 1.8 times the noise level (first number of the *Segment* command in Figure 3A). The minimum peak height was tested for values of 2.0, 2.5 and 3.0 times the noise level (second number of *Segment*, and first number of *Identify* and *Deconv* in Figure 3A). The minimal sizes in the proton and carbon dimensions were either two or three data points (″3 3″ in *Segment* and *Identify*); the minimal peak size in the nitrogen dimension was kept unchanged at a value of 2. The symmetry requirement was tested for values of 0.2, 0.3 and 0.4 (last number in *Identify*). The deconvolution function was used in order to find smaller peaks using information from larger peaks. All peaks with a quality factor lower than 0.3 were removed. The results showed that 209 positive and 169 negative expected peaks were always picked, while 224 positive and 188 negative expected peaks were picked in at least 90% of the test runs. 231 positive and 202 negative expected peaks show up in at least 50% of the 108 runs. Additional peaks not expected from the sequence but picked in 90% or more of the test runs consist of 4 positive peaks and 11 negative peaks. The missed and the additional peak in Figure 3B (see above and figure caption) were found in 72 and 48 of the 108 test runs, respectively.

Peak-picking in the other 3D spectra

For the peak lists of the other three-dimensional spectra, HCCH-TOCSY, HNHA, ^{15}N-NOESY-HSQC, ^{13}C-NOESY-HSQC, the same macro was used as for the HNCACB with the following exceptions. For the HNHA the deconvolution function (see Figure 3A) was not used. For the two NOESYs the minimal intensity for a peak maximum was lowered to 2.0 and the minimal peak size to 2*2*2 in order to also allow peak picking of small NOEs.

Improvement of the peak lists with PICS

The first task of the program PICS was to properly position the diagonal in spectra with two proton frequencies. This was done by calculating signed (left, respectively right of the diagonal) distances in ^{1}H-^{1}H planes of the positions of all peaks to the diagonal. One proton axis was then shifted to ensure that the central part of this distance distribution with distances $<$0.01 ppm was evenly distributed around zero. The size of these shift corrections was small, maximally 0.003 ppm, which corresponds to about one data point in the directly detected proton dimensions.

In a next step, the ^{15}N-^{1}H planes in the HNCACB, HNHA and ^{15}N-NOESY-HSQC 3D peak lists were calibrated with respect to the ^{15}N-HSQC peak list. For all peak pairs with one peak from each spectrum, for which the differences in ω_H and ω_N are smaller than 0.01 and 0.15 ppm, respectively, the signed differences were tabulated for each dimension. These difference distributions were used to change shifts in the 3D peak lists such that the distributions became symmetric about zero (without moving diagonal peaks away from the diagonal). The ^{13}C-NOESY-HSQC and the HCCH-TOCSY peak lists were then calibrated with respect to the other 3D peak lists. For this purpose, an artificial peak list was constructed from a copy of the HNCACB peak list by replacing the HN-frequencies with the corresponding Hα frequencies using the HNHA peak list. Distributions were defined for shift differences $<$0.01 ppm in ω_H and $<$0.3 ppm in ω_C. The 2D ^{13}C-HSQC was calibrated with respect to the ^{13}C-NOESY-HSQC. Figure 4 shows the difference distributions along ω_H and ω_C used for calibrating the ^{13}C-NOESY-HSQC peak list with respect to the artificial peak lists obtained from the HNCACB and HNHA lists. The widths of these distributions correspond closely to the resolutions along the respective dimensions in the spectra ($\sim$0.01 ppm in ω_H and $\sim$0.2 ppm in ω_C). Thus, for the combined use of several peak lists in the assignment by the program GARANT (see below), it was not necessary to increase the uncertainty of peak positions derived directly from the resolution of each spectrum. It should be noted that the cutoff values for peak selection used for the distributions in this and the above step are not critical since the distributions have a very high maximum.

Finally, peaks in the 3D peak lists, which did not have a corresponding 2D peak in the ^{15}N-HSQC peak list or the ^{13}C-HSQC peak list, were removed. Thus, all peaks in the HNCACB, HNHA and the ^{15}N-NOESY-HSQC peak lists not having a ^{15}N-HSQC peak within 0.4 ppm in the ^{15}N-dimension and 0.05 ppm in the ^{1}H-dimension were eliminated. Because the ^{13}C-HSQC is very crowded and exhibits

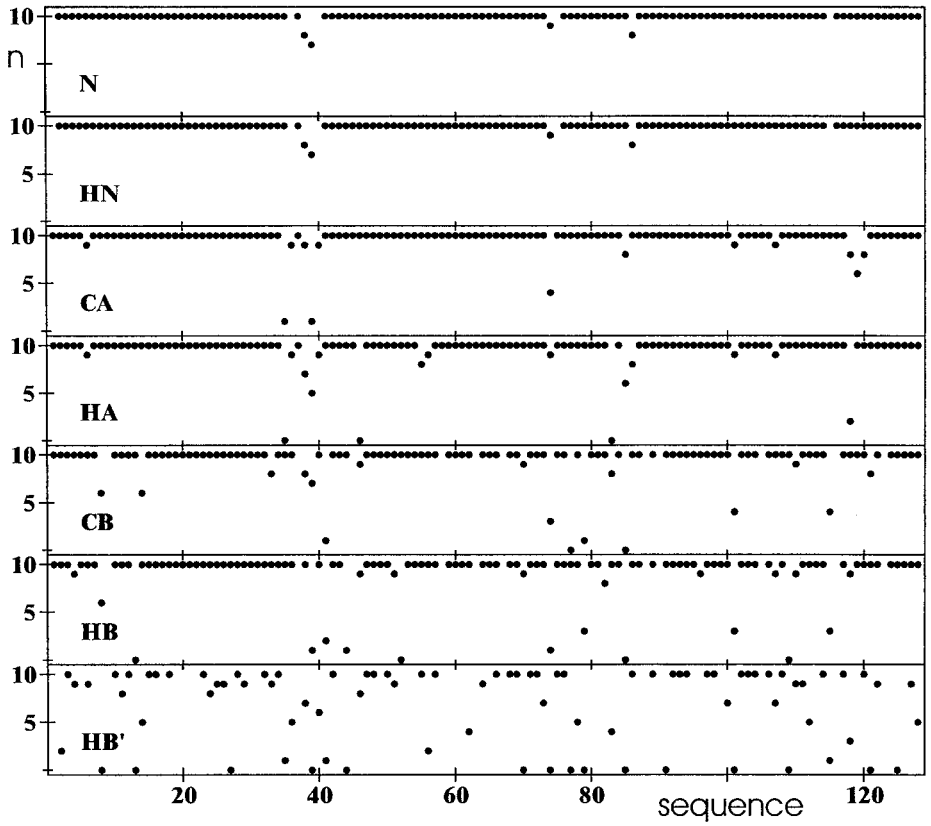

Figure 5. Backbone, Cβ and Hβ assignments by GARANT vs. the sequence. For each residue and nucleus, the position of a dot along the vertical axis indicates the number of GARANT runs with correct assignments. The total number of runs was ten.

large differences in peak intensities, the peak picking may miss small peaks in the neighborhood of large ones. Therefore, relatively large cutoffs of 1.5 ppm in the carbon dimension and 0.15 ppm in the proton dimension were used for peak removal in the HCCH-TOCSY and the ^{13}C-NOESY-HSQC peak lists with respect to the ^{13}C-HSQC peaks.

Sequence-specific assignments from GARANT

The command macro that steers a GARANT run requires only a few parameter choices. The program employs a random generator that needs an initial seed. For the genetic algorithm the size of the 'population' of resonance assignments was set to 100. The sequence of the protein and the names of the (unassigned) peak lists from the five 3D spectra after processing with PICS (Table 1) represented the main input for GARANT. The accuracy of peak positions within a spectrum was set equal to one data point, but not less than 0.01 ppm. As argued above, it is reasonable to assume after relative calibration with PICS the same accuracy between different spectra.

GARANT was run ten times using different seeds but otherwise the same input. The criterion for an assignment of a nucleus by GARANT was that six or more of the ten runs report the same shifts using cutoffs of $\Delta\delta_H = 0.05$ ppm, $\Delta\delta_N = 0.4$ ppm and $\Delta\delta_C = 0.9$ ppm. The output was compared to independent, manually obtained assignments (see Methods) and the results are summarized in Table 2. Figure 5 further documents the result for the backbone and the β-nuclei. Both the ^{15}N and the HN resonances were correctly assigned for all residues, usually in all ten runs. Individual deviations from this optimal result occurred for residues 38, 39, 74 and 86, where correct assignments were obtained in 7-9 of the ten runs. The fragment 36-40 is a particular difficult case as it begins and ends with a proline. Residue 38 has a rather unusual shift for the amide proton of $\omega_{HN} = 11.3$ ppm and very weak peaks in the HNHA spectrum. Moreover, shift degeneracy was observed in the dipeptide

Table 2. Number of assignments by AUTOPSY, PICS and GARANT

Type of nuclei	Total[a]	Correct assignments[b]	Incorrect assignments[b]	Missing assignments
^{15}N	123	123 (100)	0 (0)	0 (0)
HN	123	123 (100)	0 (0)	0 (0)
$^{13}C\alpha$	128	125 (98)	2 (1.6)	1 (0.8)
Hα	139	134 (96)	3 (2.2)	2 (1.4)
$^{13}C\beta$	117	110 (94)	3 (2.6)	4 (3.4)
H-methyl	57	48 (84)	4 (7.0)	5 (8.8)
H-ring	21	9 (43)	5 (23.8)	7 (33.3)
other CH, CH_2[c]	288	185 (64)	16 (5.6)	87 (30.2)
		222 (77)		50 (17.4)
NH_2	24	14 (58)	4 (16.7)	6 (25.0)

[a]The total number of assignments refers to all assignments found in the independently obtained reference list of chemical shifts, which lacks a few side chain assignments (see Methods).
[b]The number given are absolute number of assignments by GARANT and in parentheses the percentage relative to the total number from the second column. Missing assignments occur when GARANT does not provide a unique assignment.
[c]For 'other CH, CH_2', the first line in the columns 'Correct assignments' refers to assignments with correct residue number and correct position within the side chain, the second line includes also assignments with correct residue number but incorrect position in the side chain (see text). In the column 'Incorrect assignments' only assignments with incorrect residue numbers are considered.

38-39 for both the α-carbons and the β-carbons. For the carbons and protons at the α-positions, the final results were 125 of 128, respectively 134 of 139, nuclei with correct assignments, including all 11 glycines. Two, respectively three, assignments were incorrect, and the Cα and Hα of residue 39 and Hα of residue 83 were not assigned at all by GARANT. The problematic sequence regions for the α-nuclei coincide with the ones for the N-H moieties. In addition, an uncertainty of the Cα shifts for the tripeptide 118-120 is caused by overlap of the Cα and Cβ frequencies of residue 118 as well as of the Cα frequencies of residues 119 and 120. The difficulty for the residues preceding proline 40 is further accentuated by the unusual chemical shift of the α-proton of histidine 35 of 6.53 ppm. The assignment of Cβ resonances yielded better results for the region 36-40 than that for the α-nuclei. Instead, a few problems of Cβ resonance assignments appeared around position 80. Very similar results were obtained for the assignment of a first β-proton in any residue; note that diasterotopic protons are not discriminated as no stereospecific assignment is attempted here. GARANT sometimes confused the assignment of β-protons with other protons of the same side chain, which partly explains the result shown in Figure 5 for the second proton in β-methylene groups. This is not unexpected, since the available data stems mostly from spectra of type TOCSY and NOESY (see also below).

For a discussion of side chain assignments we return to Table 2. The 57 methyl groups were with a success rate of 84% and only four incorrect assignments, all concerning leucines, well characterized by the automated procedure. Regarding ring protons it should be noted that no spectrum specific for aromatic carbon shifts was used, which explains the large number of both missing and incorrect assignments. Table 2 provides two numbers for correct assignments for the large group of remaining carbon-bound protons ('other CH, CH_2' in Table 2). Fully correct assignments were obtained for 185 protons, while in 37 cases the procedure confused different protons from the same side chain. As stated above, this was caused by the use of only TOCSY- and NOESY-type spectra. Together, the assignment to correct side chains amounts to 77%. Of the 16 incorrect assignments, ten originated from lysines. For this 'other CH, CH_2' group, GARANT provided no assignments for 50 protons, mostly from long side chains. On the other hand, GARANT assigned six additional protons for which no chemical shifts were given in the reference. The result for side chain NH_2 groups strongly depended on residue type. From the 14 protons from asparagines, only one assignment was incorrect and one was miss-

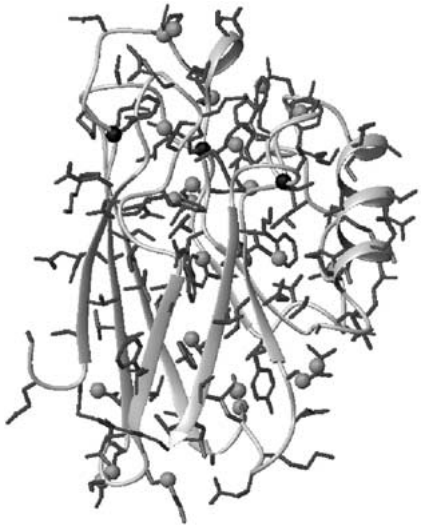

Figure 6. Structure summarizing the automated assignments by AUTOPSY, PICS and GARANT. The protein backbone is shown as a ribbon, heavy side chain atoms are indicated by black bonds. Spheres centered on heavy atoms indicate the positions of incorrect assignments for backbone (black) and side chain (gray) protons.

ing, while from the 10 protons from glutamines, five assignments were missing and three were incorrect.

Figure 6 illustrates the extent and the distribution of incorrect assignments in the azurin structure. Five residues with long side chains contain two erroneous assignments, but otherwise no clustering of errors in the 3D structure is observed. The overall error rate is 3.6% with respect to all assignments listed in the manually obtained reference, and 1.3% if one considers only the backbone and Cβ nuclei. It may be expected that further processing of this data with reasonably robust tools, e.g., structure determination programs (Güntert, 1998), would not be misled by the remaining errors but rather be able to help detect these. Also, a careful approach could initially ignore difficult assignments such as assignments for lysine side chains, aromatic protons and glutamine NH_2 groups. With the present data, this elimination would reduce the number of incorrect assignments from 35 to 19.

Conclusions

We have presented a fully automated approach for obtaining sequence-specific resonance assignments for proteins by combining the earlier presented programs AUTOPSY (Koradi et al., 1998) and GARANT (Bartels et al., 1996, 1997) with a new program, PICS, that optimizes the AUTOPSY output for use within GARANT. The relative calibration of peak lists by PICS using all peaks proved to be essential, as it allowed reducing the accuracy of peak position between different spectra to the same value as within each spectrum. By performing a complete assignment without interactive intervention we could demonstrate the quality of the individual steps performed by AUTOPSY and GARANT.

The approach is highly flexible allowing various combinations of spectra. This makes it useful for a variety of high-throughput NMR studies: for high-resolution structure determinations in structural genomics, or for the purposes of screening interactions and dynamics in functional genomics. In our example, spectra were selected that allow a complete assignment including side chains (although no specific spectra for the characterization of aromatic side chains were available). The output is expected to be sufficient for a 3D-structure determination by any of the recently presented procedures that combine NOESY assignment with structure calculation (Güntert, 1998; Linge et al., 2001; Gronwald et al., 2002; Herrmann et al., 2002).

Future improvements may include the usage in GARANT of the quality factors that AUTOPSY provides for every peak. Similarly, the interpretation of the GARANT output could be optimized. In this work the interpretation of the output was restricted to consideration of the first resonance proposed by GARANT for every nucleus. This program provides, however, more information. A more sophisticated interpretation would also consider the alternative suggestions for resonance assignments as well as the evaluation of assignment probabilities provided by GARANT. In a larger context, it would certainly be of advantage to include feedback loops that for example would allow GARANT to look back at spectral data. For more specific tasks such as structure determinations one could attempt to integrate also NOESY assignment and structure calculation in a complete approach with feedback loops and self-control. This would add one of the significant advantages of manual (interactive) resonance assignments into the automated processing.

Acknowledgements

B.G. Karlsson generously made the NMR sample with doubly labeled azurin available. Independent resonance assignments were obtained manually by

I. Bezsonova. Use of the Swedish NMR Centre facilities is gratefully acknowledged. This research is supported by the Swedish Research Council (VR grant 621-2001-3014).

References

Atreja, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comput. Chem.*, **18**, 139–149.

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

Billeter, M. (1991) In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*, Vol. 255: NATO ASI Series, Hoch J.C., Poulsen F.M. and Redfield C. (Eds.) Plenum, New York, pp. 279–290.

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

Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.P. and Kalbitzer, H.R. (2002) *J. Biomol. NMR*, **23**, 271–287.

Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sönnichsen, F.D. and Sykes, B.D. (1998) *J. Biomol. NMR*, **12**, 395–405.

Güntert, P. (1998) *Quart. Rev. Biophys.*, **31**, 145–237.

Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hajduk, P.J., Meadows, R.P. and Fesik, S.W. (1999) *Quart. Rev. Biophys.*, **32**, 211–240.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002) *J. Mol. Biol.*, **319**, 209–227.

Härd, T. (1999) *Quart. Rev. Biophys.*, **32**, 57–98.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.*, **135**, 288–297.

Korzhnev, D.M., Billeter, M., Arseniev, A.S. and Orekhov, V.Y. (2001) *Prog. Nucl. Magn. Reson. Spectrosc.*, **38**, 197–266.

Leckner, J. (2001) *Folding and Structure of Azurin – The Influence of a Metal*, Chalmers University of Technology, Göteborg, Sweden.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Li, K.B. and Sanctuary, B.C. (1997a) *J. Chem. Inf. Comput. Sci.*, **37**, 359–366.

Li, K.B. and Sanctuary, B.C. (1997b) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.

Linge, J.P., O'Donoghue, S.I. and Nilges, M. (2001) *Meth. Enzymol.*, **339**, 71–90.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Moseley, H.N., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.

Tian, F., Valafar, H. and Prestegard, J.H. (2001) *J. Amer. Chem. Soc.*, **123**, 11791–11796.

Vuister, G.W. and Bax, A (1993) *J. Amer. Chem. Soc.*, **115**, 7772–7777.

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

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.