J Biomol NMR (2008) 42:11–21
DOI 10.1007/s10858-008-9262-2

ARTICLE

Graphical interpretation of Boolean operators for protein NMR assignments

Dries Verdegem · Klaas Dijkstra · Xavier Hanoulle · Guy Lippens

Received: 31 March 2008 / Accepted: 9 June 2008 / Published online: 2 September 2008

Abstract We have developed a graphics based algorithm for semi-automated protein NMR assignments. Using the basic sequential triple resonance assignment strategy, the method is inspired by the Boolean operators as it applies "AND"-, "OR"- and "NOT"-like operations on planes pulled out of the classical three-dimensional spectra to obtain its functionality. The method's strength lies in the continuous graphical presentation of the spectra, allowing both a semi-automatic peaklist construction and sequential assignment. We demonstrate here its general use for the case of a folded protein with a well-dispersed spectrum, but equally for a natively unfolded protein where spectral resolution is minimal.

Keywords Computer-aided sequential assignment · Graphical semi-automatic protein assignment method · Boolean operators in NMR · Assignment of structured proteins · Assignment of unfolded proteins

Introduction

The first step in protein structure determination by NMR consists in the sequence specific assignment of the backbone and side chain resonances. A large number of programs have been developed over the last years to assist or automate this assignment process (Andrec and Levy 2002; Atreya et al. 2000, 2002; Bailey-Kellogg et al. 2000, 2005; Bartels et al. 1996, 1997; Bernstein et al. 1993; Buchler et al. 1997; Choy et al. 1997; Coggins and Zhou 2003; Croft et al. 1997; Eads and Kuntz 1989; Eccles et al. 1991; Eghbalnia et al. 2005; Friedrichs et al. 1994; Goddard and Kneller 1989; Görler et al. 1999; Grishaev and Llinás 2004; Gronwald et al. 1998, 2002; Güntert et al. 2000; Hare and Prestegard 1994; Helgstrand et al. 2000; Herrmann et al. 2002a, b; Hitchens et al. 2003; Hyberts and Wagner 2003; Johnson and Blevins 1994; Jung and Zweckstetter 2004; Kjaer et al. 1994; Kleywegt et al. 1991; Kobayashi et al. 2007; Kraulis 1989, 1994; Langmead and Donald 2004; Langmead et al. 2004; Leutner et al. 1998; Li and Sanctuary 1996, 1997a, b; Lin et al. 2003, 2006, 2005; Lukin et al. 1997; Malliavin et al. 1998; Malmodin et al. 2003; Masse and Keller 2005; Masse et al. 2006; Meadows et al. 1994; Morelle et al. 1995; Morris et al. 2004; Moseley and Montelione 1999; Moseley et al. 2001; Mumenthaler and Braun 1995; Mumenthaler et al. 1997; Neidig et al. 1995; Oezguen et al. 2002; Olson and Markley 1994; Orekhov et al. 2001; Oschkinat et al. 1991; Oschkinat and Croft 1994; Ou et al. 2001; Pons and Delsuc 1999; Pristovšek et al. 2002; Slupsky et al. 2003; Szyperski et al. 1998, 2002; Tian et al. 2001; van de Ven 1990; Vitek et al. 2005, 2006; Wan et al. 2003; Wan and Lin 2006; Wang et al. 2005; Wehrens et al. 1991, 1993a, b; Wu et al. 2006; Xu and Sanctuary 1993; Xu et al. 1994, 2002, 2006; Zimmerman et al. 1994, 1997; Zimmerman and Montelione 1995). One of the most common assignment strategies, on which indeed most of the mentioned methods are based, consists of a peak list construction and the subsequent matching of the C_α, C_β and CO chemical shifts between successive residues (Ikura et al. 1990; Kay et al. 1990; Montelione and Wagner 1990). Although successful for small to medium sized proteins, many programs using

D. Verdegem · X. Hanoulle · G. Lippens (✉)
Unité de Glycobiologie Structurale et Fonctionelle, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France
e-mail: guy.lippens@univ-lille1.fr

K. Dijkstra
Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747AG Groningen, The Netherlands

this strategy run into trouble when (i) overlap of the amide resonances increases due to the size or the unstructured nature of the protein, or (ii) spectral incompleteness due to intermediate line broadening or other phenomena. Both operations of peak list construction and frequency matching will suffer under those conditions, leading the operator back to the physical spectra, where one will manually try to complete the data.

We present here an assignment strategy that simultaneously allows both peak list construction and frequency matching in a semi-automatic manner, while remaining close to the initial spectra. It is based on a graphical interpretation of the Boolean AND operator, i.e. a point-by-point multiplication of 2D spectra. The main advantage is that the operator can walk graphically through the protein sequence, while maintaining a quality evaluation of the experimental data that lead to a decision on a sequential assignment. Although point-by-point operations (addition, subtraction, multiplication or division) are very commonly performed on FID's, they can also be done on frequency domain spectra and have even been introduced yet in the field of NMR spectra assignments (Masse et al. 2006). However, in our method, these operations play a more prominent role. Demonstrating the principles first on the well-folded Cyclophilin B protein, we extend its application towards a fragment of the natively unfolded Tau protein (Tau F3, amino acids 208–324), where extreme spectral overlap leads to strong degeneracies in the resonance frequencies.

Theory and methods

The assignment principle

Starting from the root $^{1}H,^{15}N$ HSQC spectrum and clicking on a certain appearing peak, our program readily extracts the corresponding $^{1}H,^{13}C$ planes from the CBCA(CO)NH and HNCO spectra (Fig. 1). On the basis of these two $^{1}H,^{13}C$ spectra, the operator defines with the mouse the carbon frequencies corresponding to the (i − 1) residue. These are automatically stored in a peak list (without

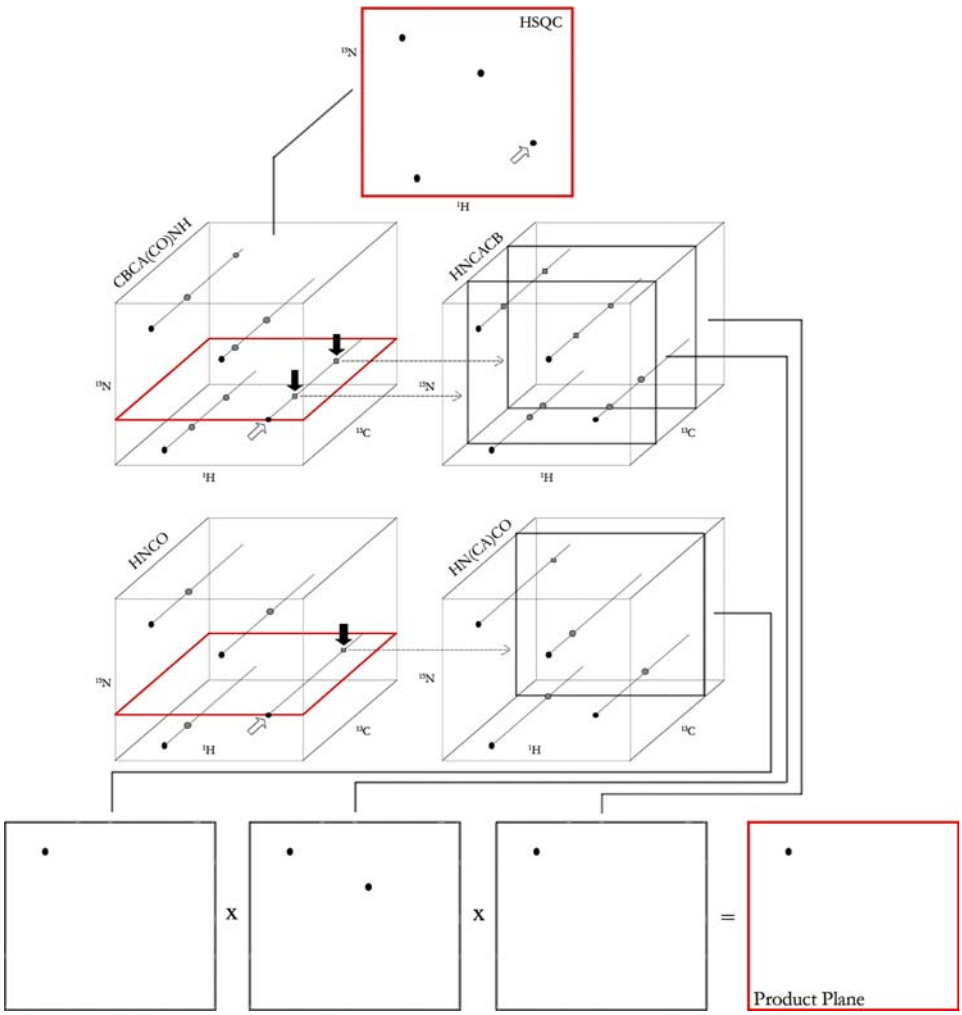

Fig. 1 The product plane approach applied to a protein subset of four consecutive residues. The planes presented on screen during execution in order to be able to click the necessary peaks are drawn in red. Clicking on the rightmost amide peak in a first step (hollow arrow) and the (i − 1) ^{13}C signals in a second step (black arrows) results in the selection of three planes whose point-by-point multiplication leaves only one major peak indicating that the leftmost amide peak is the (i − 1) residue. All 3D-spectra are joined with the HSQC spectra in front to indicate the root of each spin system. For simplicity, the smaller (i − 1) peaks occurring in the HNCACB and HN(CA)CO spectra have been left out of this scheme. It should be noted however, that these can lead to a small product plane signal in the residue (i) position

assignment at this moment), and the corresponding $^{1}H,^{15}N$ planes are extracted from the HNCACB and HN(CA)CO spectra. Rather than displaying the three corresponding planes together on screen and determine by eye the coordinates where there is simultaneous intensity, we impose this criterion by a point-by-point multiplication of the planes. This corresponds to a graphical interpretation of the Boolean AND operator, that requires simultaneous intensity in the spectra to obtain a resulting spectrum with a detectable intensity (Fig. 2). Applied to the three $^{1}H,^{15}N$ spectra extracted at the carbon frequencies of the (i − 1) residue, the point-by-point multiplication therefore defines a novel $^{1}H,^{15}N$ HSQC spectrum that contains intensity at the position of the (i − 1) residue.

Once the (i − 1) residue position has been found, it can be used as the starting point of another run of the algorithm. The repeated execution of the routine allows for an assigning walk through the spectrum towards the N-terminus of the protein.

In order to obtain product planes with constant highest peak intensities, the final product plane is initially normalized by dividing it by its maximum value (or minimum value if an odd number of negative peaks was involved in the multiplication) and is afterwards multiplied by a constant factor (e.g. 1e10) to finish with a spectrum with "natural" intensities (i.e. with peaks of more or less the same magnitude as the ones in real spectra).

Boolean operators in NMR

The assignment method is based on a graphical interpretation of the Boolean AND operator, that can be implemented as the point-by-point multiplication of spectral matrices (Fig. 2). Likewise, the OR operator would correspond with point-by-point summation. The NOT operation applied to a spectrum does not result in a new spectrum as such, but rather a 0/1 filled matrix of the same size as the original spectrum. Whether a certain element of this matrix is zero or one is determined by a chosen threshold (see light grey plane in the 2D-case of Fig. 2). If the intensity at a certain point in the original spectrum exceeds this threshold, the corresponding value in the NOT-matrix is set to zero. In the other case, the NOT-matrix value is set to one. This results in a "spectrum" that display holes at the places where the original spectrum contained peaks. Both OR and NOT operations on spectral planes will prove useful further on when trying to assign proteins with unfavorable amino acid sequences.

Input spectra

In its most basic form, the described algorithm uses the HNCACB, CBCA(CO)NH, HN(CA)CO, HNCO and of course HSQC spectra as input. When these five spectra are applied as depicted in Fig. 1, an assigning "walk" towards the N-terminus of the protein is made. It is however interesting to note that a simple exchange of sequential and intra-residue spectra in the algorithm results in the opposite functionality that allows a "walk" in the opposite direction, towards the C-terminus.

Here, the assignment of Cyclophilin B and Tau F3 spectra will be discussed.

The NMR measurements of both protein samples were performed on a Bruker Avance 600 MHz equipped with a cryogenic triple resonance probe head by using standard Bruker pulse programs. The CypB (185aa, 20.4 kDa) sample contained 600μM CypB in an aqueous buffer with 50 mM Na_2HPO_4/NaH_2PO_4, 60 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 6.35 at 293 K. The Tau F3 (124aa, 13.3 kDa) sample contained 250 μM of protein in a 25 mM Tris-D11, 25 mM NaCl, 2.5 mM EDTA, 2.5 mM DTT aqueous buffer (pH 6.8, 293 K). The acquisition

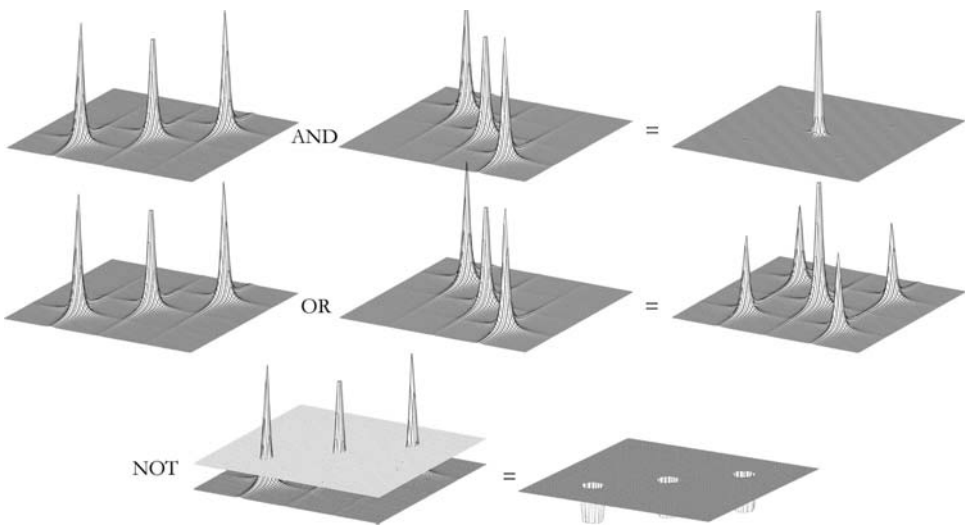

Fig. 2 The Boolean operators applied to 2D spectra presented as topographic maps

parameters for the CypB spectra were: 2048 (^{1}H) and 256 (^{15}N) complex points and 32 scans per increment for the HSQC (exp time: 2 h 41 min), 1024 (^{1}H), 68 (^{15}N) and 128 (^{13}C) complex points and 16 scans per increment for the HNCACB (exp time: 1 day 21 h 23 min), 1024 (^{1}H), 104 (^{15}N) and 142 (^{13}C) complex points and 8 scans per increment for the CBCA(CO)NH (exp time: 1 day 15 h 20 min) and 1024 (^{1}H), 104 (^{15}N) and 128 (^{13}C) complex points and 8 scans per increment for the HN(CA)CO and HNCO (exp times: 1 day 10 h 58 min and 1 day 10 h 25 min). For the Tau F3 sample, the acquisition parameters were: 2048 (^{1}H) and 256 (^{15}N) complex points and 64 scans per increment for the HSQC (exp time: 5 h 25 min), 2048 (^{1}H), 96 (^{15}N) and 232 (^{13}C) complex points and 8 scans per increment for the HNCACB (exp time: 2 days 13 h 36 min), 2048 (^{1}H), 96 (^{15}N) and 132 (^{13}C) complex points and 8 scans per increment for the CBCA(CO)NH (exp time: 1 day 11 h 42 min), 2048 (^{1}H), 92 (^{15}N) and 96 (^{13}C) complex points and 8 scans per increment for the HN(CA)CO and HNCO (exp times: 1 d 10 h 58 min and 1 d 10 h 25 min) and 2048 (^{1}H), 86 (^{15}N) and 96 (^{13}C) complex points and 8 scans per increment for the HNN (exp time: 23 h 26 min).

It is important to notice that the different spectra required for an assignment of this kind should all be recorded under the same sample conditions. Any technique based on the point-by-point multiplication of spectrum slices originating from different spectra is obviously quite sensitive to small differences in chemical shift across those spectra, which might arise if nonidentical parameters are used.

Results and discussion

Graphical walk through the triple resonance spectra

To demonstrate the procedure on a real-life example, we start from the cross peak at 7.55, 121.72 ppm in the CypB HSQC spectrum, that we previously assigned to Lys 149 (Hanoulle et al. 2007). Extracting the ^{1}H,^{15}N planes from the HN(CA)CO, HNCACB (C_α and C_β) at the (i − 1) ^{13}CO, $^{13}C_\alpha$ and $^{13}C_\beta$ carbon frequencies as defined by the HNCO and HN(CO)CACB lines at the Lys 149 position yields the three planes shown in Fig. 3. The product plane (4) (in green) comes into being as a result of their point-by-point multiplication. This plane superposed on the Cyclophilin B HSQC indisputably points out the position of the root signal of Arg 148.

When lowering the threshold, we do see other amide resonances of lower intensity, indicating that due to the limited resolution in the carbon dimension residues can have some residual intensity that matches the three required frequencies. When a given (^{1}H,^{15}N) correlation peak represents two or more residues, the operator is faced with the same problem as the number based algorithms. However, as in other semi-automated assignment programs, our method, inherent to its principle, constantly shows the relevant spectrum slices on screen. The obvious advantage is that the raw data with all the information about subtle frequency differences and/or peak forms are still available. Two real situations where only working with raw data helps to exclude ambiguity in the assignment of the CypB protein are considered here.

Figure 4 shows the plane pulled out from the cyclophilin B CBCA(CO)NH spectrum after clicking the Val 12 residue signal. This plane can then, according to the product plane (AND) algorithm, be used to select the C_α and C_β (i − 1) signals. The Val 12 (^{1}H,^{15}N) correlation peak appears in a more crowded region of the HSQC. Four ^{13}C peaks can be distinguished in the ^{1}H,^{13}C plane, but visual inspection readily allows to pair the peaks at 61.5 and 69.0 ppm. The two other signals at 41.0 and 54.0 ppm have a proton frequency that differs by 0.006 ppm from the previous pair, and would therefore probably be assigned to the same peak by automated assignment routines that commonly apply a proton uncertainty of 0.05 ppm.

A second example illustrating the advantage of having ready access to the raw data is found when trying to assign the amide peak that corresponds to Gly 31. The superposition of the HSQC and the product plane obtained after clicking Leu 32 is shown in Fig. 5. We are faced here with the extreme, but possible situation in which the authentic (i − 1) signal is not the most intense one in the product plane. To establish and overcome this problem however, a simple feedback strategy, that exploits once again the usefulness of being able to graphically present the slight chemical shift differences, is sufficient.

This feedback functionality graphically compares the set of peaks involved, as in Fig. 6 for the Leu 32 case, and reveals clearly that the Gly 138 C_α chemical shift is shifted slightly downfield compared to the Leu 32 C_α (i − 1) shift.

At any point, the spectroscopist can decide not to include one of the three plane subject to the multiplication (bottom Fig. 1). If for example a certain residue has a weak C_β peak in the HNCACB spectrum, one can exclude the corresponding C_β-plane from the (i − 1) product plane calculation (and thus treat the residue as if it were a glycine). Although this practice will in theory lead to a less selective product plane, it can in some cases avoid the situation where the product plane exhibits a too low signal/noise to be useful.

Using our graphical walk, that is in this case only interrupted when one encounters a proline residue as these do not appear in a HSQC spectrum, we were able to repeat the full assignment of CypB in a minimal time (less than a

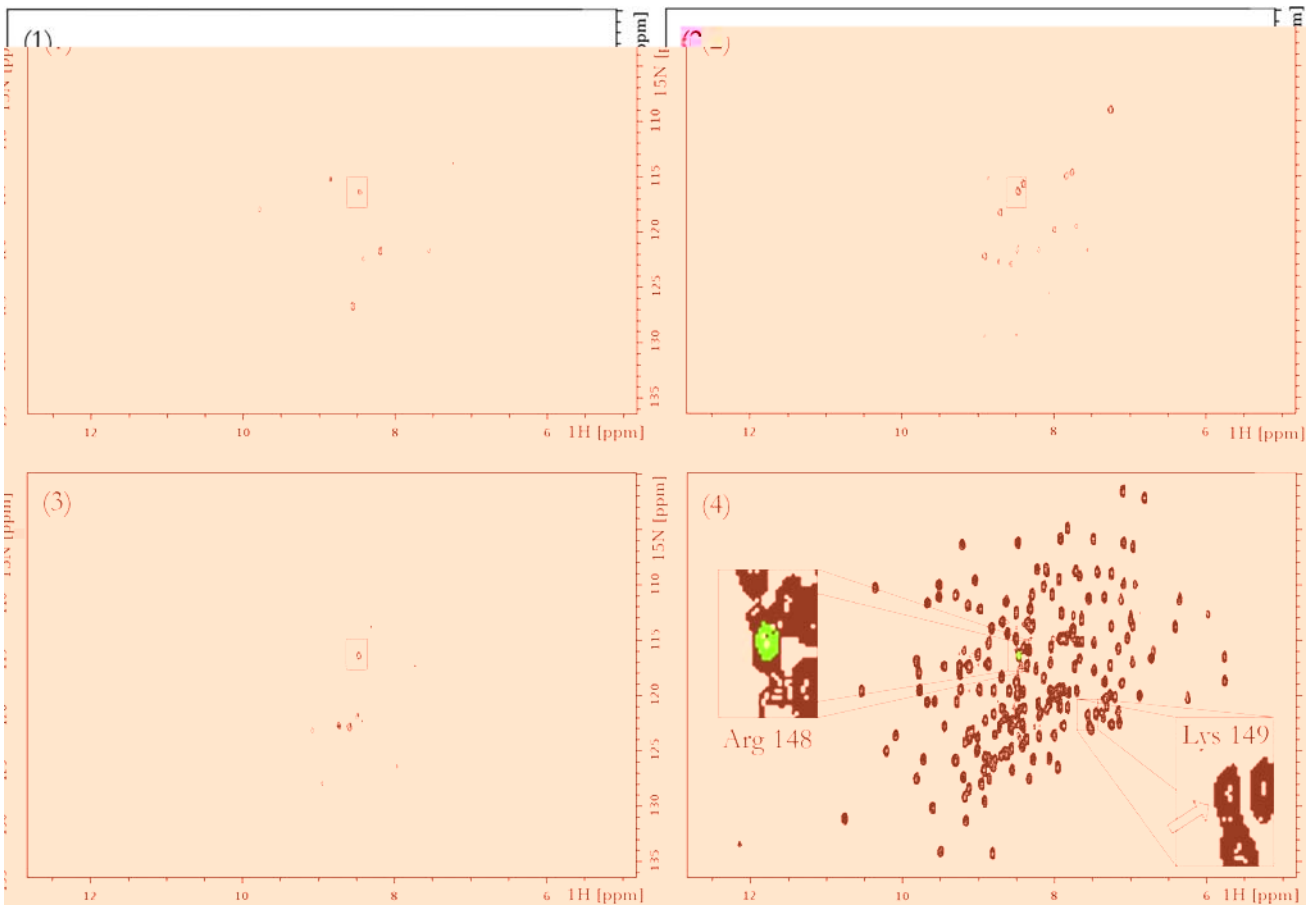

Fig. 3 When clicking the Lys 149 ^{1}H, ^{15}N signal in the CypB HSQC, automatic extraction of the corresponding ^{1}H,^{13}C planes of the HNCO and HN(CO)CACB allow the manual definition of the (i − 1) CO, C_{α} and C_{β} frequencies. Extracting the ^{1}H,^{15}N planes from the HN(CA)CO, HNCACB (C_{α} and C_{β}) at these (i − 1) ^{13}CO, $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ carbon frequencies yields the three planes (**1**), (**2**) and (**3**). The resulting product plane (in green) superposed on the original HSQC-spectrum is presented in (**4**). In it, the Arg 148 position can be clearly identified. All clicking required to obtain the first three planes can be done fairly precisely because of an available zoom function. This also enables one to backup precise backbone ^{15}N, ^{13}C, ^{1}H, and side-chain $^{13}C_{\beta}$ assignments to an output file

day), and comparison with our previous assignment based on peak lists showed perfect agreement.

Extending to natively unstructured proteins

Natively unfolded proteins represent a different challenge to assignment programs. With a reduced amide proton chemical shift range (often inferior to 1 ppm), overlap becomes very severe, leading to many branching points for the automatic matching algorithms. Therefore, manual intervention of the operator becomes even more important than in the case of folded proteins such as CypB, as it allows to alleviate possible ambiguities on the basis of subtle peak position or shape differences.

In this unstructured protein category, the algorithm was powerful enough to determine almost all the proline bordered amino acid stretches of the Tau F3 (amino acids 208–324) protein fragment. We were able to assign all residues except for the S237–S238 pair and the two GGG triplets starting at G271 and G302. The pair and triplets occur in heavily overlapped regions and are moreover preceded by a proline residue preventing the upstream graphical walk.

The performance of the product plane algorithm can however, for natively unfolded proteins, be improved when combined with the information included within the triple resonance HNN spectrum, that can be recorded with a decent sensitivity for these protein due to their narrow line widths. When a certain HSQC root is chosen with the mouse, a corresponding ^{1}H,^{15}N plane can also be pulled out of this latter 3D spectrum, in which one can find the ^{15}N chemical shift of the residues in (i − 1) and (i + 1) position. Drawing this information as horizontal lines on the product plane spectrum will, in case of doubt, reveal the correct neighbor of the clicked root signal.

Figure 7 shows the (i − 1) product plane and the (i − 1) ^{15}N and (i + 1) ^{15}N chemical shift values after executing the algorithm on Val 306. Besides the own (i) signal, the product plane contains three possible (i − 1) signals. The

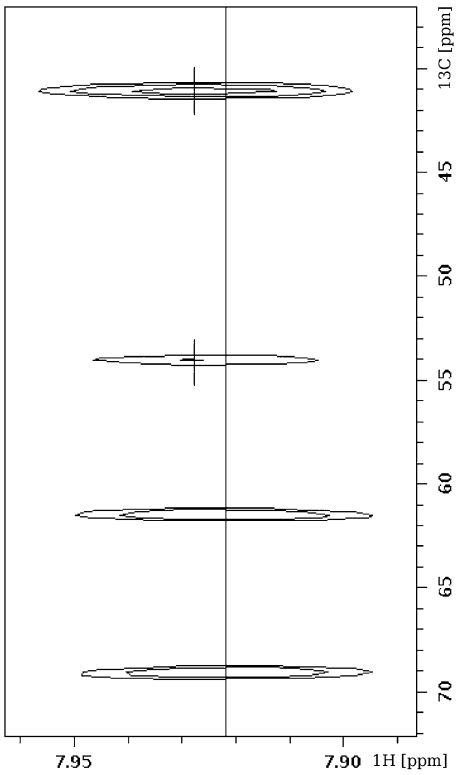

Fig. 4 A fragment of the Val 12 corresponding $^{1}H,^{13}C$ plane extracted from the cyclophilin B CBCA(CO)NH spectrum. The black vertical line indicates the proton chemical shift of the Val 12 residue. This view is projected on the screen as determined by the algorithm

feedback strategy of graphically comparing the C_α and C_β shifts involved shows that there is a perfect match with none of those three. The HNN info on the other hand,

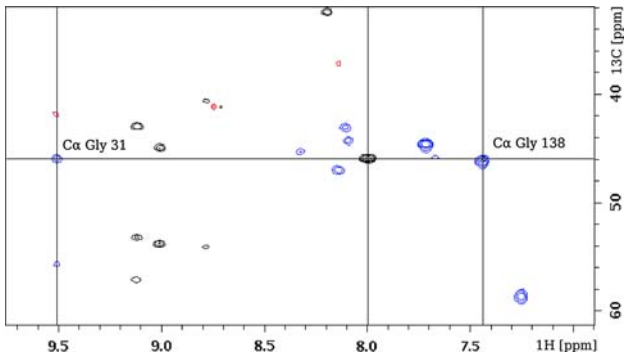

Fig. 6 A superposition of three pulled-out 2D spectra in order to graphically compare the involved signals. The two HNCACB extracted planes are determined by the ^{15}N chemical shifts of Gly 31 and Gly 138 of CypB. The CBCA(CO)NH plane corresponds to the Leu 32 ^{15}N chemical shift. As color convention we use black for the CBCA(CO)NH signals and blue (C_α) and red (C_β) for HNCACB signals. Vertical lines intercept the three comparison partners in their points of highest intensity. The horizontal line crosses the Leu 32 signal at its maximum

allows one to pinpoint the largest product plane signal as the genuine Ser 305 amide resonance. This situation arose because of the almost complete overlap of the Ser 262, Ser 293 and Ser 305 HSQC signals that all have a glycine residue in the (i − 1) position. This caused the individual HNCACB signals to have merged to three new averaged glycine C_α, serine C_α and serine C_β signals at different chemical shifts.

When the HNN was added to list of input spectra we succeeded in completely assigning the Tau F3 spectra, including the eight earlier mentioned difficult cases. For

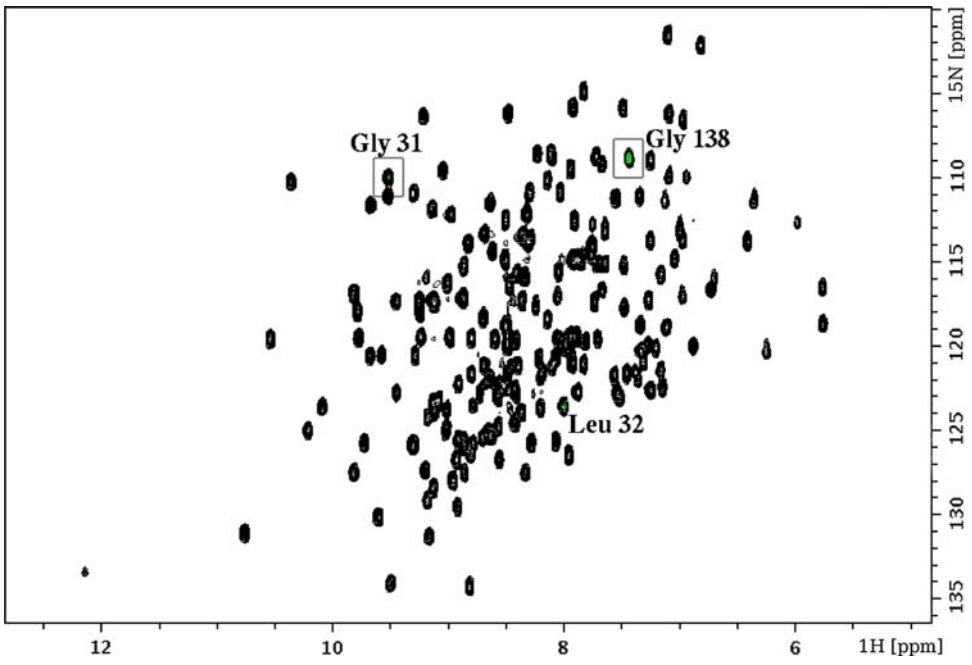

Fig. 5 The product plane after clicking Leu 32 shown on top of the Cyclophilin B HSQC. The strongest signal is actually that of Gly 138, while Gly 31 has a lower intensity at the given threshold. Also the Leu 32 peak itself shows some intensity due to the presence of minor (i − 1) signals in the HNCACB and HN(CA)CO spectra

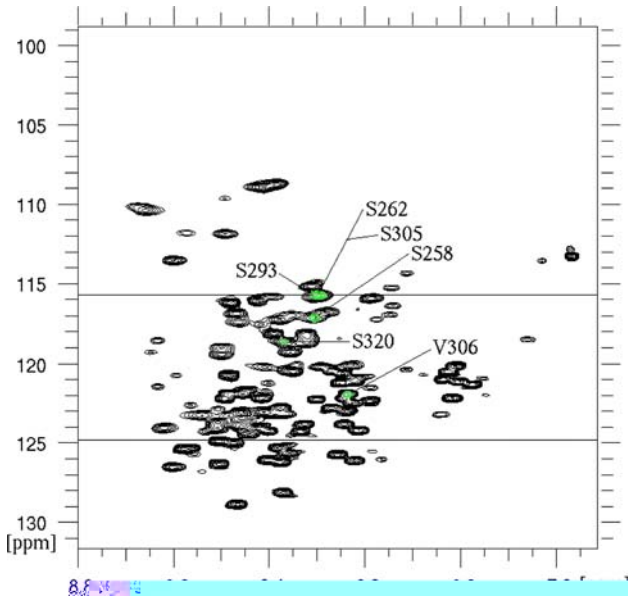

Fig. 7 For natively unstructured proteins such as Tau F3, the product plane functionality has to be reinforced by information from the HNN spectrum to be able to do complete assignments. The (i − 1) and (i + 1) ^{15}N chemical shifts it provides when following the procedure after clicking V306 (drawn on the product plane as horizontal lines), allows to assign the S305 residue

unstructured proteins, the HNN information also makes the assignment a lot faster, since it efficiently prevents the need for a signal position based feedback in the assigning walk.

Generating starting points

The above described procedure leads to the ready assignment of stretches of connected resonances. Based on the residue-type specific carbon chemical shifts, that unambiguously define residues such as Gly, Ala, Ser and Thr, those stretches can in most cases be mapped in a straightforward way onto the protein sequence. For the case of a folded protein such as CypB in our example, this information is ample, and a full assignment can be easily obtained. For the Tau fragment, however, the rapid obtention of suitable starting points helps in the procedure, and avoids problems with repeating stretches in the protein sequence. Generally, the existence of suitable starting points in the assignment procedure gives additional confidence in the method, and leads to a more rapid assignment.

The graphical interpretation of the Boolean operators as defined above can equally be used in a similar way by including the OR operator to allow for some spectral degeneracy. A first manner is to define a given residue type by the requirement that both the C_α and C_β frequencies fall within a certain range of the random coil chemical shift values for this residue type. This requirement can be obtained graphically in two steps (Fig. 8): first, the HNCACB ^{1}H,^{15}N planes with the ^{13}C chemical shift values within the defined range of the random coil values are summed, leading to a C_α- and C_β-defining plane. Formally, this sum procedure is equivalent to the Boolean OR operator. In a second stage, we multiply both resulting sum-planes to obtain a novel ^{1}H,^{15}N plane that contains intensity only for those resonances where the C_α and C_β requirement is fulfilled. This procedure is akin to the MUSIC pulse sequences, where one combines carbon selective pulses and multiple quantum filtering to obtain residue-type specific subspectra (Schubert et al. 1999, 2001a, b). However, the present method does not require novel experiments, as it is a post-processing method based only on the existing HNCACB experiment, and thereby does not suffer from the relaxation losses that inevitably accompany the longer pulse lengths required for selectivity. This is a distinct advantage for larger proteins, but also for unfolded proteins where the selectivity of the post-music procedure can easily be fine-tuned on the basis of the same experiment, without requiring the recording of novel experiments.

Following this procedure starting from the HNCACB spectrum, one obtains (i, i + 1) subspectra, as the given residue (i) will also be seen from the (i + 1) amide resonance because of the (weaker) N(i)-C_α(i − 1) coupling constant. The same principle can equally be applied to the CBCA(CO)NH spectrum, and thereby leads to a subspectrum of only those residues that have the required residue type as their downstream neighbor (i + 1). Applying the third Boolean operator described in Fig. 2, both the HNCACB and CBCA(CO)NH can thus be used to generate

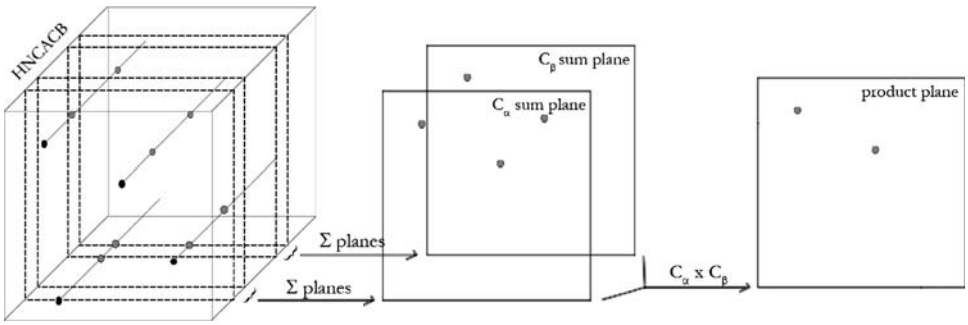

Fig. 8 Boolean operators applied to obtain type-selective ^{1}H,^{15}N spectra. Windows are defined (indicated by the dashed lines) around a residue type characteristic C_α and C_β chemical shift value ($\overline{C_\alpha} \pm x_\alpha$ and $\overline{C_\beta} \pm x_\beta$) and the total of planes enclosed are summed. Subsequently, the resulting sum planes are multiplied to yield the type specific HSQC

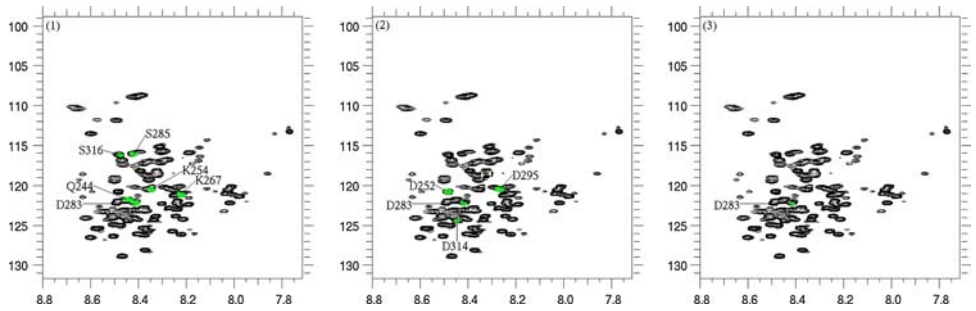

Fig. 9 (**1**) Represents the Leu (i + 1) HSQC, generated from Tau F3's CBCA(CO)NH. For it, windows of 55.1 ± 0.5 ppm (C_α) and 42.4 ± 0.5 ppm (C_β) were chosen. All Leu (i + 1) signals are present in the spectrum. In order to generate the pure Asp (i) HSQC (**2**), the C_α and C_β windows were 54.2 ± 0.5 and 41.1 ± 0.5 ppm, respectively. The cutoff threshold for the NOT operation (Fig. 2) was put at 0.001% of the most intense CBCA(CO)NH derived product plane peak. The Asp type-specific subspectrum contains all four Asp residues in the Tau F3 sequence. Finally, a Boolean AND operation between (**1**) and (**2**) results in a spectrum that only contains intensity at the position of the Asp residue in the unique (L)D283 dipeptide (**3**). The spectrum manipulations that lead to (**3**) are done in a few seconds and thus this procedure provides a very fast generation of starting points. Again, as was the case in the assignment method, all planes are normalized before multiplication and multiplied by a constant factor (1e10) after, as to maintain constant intensity. The scales are values in ppm

pure (i) type specific subspectra. Indeed, (i,i + 1) AND NOT (i + 1) equals (i). We typically do a dipeptide scan over the protein sequence to determine those dipeptides that are unique in the sequence. For the first amino acid of such a dipeptide, the (i + 1) residue-type selective HSQC is calculated, while for the second amino acid, we generate the (i) selective subspectrum. The product plane of those two HSQC's will contain only one major peak, indicating the position of the second residue of the dipeptide.

In a concrete example as the Tau fragment, a simple scan found that 56 residues are in a unique pattern, and this despite the fact that the overall amino acid sequence of Tau is largely unfavorable, with five amino acids making up for over 55% of the sequence. The (L)D283 dipeptide is unique, and the Leu (i + 1) selective spectrum combined with the Asp (i) specific spectrum readily defines it as the peak at 8.41, 122.20 ppm (Fig. 9).

A total of 42 out of 56 residues in a unique pattern, where we note that the pattern XY can be differentiated from XYP because of the proline-directed effect (prolines in (i + 1) induces a −2 ppm chemical shift for the C_α (Wishart et al. 1995)), were immediately assignable. Some residues were unable to be found because of one of three reasons: (i) carbon signals not included in the defined windows, (ii) weak corresponding HNCACB and/or CBCA(CO)NH signals or (iii) both residues of the unique pair are of the same type, which causes the signal to disappear in the (i) type selective subspectrum. All three effects lead to empty product planes at the usual contour threshold and a large amount of meaningless noise peaks at lower thresholds. Reason (ii) is related to the fact that a summation of a number of N subspectra by the OR operator will lead to a decrease in signal/noise of about sqrt(N) (depending on the peak widths) as many planes will contribute to the noise and only a few to the actual signal. However, we found that for unfolded proteins such as Tau, where the defined windows can be kept reasonably small (e.g. 1 ppm) because of the smaller C_α and C_β chemical shift spreads, this signal/noise reduction is disturbing in only a minor number of cases. Thus, the rapid determination of pivotal points, whereby we can even allow for some ambiguous assignments, greatly enhances not only the initial stages of the assignment procedure. As it provides for suitable anchoring points, it facilitates to connect the sequential stretches to the protein sequence. The complete assignment of Tau F3, using our complete package of assignment tools, was done in 1 day time.

Discussion

We have shown here a graphical implementation of the traditional assignment procedure based on connecting complementary triple resonance experiments. The main advantage of the procedure is that the operator remains very close to the experimental spectra at every moment, without relying on peak lists. Whereas the latter allow a rapid computer-assisted assignment in favorable cases, spectral overlap or differential quality of the data in different zones of the spectra can introduce errors that inevitably will lead to problems requiring manual intervention. The product planes as defined in this work represent the Boolean AND operator in its most simple fashion: point-by-point multiplication guarantees that the only remaining intensity comes from planes that both had intensity at the given resonance position. We showed that even for crowded spectra such as obtained for the natively unfolded Tau protein, this graphical procedure can greatly facilitate the assignment process. When complemented in a straightforward way with the HNN experiment, that is particularly favorable for such samples because of their

sharp lines, the assignment becomes as trivial as for a folded protein. An extension to the Boolean OR operator allows to define amino-acid specific subspectra based on the original HNCACB and CBCA(CO)NH spectrum. When compared to the experimental MUSIC pulse sequences, this procedure does not suffer from additional relaxation losses due to the selective and hence longer carbon pulses, does not require novel experiments and can readily build in differential ^{13}C selectivity, but can evidently not reproduce the multiple quantum filtering as was done in the some MUSIC sequences. These residue selective subspectra constitute the input of a very straightforward starting point generation method. We showed that this latter procedure is particularly suitable for unfolded proteins, where the random coil ^{13}C chemical shifts by definition provide an excellent center point for the chemical shift range to be considered. We are currently exploring how the procedure can be combined with quantum-mechanical or semi-empirical chemical shift calculations in order to provide a rapid assignment of the HSQC spectra of proteins with a known 3D structure.

All methods described in this paper were developed using python scripts with the NMR python library functionality (http://linuxnmr02.chem.rug.nl/~dijkstra/NMRpy/). However, we have also implemented them in the CcpNmr software suite (Vranken et al. 2005) as an extension to Analysis for wide distribution. They will become available in the next release (Analysis2.0)

Acknowledgments We thank Dr. I. Landrieu for sample preparation, Dr. J.-M. Wieruszeski for collecting the NMR spectra and Dr. T. Stevens and W. Boucher of the University of Cambridge, Department of Biochemistry for implementing our protein NMR assignment tools in the CcpNmr software suite. The 600 MHz facility used in this study was funded by the Région Nord—Pas de Calais (France), the CNRS and the Institut Pasteur de Lille. Part of this work was funded by a grant of the Agence National de la Recherche (ANR 05 BLAN 0320-0; Tau:Tubulin). D.V. received a predoctoral grant of the French Ministère de la Recherche.

References

Andrec M, Levy RM (2002) Protein sequential resonance assignments by combinatorial enumeration using $^{13}C_\alpha$ chemical shifts and their (i,i − 1) sequential connectivities. J Biomol NMR 23: 263–270

Atreya H, Chary K, Govil G (2000) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. Curr Sci 83:1372–1376

Atreya H, Sahu S, Chary K, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17:125–136

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparce unassigned NMR data. J Comput Biol 7:537–558

Bailey-Kellogg C, Chainraj S, Pandurangan G (2005) A random graph approach to NMR sequential assignment. J Comput Biol 12:569–583

Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7:207–213

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18: 139–149

Bernstein R, Cieslar C, Ross A, Oschkinat H, Freund J, Holak TA (1993) Computer-assisted assignment of multidimensional NMR spectra of proteins: application to 3D NOESY-HMQC and TOCSY-HMQC spectra. J Biomol NMR 3:245–251

Buchler NE, Zuiderweg ER, Wang H, Goldstein RA (1997) Protein heteronuclear NMR assignments using mean-field simulated annealing. J Magn Reson 125:34–42

Choy W, BC S, Zhu G (1997) Using neural network predicted secondary structure information in automatic protein NMR assignment. J Chem Inf Comput Sci 37:1086–1094

Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 26:93–111

Croft D, Kemmink J, Neidig KP, Oschkinat H (1997) Tools for the automated assignment of high-resolution three-dimensional protein NMR spectra based on pattern recognition techniques. J Biomol NMR 10:207–219

Eads C, Kuntz I (1989) Programs for computer-assisted sequential assignment of proteins. J Magn Reson 82:467–482

Eccles C, Güntert P, Billeter M, Wüthrich K (1991) Efficient analysis of protein 2D NMR spectra using the software package EASY. J Biomol NMR 1:111–130

Eghbalnia HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). J Biomol NMR 32:219–233

Friedrichs M, Mueller L, Wittekind M (1994) An automated procedure for the assignment of protein 1HN, 15N, 13C alpha, 1H alpha, 13C beta and 1H beta resonances. J Biomol NMR 4:703–726

Goddard T, Kneller D (1989) Sparky 3. University of California, San Francisco

Görler A, Gronwald W, Neidig KP, Kalbitzer HR (1999) Computer assisted assignment of ^{13}C and ^{15}N edited 3D-NOESY-HSQC spectra using back calculated and experimental spectra. J Magn Reson 137:39–45

Grishaev A, Llinás M (2004) BACUS: a Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. J Biomol NMR 28:1–10

Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR 12:395–405

Gronwald W, Moussa S, Elsner R, Jung A, Ganslmeier B, Trenner J, Kremer W, Neidig KP, Kalbitzer HR (2002) Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). J Biomol NMR 23:271–287

Güntert P, Salzmann M, Braun D, Wüthrich K (2000) Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. J Biomol NMR 18: 129–137

Hanoulle X, Melchior A, Sibille N, Parent B, Denys A, Wieruszeski JM, Horvath D, Allain F, Lippens G, Landrieu I (2007) Structural and functional characterisation of the interaction between cyclophilin B and a heparin derived oligosaccharide. J Biol Chem 282:34148–134158

Hare BJ, Prestegard JH (1994) Application of neural networks to automated assignment of NMR spectra in proteins. J Biomol NMR 4:35–46

Helgstrand M, Kraulis P, Allard P, Härd T (2000) Ansig for Windows: an interactive computer program for semiautomatic assignment of protein NMR spectra. J Biomol NMR 18:329–336

Herrmann T, Güntert P, Wüthrich K (2002a) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Biomol NMR 319:209–227

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR 24: 171–189

Hitchens T, Lukin JA, Zhan YP, McCallum SA, Rule GS (2003) MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR 25:1–9

Hyberts SG, Wagner G (2003) IBIS—A tool for automated sequential assignment of proteins spectra from triple resonance experiments. J Biomol NMR 26:335–344

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667

Johnson BA, Blevins RA (1994) NMR view: a computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614

Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Kjaer M, Andersen K, Poulsen F (1994) Automated and semiautomated analysis of homo- and heteronuclear multidimensional nuclear magnetic resonance spectra of proteins: the program PRONTO. Methods Enzymol 239:288–308

Kleywegt GJ, Boelens R, Cox M, Llinás M, Kaptein R (1991) Computer-assisted assignment of 2D ^{1}H NMR spectra of proteins: basic algorithms and application to phoratoxin B. J Biomol NMR 1:23–47

Kobayashi N, Iwahara J, Koshiba S, Tomizawa T, Tochio N, Güntert P, Kigawa T, Yokoyama S (2007) KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR studies. J Biomol NMR 39:31–52

Kraulis P (1989) ANSIG: a computer program for the assignment of 1H NMR spectra by interactive computer graphics. J Magn Reson 84:627–633

Kraulis P (1994) Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space ab initio approach. J Mol Biol 243:696–718

Langmead CJ, Donald BR (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR assignments. J Biomol NMR 29:111–138

Langmead CJ, Yan A, Lilien R, Wang L, Donald BR (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. J Comput Biol 11:277–298

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11:31–43

Li KB, Sanctuary B (1996) Automated extracting of amino acid spin systems in proteins using 3D HCCH-COSY/TOCSY spectroscopy and constrained partioning algorithm. J Chem Inf Comput Sci 36:585–593

Li KB, Sanctuary B (1997a) Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. J Chem Inf Comput Sci 37:359–366

Li KB, Sanctuary B (1997b) Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. J Chem Inf Comput Sci 37: 467–477

Lin G, Xiang W, Tegos T, Li Y (2006) Statistical evaluation of NMR backbone resonance assignment. Int J Bioinform Res Appl 2:147–160

Lin G, Xu D, Chen ZZ, Jiang T, Wen J, Xu Y (2003) Computational assignment of protein backbone NMR peaks by efficient bounding and filtering. J Bioinform Comput Biol 1:387–409

Lin HN, Wu KP, Chang JM, Hsu WL (2005) GANA—a genetic algorithm for NMR backbone resonance assignment. Nucleic Acids Res 33:4593–4601

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (13C, 15N)-labeled proteins). J Biomol NMR 9:151–166

Malliavin T, Pons J, Delsuc M (1998) An NMR assignment module implemented in the Gifa NMR processing program. Bioinformatics 14:624–631

Malmodin D, Papavoine CH, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J Biomol NMR 27:69–79

Masse JE, Keller R (2005) AutoLink: automated sequential reso nance assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. J Magn Reson 174:133–151

Masse JE, Keller R, Pervushin K (2006) SideLink: automated side-chain assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. J Magn Reson 181:45–67

Meadows RP, Olejniczak ET, Fesik SW (1994) A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. J Biomol NMR 4:79–96

Montelione GT, Wagner G (1990) Conformation-independent sequential NMR connections in isotope-enriched polypeptides by ^{1}H-^{13}C-^{15}N triple resonance experiments. J Magn Reson 87:183–188

Morelle N, Brutscher B, Simorre JP, Marion D (1995) Computer assignment of the backbone resonances of labelled proteins using two-dimensional correlation experiments. J Biomol NMR 5:154–160

Morris LC, Valafar H, Prestegard JH (2004) Assignment of protein backbone resonances using connectivity, torsion angles and $^{13}C^{\alpha}$ chemical shifts. J Biomol NMR 29:1–9

Moseley H, Montelione G (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Moseley H, Monleon D, Montelione G (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol 339:91–108

Mumenthaler C, Braun W (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. J Mol Biol 254:465–480

Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY and three-dimensional protein structure determination. J Biomol NMR 10:351–362

Neidig KP, Geyer M, Görler A, Antz C, Saffrich R, Beneicke W, Kalbitzer HR (1995) AURELIA, a program for computer-aided

analysis of multidimensional NMR spectra. J Biomol NMR 6:255–270

Oezguen N, Adamian L, Xu Y, Rajarathnam K, Braun W (2002) Automated assignment and 3D structure calculations using combinations of 2D homonuclear and 3D heteronuclear NOESY spectra. J Biomol NMR 22:249–263

Olson JB, Markley JL (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances: a demonstration of the connectivity tracing assignment tools (CONTRAST) software package. J Biomol NMR 4: 385–410

Orekhov VY, Ibraghimov V, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Oschkinat H, Croft D (1994) Automated assignment of multidimensional nuclear magnetic resonance spectra. Methods Enzymol 239:308–318

Oschkinat H, Holak T, Cieslar C (1991) Assignment of protein NMR spectra in the light of homonuclear 3D spectroscopy: an automatable procedure based on 3D TOCSY-TOCSY and 3D TOCSY-NOESY. Biopolymers 31:699–712

Ou HD, Lai HC, Serber Z, Dötsch V (2001) Efficient indentification of amino acid types for fast protein backbone assignments. J Biomol NMR 21:269–273

Pons J, Delsuc M (1999) RESCUE: an artificial neural network tool for the NMR spectral assignment of proteins. J Biomol NMR 15:15–26

Pristovšek P, Rüterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure–the program *st2nmr*. J Comput Chem 23:335–340

Schubert M, Smalla M, Schmieder P, Oschkinat H (1999) MUSIC in triple-resonance experiments: amino acid type-selective $^{1}H,^{15}N$ correlations. J Magn Reson 141:34–43

Schubert M, Oschkinat H, Schmieder P (2001a) MUSIC and aromatic residues: amino acid type-selective $^{1}H,^{15}N$ correlations, III. J Magn Reson 153:186–192

Schubert M, Oschkinat H, Schmieder P (2001b) MUSIC, selective pulses, and tuned delays: amino acid type-selective $^{1}H,^{15}N$ correlations, II. J Magn Reson 148:61–72

Slupsky CM, Boyko RF, Booth VK, Sykes BD (2003) Smartnotebook: a semi-automated approach to protein sequential NMR resonance assignments. J Biomol NMR 27:313–321

Szyperski T, Banecki B, Braun D, Glaser RW (1998) Sequential resonance assignment of medium-sized $^{15}N/^{13}C$-labeled proteins with projected 4D triple resonance NMR experiments. J Biomol NMR 11:387–405

Szyperski T, Yeh DC, Sukumaran DK, Moseley HN, Montelione GT (2002) Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 99:8009–8014

Tian F, Valafar H, Prestegard J (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. J Am Chem Soc 123: 11791–11796

van de Ven FJ (1990) PROSPECT, a program for automated interpretation of 2D NMR spectra of proteins. J Magn Reson 86:633–644

Vitek O, Bailey-Kellogg C, Craig B, Kuliniewicz P, Vitek J (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. Bioinformatics 21:230–236

Vitek O, Bailey-Kellogg C, Craig B, Vitek J (2006) Interential backbone assignment for sparse data. J Biomol NMR 35: 187–208

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinás M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59:687–696

Wan X, Lin G (2006) A graph-based automated NMR backbone resonance sequential assignment. In: Computational systems bioinformatics 2006 conference proceedings, vol 4, pp 55–66

Wan X, Xu D, Slupsky CM, Lin G (2003) Automated protein NMR resonance assignments. In: Proceedings of the IEEE computer society conference on bioinformatics, vol 2, pp 197–208

Wang J, Wang T, Zuiderweg ER, Crippen GM (2005) CASA: an efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. J Biomol NMR 33:261–279

Wehrens R, Buydens L, Kateman G (1991) Validation and refinement of expert systems—interpretation of NMR-spectra as an application in analytical-chemistry. Chemometr Intell Lab Syst 12:57–67

Wehrens R, Lucasius C, Buydens L, Kateman G (1993a) HIPS, a hybrid self-adapting expert system for nuclear magnetic resonance spectrum interpretation using genetic algorithms. Anal Chim Acta 277:313–324

Wehrens R, Lucasius C, Buydens L, Kateman G (1993b) Sequential assignment of 2D-NMR spectra of proteins using genetic algorithms. J Chem Inf Comput Sci 33:245–251

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigation of nearest-neighbour effects. J Biomol NMR 5:67–81

Wu KP, Chang JM, Chen JB, Chang CF, Wu WJ, Huang TH, Sung TY, Hsu WL (2006) RIBRA—an error-tolerant algorithm for the NMR backbone assignment problem. J Comput Biol 13:229–244

Xu J, Sanctuary B (1993) CPA: constrained partitioning algorithm for initial assignment of protein ^{1}H resonances from MQF-COSY. J Chem Inf Comput Sci 33:490–500

Xu J, Strauss S, Sanctuary B, Trimble L (1994) Use of fuzzy mathematics for complete automated assignment of peptide 1H 2D NMR spectra. J Magn Reson B103:53–58

Xu Y, Xu D, Kim D, Olman V, Razumovskaya J, Jiang T (2002) Automated assignment of backbone NMR peaks using constrained bipartite matching. Comput Sci Eng 4:50–62

Xu Y, Wang X, Yang J, Vaynberg J, Qin J (2006) PASA—a program for automated protein NMR backbone signal assignment by pattern-filtering approach. J Biomol NMR 34:41–56

Zimmerman DE, Montelione GT (1995) Automated analysis of nuclear magnetic resonance assignments for proteins. Curr Opin Struct Biol 5:664–673

Zimmerman D, Kulikowski C, Wang L, Lyons B, Montelione GT (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. J Biomol NMR 4:241–256

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien Cy, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610