J Biomol NMR (2009) 44:43–57
DOI 10.1007/s10858-009-9314-2

ARTICLE

Highly automated protein backbone resonance assignment within a few hours: the «BATCH» strategy and software package

Ewen Lescop · Bernhard Brutscher

Received: 15 December 2008 / Accepted: 20 March 2009 / Published online: 15 April 2009

Abstract Sequential resonance assignment represents an essential step towards the investigation of protein structure, dynamics, and interaction surfaces. Although the experimental sensitivity has significantly increased in recent years, with the availability of high field magnets and cryogenically cooled probes, resonance assignment, even of small globular proteins, still generally requires several days of data collection and analysis using standard protocols. Here we introduce the BATCH strategy for fast and highly automated backbone resonance assignment of ^{13}C, ^{15}N-labelled proteins. BATCH makes use of the fast data acquisition and analysis tools BEST, ASCOM, COBRA, and HADAMAC, recently developed in our laboratory. An improved Hadamard encoding scheme, presented here, further increases the performance of the HADAMAC experiment. A new software platform, interfaced to the NMRView software package, has been developed that enables highly automated NMR data processing and analysis, sequential resonance assignment, and ^{13}C chemical shift extraction. We demonstrate for four small globular proteins that sequential resonance assignment can be routinely obtained within a few hours, or less, in a highly automated and robust way.

Keywords Protein · Fast NMR · Resonance assignment · Chemical shift · Amino-acid type discrimination · Algorithm

Introduction

Protein resonance assignment is an essential prerequisite for NMR investigation of molecular structure and dynamics. It is also the basis for powerful chemical shift mapping approaches routinely used for identifying binding interfaces in stable or transient molecular assemblies, or for fragment-based ligand-screening approaches by NMR. Backbone resonance assignment is usually achieved through the collection of a set of multidimensional (at least 3D) triple-resonance H–N–C experiments that correlate neighboring nuclear spins within the same residue or from consecutive residues. The frequency lists obtained from these spectra containing the coordinates (chemical shifts) of the individual correlation peaks are then analyzed either manually or by appropriate computer tools in order to (1) build fragments of sequentially connected spin systems, and (2) identify possible amino-acid types for each ^{1}H–^{15}N frequency pair (residue). Finally, the sequential assignment problem consists in finding the best match between the protein amino acid sequence and these two types of experimental constraints.

Generally, a complete set of NMR spectra is recorded before data analysis is started. The recording of a set of 3D spectra using standard methods requires at least a few days of data acquisition time. Often even longer experimental times are chosen to ensure that enough information is

Electronic supplementary material The online version of this article (doi:10.1007/s10858-009-9314-2) contains supplementary material, which is available to authorized users.

E. Lescop (✉)
Laboratoire de Chimie et Biologie Structurales, Institut de Chimie des Substances Naturelles, CNRS UPR 2301,
1, Avenue de la Terrasse, 91190 Gif-sur-Yvette, France
e-mail: ewen.lescop@icsn.cnrs-gif.fr

B. Brutscher (✉)
Laboratoire de RMN, Institut de Biologie Structurale—Jean-Pierre Ebel, UMR5075 CNRS-CEA-UJF, 41, Rue Jules Horowitz, 38027 Grenoble Cedex, France
e-mail: bernhard.brutscher@ibs.fr

available from these data for complete, or as complete as possible resonance assignment. This standard approach imposes a time stability of the NMR sample of several days to weeks, excluding numerous biologically interesting protein samples from NMR investigation because of their short life times. Therefore, many efforts have been made in recent years to speed up this assignment step by developing alternative fast multidimensional NMR data acquisition schemes (Brutscher et al. 1994; Frydman et al. 2002, 2003; Kim and Szyperski 2003; Kupce and Freeman 2003a, b; Atreya and Szyperski 2004; Brutscher 2004; Hiller et al. 2005; Cornilescu et al. 2007; Gal et al. 2007) together with new processing approaches (Orekhov et al. 2001, 2003; Bruschweiler 2004; Rovnyak et al. 2004; Marion 2005; Jaravine et al. 2006; Kazimierczuk et al. 2006). Concomitantly, new computational tools that allow fast and automated data analysis without or with very little user intervention required (Friedrichs et al. 1994; Olson and Markley 1994; Bartels et al. 1997; Lukin et al. 1997; Zimmerman et al. 1997; Leutner et al. 1998; Lin et al. 2002, 2005, 2006; Monleon et al. 2002; Coggins and Zhou 2003; Jung and Zweckstetter 2004; Vitek et al. 2004, 2005, 2006; Masse and Keller 2005; Wan and Lin 2006, 2007; Wu et al. 2006; Volk et al. 2008) were introduced. These developments open the way for new iterative assignment protocols where NMR data are acquired in a piecewise manner, and analyzed while recording the next piece of data until sufficiently complete resonance assignment is obtained. Iterative approaches are expected to make optimal use of the intrinsic sensitivity of the experimental setup (protein sample and NMR spectrometer), and avoid unnecessarily long experimental times. Recently, such an iterative method for the fully automated backbone resonance assignment through the stepwise collection of NMR data was introduced by Jaravine and Orekhov (2006) and Jaravine et al. (2008), and its good performance was demonstrated for the protein ubiquitin, as well as for a 13 kDa unstructured polypeptide chain. This method, as well as most other commonly used approaches, are based on the retrieval of the sequential connectivity and amino-acid type information from the same set of triple resonance experiments. ^{13}C chemical shifts, however, are poorly discriminative in terms of amino-acid type determination, and only Ala, Gly and Ser/Thr residues can be unambiguously identified. As a consequence, a high degree of unambiguous connectivity information is required for successful automated backbone resonance assignment.

Here we present an alternative, integrated assignment strategy that allows iterative data acquisition with minimal experimental times of <1 h. The BATCH, for BEST/ASCOM/Targeted-sampling/COBRA/HADAMAC, assignment strategy is based on a number of tools recently developed in our laboratory aiming at shortening, and facilitating the different steps of resonance assignment for small globular proteins. In the BATCH strategy, amino-acid-type identification is obtained from a HADAMAC experiment (Lescop et al. 2008) that can be recorded in ~30 min, and provides significantly higher amino-acid-type discrimination than the usual statistical analysis of ^{13}C chemical shifts. The sequential connectivity information is retrieved from a small set of triple resonance experiments, recorded using longitudinal relaxation enhanced (Pervushin et al. 2002; Deschamps and Campbell 2006) BEST-type pulse sequences (Schanda et al. 2006; Lescop et al. 2007a). In addition, optimized ^{15}N spectral aliasing using ASCOM (Lescop et al. 2007b), and targeted ^{13}C time-domain sampling (Lescop and Brutscher 2007) is used to reduce the number of sampled data points in the indirect dimensions. Fragments of sequentially connected spin systems are extracted from the NMR data using the COBRA method (Lescop and Brutscher 2007), a computational tool that directly uses the NMR data as input, and does not require any high dimensional peak picking. A new software platform has been developed that allows automated processing of all recorded NMR data, followed by resonance assignment and chemical shift extraction. Applications to several small proteins demonstrate the experimental performance of the BATCH protocol under various conditions, and its ability to obtain resonance assignments within a few hours in a highly automated manner.

The BATCH strategy and software

NMR data acquisition

Three types of experiments are required for the BATCH strategy. (1) First, a BEST-type ^{1}H–^{15}N HSQC spectrum is recorded without spectral aliasing in the ^{15}N dimension. This spectrum is peak-picked and subjected to on-fly ^{15}N spectral width optimization using ASCOM (Lescop et al. 2007b). The ASCOM software is available as a stand-alone version, and as a script running on Varian spectrometers (http://www.icsn.cnrs-gif.fr/download/nmr). (2) Secondly, a HADAMAC-2 experiment is performed, yielding six ^{1}H–^{15}N spectra with correlation peaks from seven different amino-acid groups: Ala–Val–Ile (*AVI*), Gly (*Gly*), Ser (*Ser*), Thr (*Thr*), Asn–Asp (*Asx*), Cys–Phe–His–Tyr–Trp (*Cys-Arom*) and Arg, Glu, Lys, Pro, Gln, Met, Leu (*Rest*). Of note, residues from the *Cys-Arom* and *Rest* groups appear in the same subspectrum but with opposite signs (by convention positive and negative respectively). Compared to the originally proposed HADAMAC experiment (Lescop et al. 2008), HADAMAC-2 uses a slightly modified Hadamard encoding scheme, explained in more detail in the next section. (3) Finally, pairs of sequential and intra-residue H–N–C (H–N–CA, H–N–CB, and H–N–CO) correlation experiments are recorded. Intra-residue experiments are preferred with respect to their more common bi-directional counterparts because they yield only a single correlation peak per residue that is of advantage for COBRA analysis. For the same reason correlation experiments with C^{β} frequency labeling are tuned in a way that only correlations with C^{β} (but not C^{α}) are observed in the final spectra for non-glycine residues. For all 3D H–N–C experiments, the ^{15}N spectral width is set to the ASCOM-derived optimal value. For $^{13}C^{\alpha}$ frequency labeling, a targeted regular sampling scheme is used as described previously (Lescop and Brutscher 2007). Typically, 10 complex points are regularly collected from $t_1 = 0$ to 3 ms, and 10 additional points are collected from $t_1 = 25$ to 28 ms.

The HADAMAC-2 experiment

The HADAMAC experiment (Lescop et al. 2008), based on a HBCBCACONH coherence transfer pathway, has been setup for maximal amino-acid-type discrimination in a single experiment, and thus in minimal experimental time. For proper amino-acid-type editing, the HACACONH coherence pathway that may also be detected in this type of experiment, needs to be suppressed for all but glycine residues. In the original HADAMAC experiment this was realized by adjusting the $^{13}C^{\beta} \rightarrow {}^{13}C^{\alpha}$ transfer delay 2ζ to $1/2/J_{CC} = 14.2$ ms. In the modified HADAMAC-2 experiment, the undesired coherence transfer pathways are filtered out by an appropriate Hadamard encoding scheme, as proposed recently by Pantoja-Uceda and Santoro (2008). Although only four different spin manipulations are used for sign inversion, an eight-step Hadamard encoding can be realized that separates the six amino-acid bands, as well as signals from two different unwanted HACACONH coherence pathways (Tables 1, 2). The eight HADAMAC-2 sub-spectra recorded for ubiquitin are shown in Fig. 1. HADAMAC-2 allows shortening the transfer delay 2ζ (typically set to $2\zeta = 9$–10 ms) in order to limit relaxation-induced signal loss. Consequently, HADAMAC-2 is expected to yield increased sensitivity for larger, slower tumbling molecules. A HADAMAC-2 spectrum of the 16.7 kDa protein calmodulin (with an experimentally estimated molecular tumbling correlation time of $\tau_c = 8$ ns) is shown in Fig. S1 of the Supplementary material.

Table 1 Hadamard matrix for band encoding and decoding in HADAMAC-2 experiment

	Gly	*AVI*	*Ser*	*Thr*	*Asx*	*Cys-Arom/ REST*	CA pathway I (<55 ppm)[a]	CA pathway II (55–85 ppm)[a]
Exp 1	+	+	+	+	+	+	+	+
Exp 2	−	+	+	+	−	+	−	−
Exp 3	−	+	−	−	+	+	+	−
Exp 4	+	+	−	−	−	+	−	+
Exp 5	−	+	−	+	−	−	+	+
Exp 6	+	+	−	+	+	−	−	−
Exp 7	+	+	+	−	−	−	+	−
Exp 8	−	+	+	−	+	−	−	+

[a] The CA pathways I and II correspond to the $H\alpha_{i-1}$–$C\alpha_{i-1}$–CO_{i-1}–N_i–HN_i transfer. The Ser/Thr C_{β} inversion pulse also affects sign encoding for residues with C_{α} chemical shift values in the 55–85 ppm range leading to the creation of the "CA pathway" groups I and II

Table 2 Experimental realization of Hadamard matrix of Table 1

	Inversion of CB coupled to CO	*Ser/Thr* inversion	CH_2 (CB) inversion	CH_2 (CA) inversion
Exp 1	+	+	+	+
Exp 2	−	+	+	+
Exp 3	+	−	+	−
Exp 4	−	−	+	−
Exp 5	+	+	−	+
Exp 6	−	+	−	+
Exp 7	+	−	−	−
Exp 8	−	−	−	−

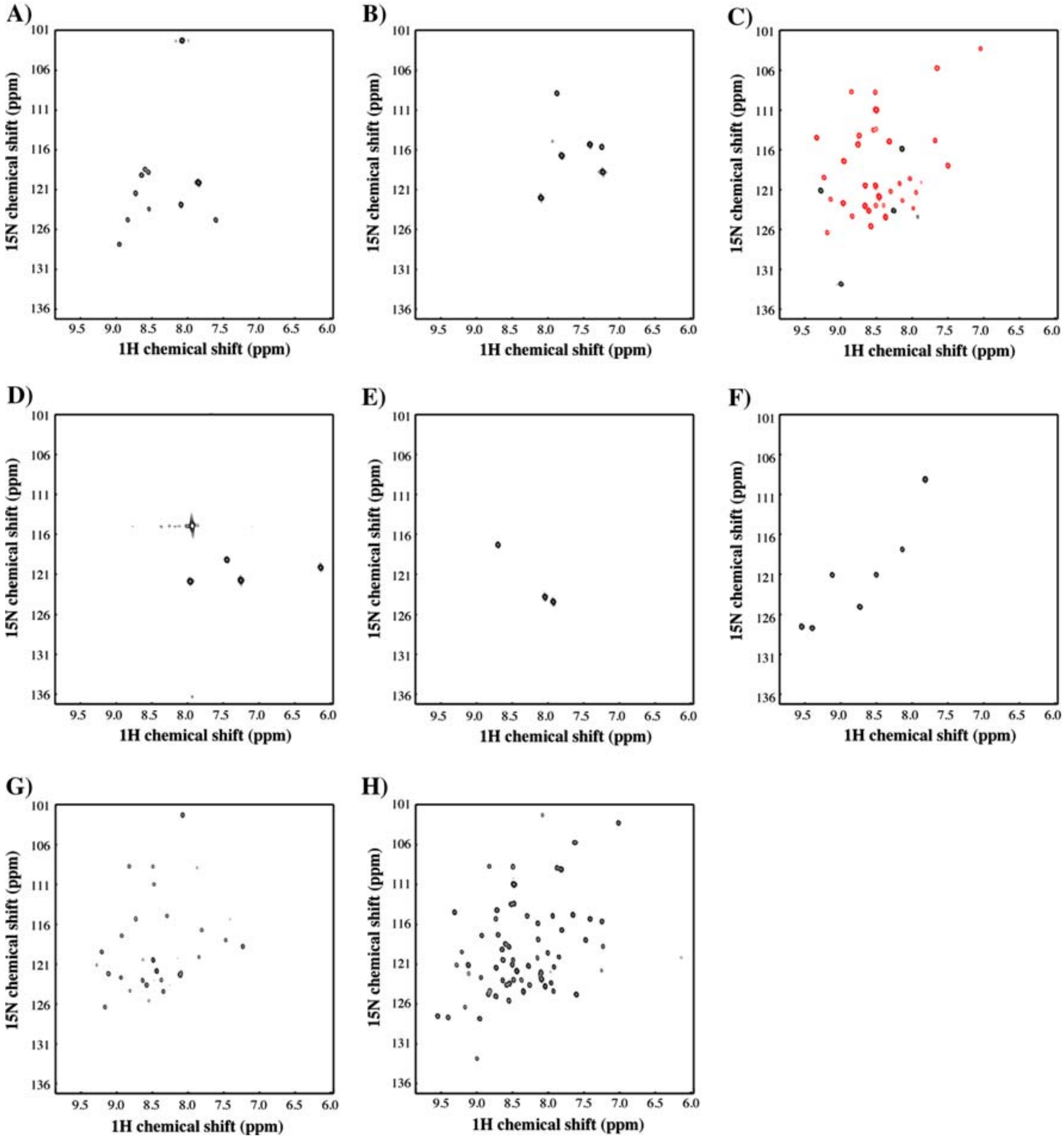
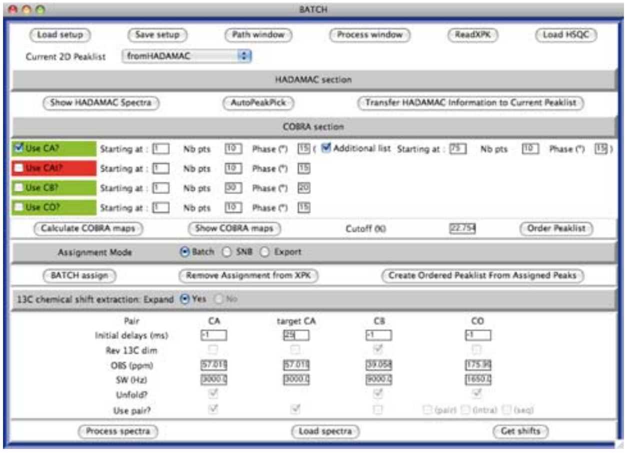


Fig. 1 HADAMAC-2 spectra of ubiquitin. The eight sub-spectra correspond to the amino-acid groups: **A** *AVI*, **B** *Asx*, **C** *Cys-Arom* (positive) and *Rest* (negative), **D** *Gly*, **E** *Ser*, **F** *Thr*. Panels **G** and **H** contain the subspectra corresponding to the HACACONH pathways that are differently affected by the Ser/Thr band selective inversion pulse. Positive and negative contours are colored in *black* and *red*, respectively. The HADAMAC-2 experiment was carried out with the minimal two-step phase cycling in 20 min on a 1 mM ubiquitin sample on a 600 MHz spectrometer equipped with a cryogenically cooled probe

Overview of the BATCH software

The BATCH software, written in Tcl language, has been interfaced to the NMRView software package (Blevins and Johnson 1994) to make use of all NMRView functionalities. BATCH provides an integrated software platform for all steps of NMR data processing and analysis required within the BATCH resonance assignment strategy. All spectral manipulations are carried out with the widely used NMR-Pipe processing software (Delaglio et al. 1995) and are controlled by scripts generated by BATCH. The main BATCH window is shown in Fig. 2. It consists of several buttons, each achieving a precise task, and is separated in five sections from top to bottom, corresponding to the different steps of the BATCH resonance assignment. The first section is used for setting general parameters and for NMR data processing. All current parameters can be saved in a separate file (“Save Setup”) and reloaded later (“Load Setup”). Paths to NMRPipe scripts, protein amino-acid sequence and NMR data are set within the “Path Window”



Current 2D Peaklist fromHADAMAC

☑ Use CA? Starting at : 1 Nb pts 10 Phase (°) 15 (☑ Additional list Starting at : 75 Nb pts 10 Phase (°) 15)
☐ Use CAi? Starting at : 1 Nb pts 10 Phase (°) 15
☐ Use CB? Starting at : 1 Nb pts 30 Phase (°) 20
☐ Use CO? Starting at : 1 Nb pts 10 Phase (°) 15

Cutoff (%) 22.754

Assignment Mode ● Batch ○ SNB ○ Export

13C chemical shift extraction: Expand ● Yes ○ No

Pair	CA	target CA	CB	CO
Initial delays (ms)	-1	25	-1	-1
Rev 13C dim	☐	☐	☑	☐
OBS (ppm)	57.019	57.019	39.054	175.9
SW (Hz)	3000.0	3000.0	9000.0	1650.0
Unfold?	☑		☑	☑
Use pair?	☑	☑	☐	☐ (pair) ☐ (intra) ☐ (seq)

Fig. 2 Graphical interface of the BATCH software. BATCH represents a new module in NMRView. The main BATCH window is composed of several sections, devoted to HADAMAC analysis, COBRA analysis, automated sequential resonance assignment, and ^{13}C chemical shift extraction. Buttons are used to execute a variety of functionalities to simplify the data analysis and calculations

(Fig. S2). The NMR data processing can be achieved from the "Process Window". The 2D ^{1}H–^{15}N peak list used for the current COBRA and HADAMAC analysis is also defined in this section. The second section aims at facilitating the analysis of the HADAMAC experiment. The third section contains the parameters for COBRA analysis of the different pairs of experiments. The fourth section allows performing sequential backbone resonance assignment, either from the Best-first principle (completely automated), or using Smartnotebook (Slupsky et al. 2003). Finally, the fifth section is used for ^{13}C chemical shift extraction from the recorded NMR data. The BATCH software can be freely downloaded (academic users only) from the website http://www.icsn.cnrs-gif.fr/download/nmr.

NMR data processing

Backbone resonance assignment requires the collection of several datasets, and usually involves the manipulation of many processing scripts. In order to make the processing step fast and user-friendly, the BATCH software provides an interface with NMRPipe. The "Process Window" contains three buttons for each type of experiment that allow setup of the conversion ("fid.com") and processing ("nmrproc.com") scripts, and visualization of the processed spectra in NMRView ("Load Spectra"). Since each experiment can always be processed in the same way for a given setup (spectrometer, pulse sequence), BATCH makes use of a library of default processing scripts for each type of experiment. The COBRA analysis only requires Fourier transformation of ^{1}H and ^{15}N dimensions of the triple resonance experiments. However, the ^{13}C dimension is also Fourier transformed for visualization in NMRView. The most time-consuming step during processing is the linear prediction (LP) in the ^{15}N dimension prior to Fourier transformation. Therefore, the processing parameters are first adjusted without LP, and the "Process All with LP" button then allows automated reprocessing of all available datasets using additional LP in the ^{15}N dimensions.

^{1}H–^{15}N peak list and HADAMAC analysis

One unique 2D ^{1}H–^{15}N peak list is required for COBRA and HADAMAC analysis. This list can be directly obtained from the peak picking of the ^{1}H–^{15}N HSQC spectrum. However, peak picking procedures are not very efficient in case of partial peak overlap. In contrast, a much higher number of ^{1}H–^{15}N peaks are well resolved in the HADAMAC experiment due to the spreading of the peaks along the amino-acid-type dimension. Therefore we developed a tool for the automated peak picking and subsequent filtering of the six individual HADAMAC subspectra, resulting in a single 2D ^{1}H–^{15}N peak list. The peak-picking algorithm is accessible by the "AutoPickPeak" button and the resulting peak list can be checked against the ^{1}H–^{15}N HSQC for the identification of possible aliased and missing peaks in the HADAMAC experiment. Figure 3 illustrates the efficiency of a HADAMAC-based peak picking procedure in case of partially overlapping correlation peaks in a ^{1}H–^{15}N HSQC spectrum.

The 2D peak list is then used for the automatic extraction of the amino-acid-type information contained in the HADAMAC spectra. Assigning a ^{1}H–^{15}N cross peak to a given amino acid group essentially consists in finding the subspectrum where the (absolute value) intensity is the highest at a given ^{1}H–^{15}N chemical shift position. However, in order to take into account possible ambiguities in amino-acid group identification due to partial overlap or for high dynamic signal range, ^{1}H–^{15}N peaks that are too close to another one are excluded from the automated analysis. In practice, the algorithm initially filters out all ^{1}H–^{15}N peaks whose boxes overlap, and that are left unassigned. Boxes can be manually adjusted by inspection of the spectra if necessary. Any mistake in assigning the amino-acid type may have disastrous effects on the final assignment. Therefore, the assignment is only accepted if the maximum intensity (absolute) value is at least twice as high as the second ranking intensity value. This criterion has the additional advantage of leaving ^{1}H–^{15}N peaks with weak or absent HADAMAC signals unassigned. Finally, the amino-acid type assignment is stored in the current 2D peaklist for later usage.

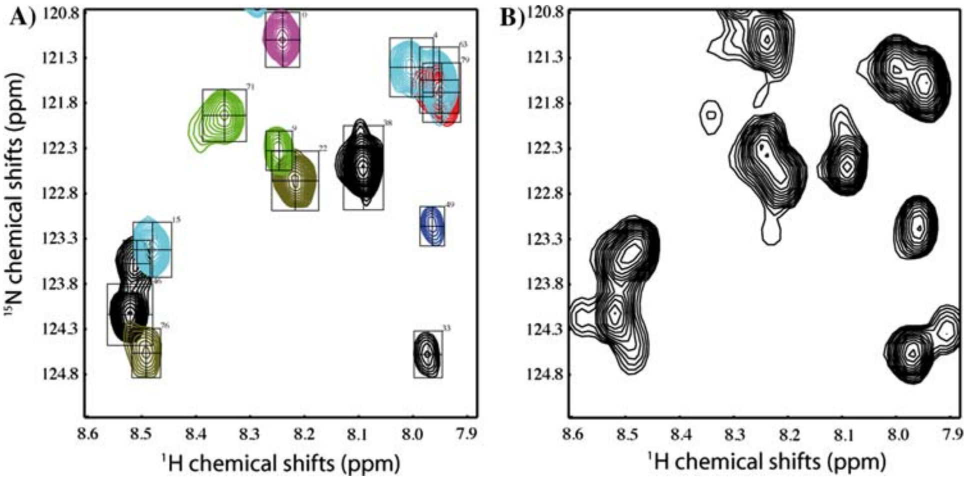


Fig. 3 HADAMAC utilities and 2D peak picking. **A** HADAMAC spectrum of Hyl1 with the six subspectra superposed and color-coded using the following default color code: *AVI* (*black*), *Gly* (*red*), *Thr* (*brown*), *Ser* (*blue*), *Cys-Arom* (*green*), *Asx* (*magenta*), and *Rest* (*cyan*). Boxes represent the result of the automated peak picking/filtering procedure of the HADAMAC subspectra. The visible region corresponds to the crowded central part of the spectrum. **B** The same region of the ^{1}H–^{15}N HSQC spectra of the Hyl1 protein, illustrating the efficiency of the HADAMAC peak analysis to distinguish partially overlapping peaks, and to create a high-quality 2D ^{1}H–^{15}N peak list for further BATCH analysis

COBRA analysis

COBRA (COrrelation-Based Reconstruction Algorithm) calculates a connectivity map from pairs of triple resonance experiments for the n ^{1}H–^{15}N frequency pairs (residues) present in the 2D peak list. The COBRA $i-1/i$ connectivity map is a $n \times n$ matrix where each element M_{ij} ($0 \leq M_{ij} \leq 100\%$) represents the "probability" of a sequential connectivity between residues with ^{1}H–^{15}N frequencies numbered i and j. One COBRA map is calculated for each pair of H–N–C experiments. When several pairs are available, the final COBRA map is computed as the element-by-element product of all individual maps. Ambiguity in sequential connectivity results from degenerate ^{13}C chemical shifts or incomplete frequency discrimination, and translates into non vanishing elements in the final COBRA map. For each pair of experiments, the elements M_{ij} are real numbers and are calculated as the phase weighted (real part of the) linear correlation coefficient of the ^{13}C complex time domain signals extracted from the sequential and intra-residue triple resonance experiments at the ^{1}H–^{15}N chemical shift positions of crosspeaks i and j, respectively (Lescop and Brutscher 2007). The frequency discrimination in COBRA is adjusted by the phase cutoff parameter ϕ_0 that is the unique parameter in the phase weighting factor and that is user-defined for each individual pair of experiments. Choosing a small ϕ_0 value yields higher frequency discrimination. In the case of a low signal to noise ratio (SNR), however, a small phase cutoff may also result in the loss of correct sequential correlations in the COBRA maps. Typical ϕ_0 values used are 1° to 30° for experiments with high SNR, while ϕ_0 values between 30° and 50° are preferred for less sensitive datasets. In the current version of the BATCH software, the same phase parameter ϕ_0 is used to calculate the coefficients M_{ij} for all (i, j) pairs and is not optimized to the SNR of individual NMR signals. For minimal sequential ambiguities, the final map should contain about one unique finite element for each row and column. As a rule of thumb, too many (too few) finite elements in the map indicate a ϕ_0 value chosen too large (too small). From our experience, 15° for time unshifted H–N–CA and H–N–CB datasets and 45° for time-shifted H–N–CA represented very good starting values. Since the COBRA map can be calculated in a few seconds, a few trials should be enough to obtain a reasonable set of ϕ_0 values.

The COBRA maps are calculated within the BATCH software as a background NMRPipe task and can be loaded for visualization ("Show COBRA maps"). Initially, the ^{1}H–^{15}N cross-peaks are ordered randomly (Fig. 4A–C). When a COBRA map with little ambiguity is obtained, the ^{1}H–^{15}N cross-peaks can be permutated such that fragments of unambiguously sequentially connected cross-peaks become apparent ("Order peaklist" button; Fig. 4D). For this operation, the product COBRA map (Fig. 4C) is first converted to a binary matrix $\tilde{M}$, containing only 0 and 1 values, using the intensity cutoff level defined in the main BATCH window. An optimized initial cutoff value is calculated when the COBRA maps are loaded from the intensity distribution in the product COBRA map. Tests on several data sets showed that the cutoff value calculated by the BATCH program provides an excellent starting point for the subsequent automatic resonance assignment step. If needed this initial value can be manually adjusted by

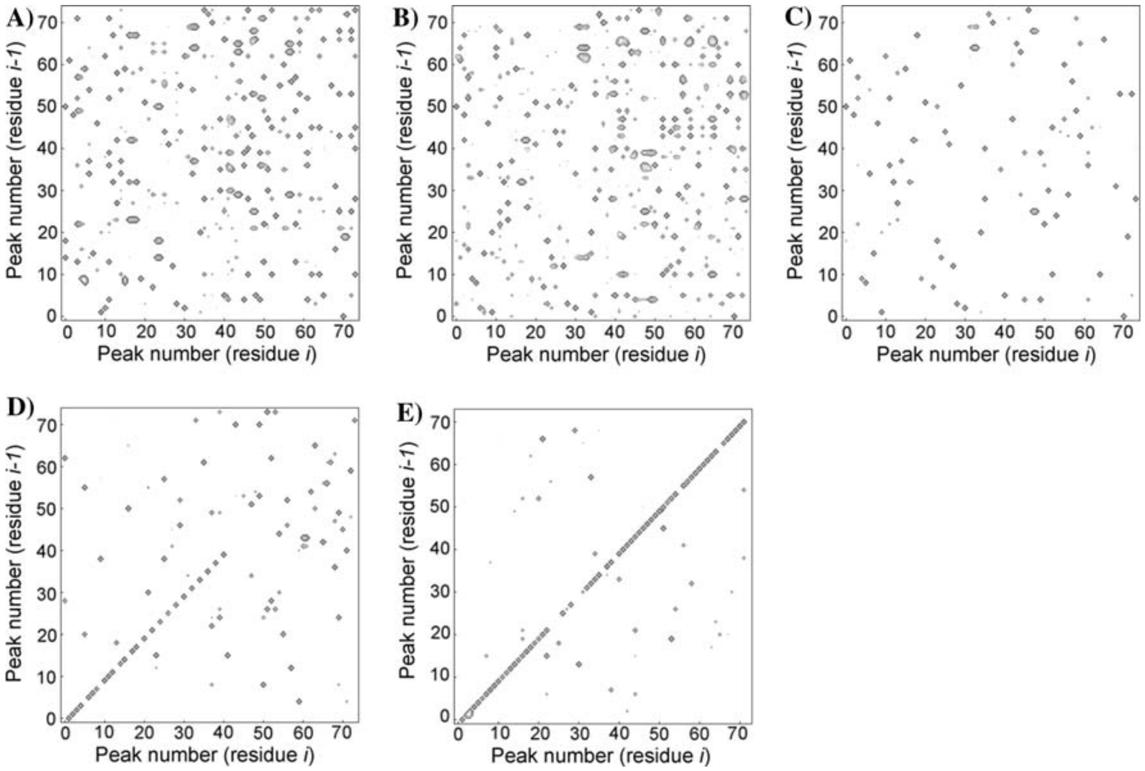


Fig. 4 COBRA maps calculated from a pair of H–N–CA experiments collected with standard ($t = 0$–3 ms) (**A**) and time-shifted ($t = 25$–28 ms) (**B**) ^{13}C time-domain sampling on the HMR sample. Panel **C** represents the product COBRA map obtained by point-by-point multiplication of the maps shown in panels **A** and **B**. The numbering along the *x* and *y* axes of the COBRA maps in panel **A**–**C** corresponds to the randomly ordered peak list generated from the automatic peak picking of the HADAMAC experiment. After cross-peak permutation and COBRA calculation, the unambiguously connected fragments of various lengths are clearly apparent as a shifted diagonal of non-vanishing COBRA elements in panel **D**. In the absence of assignment, the fragments are ordered from the longest to the shortest ones. Panel **E** represents the final COBRA map after the permutation based on the final resonance assignment. All these maps were calculated using default phase cutoff parameters of 15° and 45° for the standard and time-shifted H–N–CA data, respectively

simply changing the display level of the final COBRA map. Then, each pair of ^{1}H–^{15}N cross-peaks *i* and *j* is judged as unambiguously connected, if $\tilde{M}_{ij} = 1$ and if no other element of $\tilde{M}$ along column *i* and row *j* has an intensity of 1. Finally, the unambiguously connected fragments are assembled and ordered from the longest fragment to the shortest one. From this newly calculated COBRA map, the fragments appear as a long connected series of non-vanishing elements along the pseudo ($i - 1/i$) diagonal. This permutated COBRA map offers a convenient way to assess the quality of the NMR data in terms of the length and number of unambiguously identified protein fragments.

Automated sequential resonance assignment

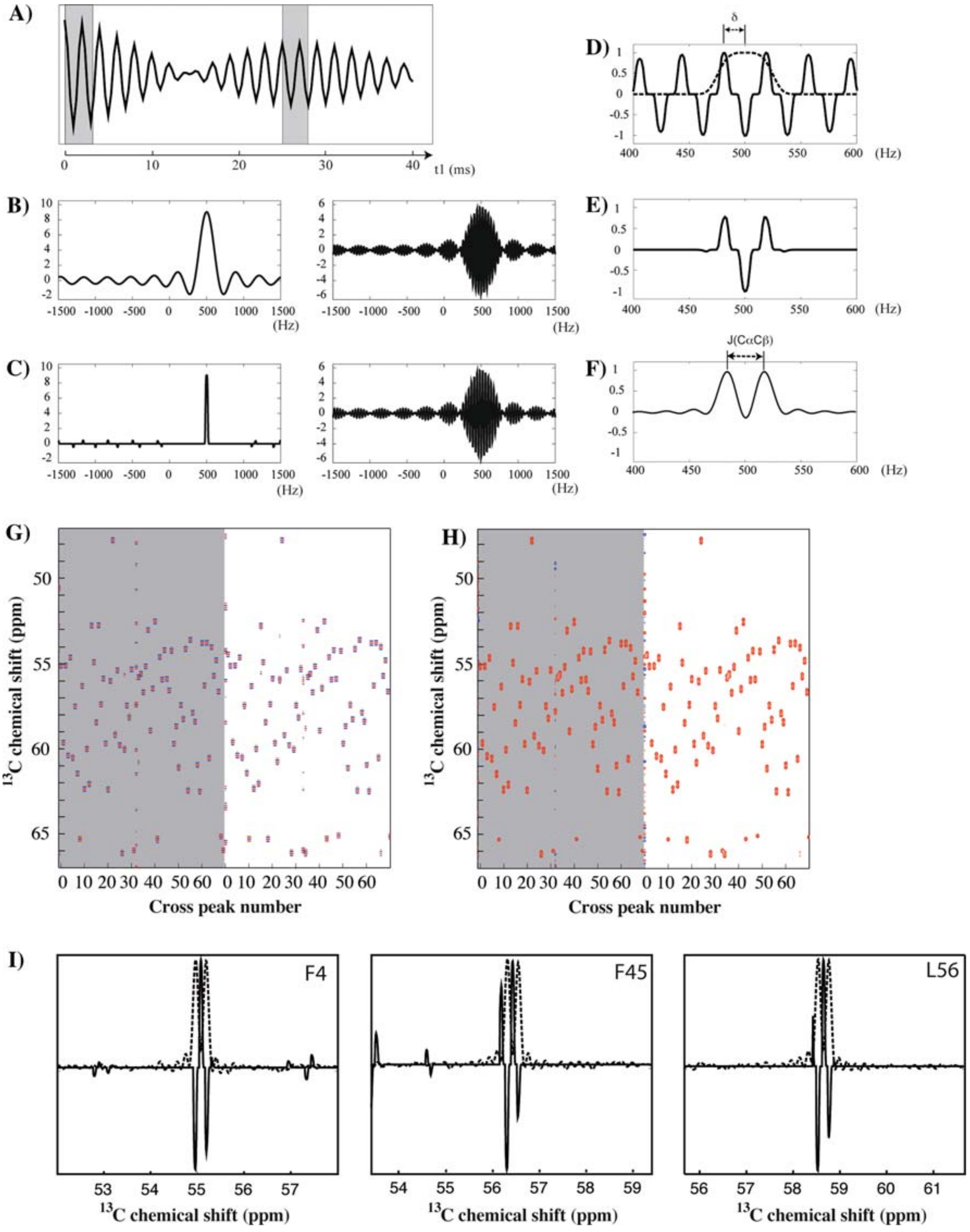
The HADAMAC and the COBRA analysis steps provide the amino-acid-type and sequential connectivity information required for sequential resonance assignment. The BATCH software includes an algorithm for automated assignment that is based on an efficient, and very fast best-first approach. Initially, fragments of unambiguously connected cross peaks are built in the same way as described in the previous paragraph. In a first step, a fragment is assigned to a position in the primary sequence if this position is unique for the given fragment in terms of correct amino-acid-type match for the connected residues. This step is repeated until no new fragment can be assigned in one run. In step 2, all positions allowed by the amino-acid composition of a still unassigned fragment are tested. If both extremities of the tested fragment can be sequentially connected to two (and only two) previously assigned residues, and if no other unassigned fragment can be connected at the same position, the fragment is assigned. Step 2 and step 1 are repeated sequentially until no new fragment can be assigned. Step 3 is identical to step 2, except that only one extremity is required to be connected with one (and only one) previously assigned residue for the fragment to be assigned. Again, step 3 and step 1 are repeated sequentially until no new fragment can be assigned. The whole procedure consists in applying the previous procedure three times. The extent of the assigned residues is monitored in a separate window (Fig. S3A) together with a confidence level of the assignment. This latter is higher when the corresponding fragment was assigned early in the assignment loop, and is reduced when it was assigned during one of the later steps. The assignment is also reported in the current 2D peak list. An additional window allows the user to obtain information about the unassigned fragments (Fig. S3B). The possible sequential connectivities, that are in agreement with the COBRA map, as well as the possible amino-acid-type-composition based locations onto the protein sequence are reported. The automated assignment algorithm is applied on the basis of a unique COBRA map, i.e. a unique set of cutoff phase parameters. However, it can be interesting to run the assignment algorithm with a different set of phase parameters, or after adjusting the amino-acid-type assignment, while keeping previously assigned residues fixed. Therefore, the cross-peaks that are assigned in the current 2D peaklist are kept assigned whenever the algorithm is run again. These assignments are marked as “Fix” in the report window. The additional button labeled “Remove Assignment from XPK” removes all assignments from the current 2D peaklist, which may be used to start the assignment again from scratch. Finally, the current 2D peaklist can be reordered according to the protein amino-acid sequence by clicking on “Create Ordered Peaklist From Assigned Peaks”. Calculating again the corresponding COBRA map (Fig. 4E) provides a good estimate of the overall exploitation of the experimental data and remaining unexploited connectivities.

In some cases, one may also want to use third party programs to achieve the assignment. All available data from COBRA and HADAMAC can be imported into the Smartnotebook (Slupsky et al. 2003) program that represents an independent NMRView module. Importing the data to Smartnotebook requires the initialization (“Initialize SNB”) before starting the application (“Launching SNB”). The assignment output from Smartnotebook can be transferred back to BATCH.

Fig. 5 Illustration of the algorithm for ^{13}C chemical shift extraction ▶ using targeted time domain sampling and phase-weighted Fourier transformation. **A** ^{13}C time domain data simulated for one signal resonating at 500 Hz with a damping rate of $R = 20\ \mathrm{s}^{-1}$, sampled each 3.33 ms (spectral width of SW = 3,000 Hz) from $t = 0$ to 40 ms. An additional cosine modulation was added to account for evolution under scalar coupling ($J = 35$ Hz). Only the real part of the signal is displayed. **B** Two regions corresponding to standard ($t = 0$–3 ms on the *left panel*), and time-shifted (25–28 ms on the *right panel*) data acquisition are extracted and Fourier transformed after 2 k zero-filling (without apodization). Only the real part of the data is displayed. **C** Same as **B** but after phase-weighted Fourier transformation with the phase parameters of 15° and 45° for the standard and time-shifted data, respectively. After phase-weighted Fourier transformation, the data are purely real. **D** Close-up view of panel **C** on the central peak (400–600 Hz region). The spectra corresponding to the standard and time shifted data are represented as *dashed* and *continuous lines*, respectively. The sign alternating peaks occur at $\delta = 1/[2\Delta + (N - 1)/\mathrm{SW}]$ (where $N = 10$ is the number of sampled points and $\Delta = 25$ ms) frequency. **E** Point-by-point multiplication of the two phase-weighted signals shown in **D**. **F** Spectrum obtained by 2 k zero-filling followed by Fourier transformation of the whole time domain data ($t = 0$–40 ms). Only the real part is shown. **G–H** Application to real NMR data recorded for ubiquitin. A pseudo 2D spectrum is built from the intra-residue and sequential H–N–CA experiments as explained in the text. The *grey* (*white*) area corresponds to signals from the intra-residual (sequential) experiment. The result of phase-weighted Fourier transformation of 20 complex points collected for times $t = 0$–3 ms, and $t = 25$–28 ms is shown in the panel **G**, while the spectrum of 2 k zero-filled Fourier transformed data, consisting in 110 complex points collected from $t = 0$ to 37 ms, is shown in the panel **H**. *Red* and *blue contours* correspond to positive and negative intensities, respectively. **I** 1D traces extracted from the pseudo-spectra (intraresidual) shown in **G** (*plain line*) and **H** (*dotted line*) corresponding to cross-peaks 3 (F4), 40 (F45) and 50 (L56). All 1D spectra were plotted so that the central peak has positive intensity

Extraction of ^{13}C chemical shifts

^{13}C chemical shifts represent an important source of secondary structural information. In contrast to other assignment methods, the BATCH strategy does not directly exploit the ^{13}C chemical shifts for resonance assignment. Although the 3D spectra can be fully Fourier transformed and peak-picked, this solution is under optimal. Furthermore, the classical Fourier transformation is not able to deal with targeted acquisition data as used here for the ^{13}Cα nucleus. We therefore developed an additional tool to automatically extract ^{13}C chemical shifts from the recorded H–N–C experiments. The algorithm is illustrated in Fig. 5 for simulated and experimental ^{13}C time domain data recorded for $t_1 = 0$–3 and 25–28 ms. First, each data set is Fourier transformed along the ^{13}C time domain (Fig. 5B). An additional first-order phase correction is automatically calculated and applied to the time-shifted data to compensate for the delayed first point acquisition. In analogy to the COBRA method, application of an additional phase weighting to each complex point of the spectrum increases the apparent spectral resolution (“phase-weighted Fourier transformation”). In order to take into account the possibility of signals of opposite signs due to e.g. the presence or absence of C–C coupling evolution (i.e. Gly vs. all other residues), a slightly modified phase-weighting procedure is used here. We use the following expression for the phase-weighted Fourier transformation: $\mathrm{FT}_{\mathrm{pw}}(\omega) = \Re(\mathrm{FT}(\omega))\Xi(\phi_1, \phi_0)$, with $\phi_1 = \phi$, if $-90° < \phi < 90°$ and $\phi_1 = |\pi - \phi|$, if $\phi < -90°$ or $\phi > 90°$. In this equation, ϕ and $\Re(\mathrm{FT}(\omega))$ are the angular phase and the real part of $\mathrm{FT}(\omega)$, respectively. The phase-dependent weighting function was defined here as $\Xi(\phi_1, \phi_0) = \mathrm{e}^{-(|\phi_1|/\phi_0)^4}$ and depends on the phase cutoff parameter ϕ_0. The same cutoff phase parameters are used as previously optimized for the COBRA analysis. For standard time-domain sampling



starting from $t = 0$ a unique sharp peak is observed (Fig. 5C, left spectrum and Fig. 5D, dashed line), as well as small artifacts that are due to truncation effects in the absence of any signal apodization. For delayed time-domain sampling starting from $t = \Delta$ with an increment of 1/SW, a series of peaks is obtained with alternating sign that are separated by $\delta = 1/[2\Delta + (N{-}1)/\mathrm{SW}]$ (Fig. 5C, right spectrum and Fig. 5D, continuous line). The intensity envelope of these peaks corresponds to the magnitude spectrum obtained from non phase-weighted Fourier transformation of standard unshifted time-domain data of the same length. The central peak is negative due to the cosine modulation under homonuclear ^{13}C–^{13}C scalar coupling evolution [$\cos(\pi Jt) \approx -1$ for t = 25–28 ms and

$J = 30$–40 Hz]. The spectra obtained from the standard and time-shifted data sets are then combined by simple point-by-point multiplication (Fig. 5E). This procedure provides an apparent higher frequency resolution (sharper peaks) compared to a spectrum obtained by classical Fourier transformation of a complete time-domain data set recorded from $t = 0$–37 ms (Fig. 5F). Similar results were obtained on experimental data (Fig. 5G–I). In addition, no line splitting due to homonuclear ^{13}C–^{13}C scalar couplings is observed. It should be noted that the degree of the removal of the artifacts introduced by the delayed acquisition (fast modulation) depends on the cutoff phase chosen for the processing of the unshifted time domain data. A larger cutoff phase leads to broader peaks and thus to more incomplete artefact suppression after multiplication. In the current conditions, residual sign alternating artifacts occur at the frequency $\delta = 18.8$ Hz that is incidentally close to half of the ^{13}C–^{13}C scalar coupling ($\sim$35 Hz). For most ^{13}C nuclei, two different traces are available: one in the sequential experiment at the ^{1}H, ^{15}N frequencies of the following residue, and one in the intra-residue experiment at the ^{1}H, ^{15}N frequencies of the same residue. In order to remove contributions of other ^{13}C nuclei possibly observed in these spectra due to partial ^{1}H–^{15}N overlap, the spectra calculated separately from the two time domains are multiplied together prior to chemical shift extraction. Finally, chemical shifts are extracted as the point of maximum absolute intensity in each spectrum.

In practice, a pseudo-spectrum of size $2n \times S$ is calculated for each pair of H–N–C experiments, where n is the total number of ^{1}H–^{15}N correlation peaks, and S is the size of the 1D ^{13}C spectrum (Fig. 5G). The n 1D ^{13}C spectra obtained from the intra-residue experiment are stacked together followed by the n ^{13}C 1D spectra obtained from the sequential experiment. The ^{13}C chemical shifts extraction section of the main BATCH window permits the control of the first order phase correction for the time shifted acquisition (calculated from the value of the first delay), of spectral reversion, and spectral calibration. The BATCH software generates scripts for data processing in NMRPipe by clicking on the button “Process Spectra”. Due to the very narrow lineshape after phase-weighted FT (a few Hz, see Fig. 5E), the zero filling should be extended, typically to 2 k points. The ^{13}C chemical shift values are automatically extracted (“Get Shifts” button) and the ^{1}H, ^{15}N and ^{13}C chemical shifts assignment are stored in a table. If needed, this routine also enables automated “unfolding” of aliased 3D cross-peaks along the ^{13}C dimension based on the expected amino-acid dependent ^{13}C chemical shifts range. Of note, the “Get Shifts” procedure works on pairs of sequential and intra-residue H–N–C experiments, as well as on a single H–N–C data set.

The robustness of this automated ^{13}C chemical shift extraction module was evaluated on a ubiquitin dataset comprising sequential and intra-residue HNCA data, recorded for the time intervals $t = 0$–3 ms and $t = 25$–28 ms. As a reference we used chemical shifts that were extracted by a peak picking routine from H–N–CA spectra recorded in the usual way with high resolution along the ^{13}C dimension ($t_1^{\max} = 37$ ms) using the same pulse sequences. The C^{α} chemical shifts were then obtained as the average position of the resolved peak doublets along the ^{13}C dimension. The difference in ^{13}C chemical shifts obtained by the two methods is 0.008 ± 0.024 ppm. This result illustrates the robustness and accuracy of the BATCH-embedded protocol for ^{13}C chemical shift extraction. This method requires only a few seconds of time and provides reliable chemical shift values for later usage.

Examples of protein resonance assignment using BATCH

The minimal amount of information required for protein resonance assignment depends on the experimental sensitivity (NMR spectrometer, protein size and concentration), but also on the peptide sequence, amino-acid-type distribution, chemical shifts dispersion, and local dynamics of the protein under investigation. It is therefore generally impossible to decide at the beginning of an NMR study how many different experiments are required, and what experimental time is needed for each of them to detect all or most of the expected correlation peaks. In the BATCH strategy, NMR data are recorded in a stepwise manner, thus allowing adjustment of the total experimental time to the intrinsic properties of the experimental setup. The general flowchart proposed for BATCH data collection is shown in Fig. 6. The first minimal data set used for BATCH assignment consists in a ^{1}H–^{15}N HSQC, a HADAMAC experiment, and a pair of sequential and intra-residue 3D H–N–CA spectra. For the latter, ^{13}C time evolution is sampled from 0 to 3 ms. This initial set of experiments can be recorded in <1 h using the fast data acquisition schemes described above and a minimal phase cycle of two scans. Once a data set has been recorded, BATCH data processing and analysis is performed, and the final result is analyzed in terms of completeness of the assignment. In case of incomplete assignment, a decision needs to be made whether the BATCH assignment failed because of limited SNR in the recorded data set, or because of significant ^{13}C chemical shift degeneracy. Low SNR leads to missing amino-acid type determination from the HADAMAC spectra, and missing connectivities in the COBRA map, while ^{13}C chemical shift degeneracy is responsible for multiple correlation peaks in the COBRA connectivity

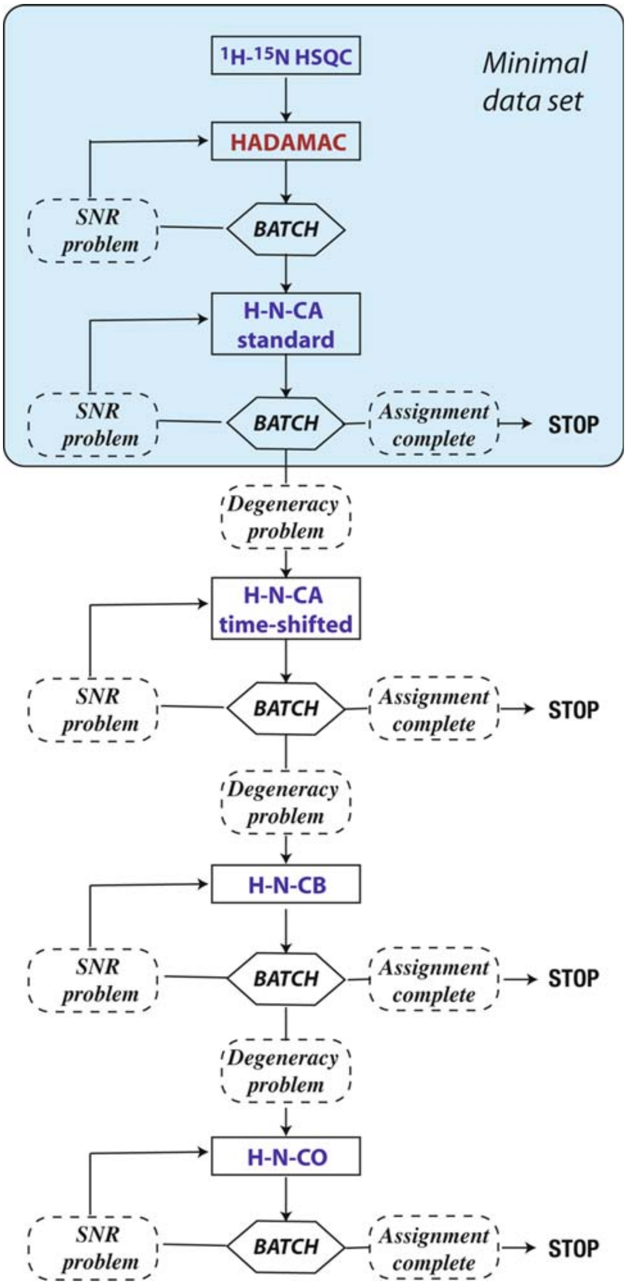


Fig. 6 Flowchart for iterative NMR data collection used in the BATCH strategy

map. If SNR is the main problem, the last set of experiments is run again and added to the previously collected dataset. If ^{13}C chemical shift degeneracy has been identified as the bottleneck, additional connectivity information is obtained from pairs of time-shifted H–N–CA, H–N–CB, or H–N–CO experiments. This procedure is repeated until sufficiently complete resonance assignment is obtained.

The strategy described above is well suited for protein studies that only require the assignment of the ^{1}H–^{15}N correlation spectra, but not necessarily side chain assignments. This applies to the study of protein backbone dynamics, chemical shift mapping of molecular interactions, and protein fold determination based on sparse NMR data such as backbone RDCs and amide ^{1}H–^{1}H nOes. If the main objective is the determination of a high-resolution structure, additional experiments need to be recorded once the sequential resonance assignment has been obtained, in order to extend the resonance assignment to the side chain ^{13}C and ^{1}H nuclei.

In order to demonstrate its performance, we have applied the BATCH assignment strategy to four different protein samples: (a) a 82 residue protein involved in heavy-metal resistance in *Cupriavidus metallidurans* (HMR, 82 residues, 1 mM, 50 mM MES, pH 6.0), (b) ubiquitin (76 residues, 1.9 mM, pH 6.7), (c) a C-terminal fragment of the Vesicular Stomatitis Virus phosphoprotein (VSV-P$_{CTD}$, 73 residues, 1.2 mM, pH 7.5) (Ribeiro et al. 2008), and (d) an RNA-binding domain of *Arabidopsis thaliana* Hyl1 (77 residues, 0.7 mM, pH 7.0). All experiments were recorded at 25°C on a 600 MHz spectrometer equipped with a cryogenically cooled probe. The experimental settings together with the total acquisition times are detailed in Tables T1–T4 in the Supplementary material. The processing and final resonance assignment was done on a PowerBook G4 laptop.

HMR: a typical example of a small globular protein

The 82-residue protein HMR, currently under structural investigation in our laboratory, has been chosen as a first real test case without known resonance assignment prior to the BATCH assignment described here. A total of 72 (out of 73 expected) correlation peaks were detected from the analysis of the HADAMAC spectra (Fig. S1C in Supplementary material). Unambiguous amino-acid type assignment from the HADAMAC spectrum, recorded in 22 min, was obtained for 59 residues (cross peaks) by the automated BATCH analysis. The amino-acid types of the remaining 13 residues were then added from manual inspection of the HADAMAC spectra. Connectivity information was obtained from pairs of sequential and intra-residue H–N–CA experiments. Standard (t = 0–3 ms) and time-shifted (t = 25–28 ms) data sets were recorded in 30 and 70 min, respectively, and default phase cutoff values of 15° and 45° were used for the COBRA analysis. The resulting individual COBRA maps, as well as their combination by point-by-point multiplication, are shown in Fig. 4A–C. Using the automatically determined intensity cutoff value, 16 fragments were identified with the longest fragment composed of five residues, and the vast majority of two residues only (Fig. 4C). Despite this highly ambiguous connectivity information, the combination with the HADAMAC-derived amino-acid type information proved sufficient for automatically assigning all the

72 detected ^{1}H–^{15}N correlation peaks to unambiguous positions in the peptide sequence, leaving as the only unassigned positions, the eight proline residues, the N-terminal residue, and residue Q74. The assignment was confirmed by the manual analysis of a set of standard 3D triple-resonance spectra (H–N–CA and H–N–CB). This analysis also revealed the reason for the missing residue Q74 in the BATCH assignment. The neighboring residues Q74 and V75 have degenerate ^{1}H and ^{15}N chemical shifts, giving rise to a single correlation peak in the HSQC spectrum. As the preceding residues P73 and Q74 belong to the same amino-acid class (*Rest*), this degeneracy could not be resolved in the HADAMAC spectra that also revealed a single peak. This example of overlapping ^{1}H–^{15}N correlation peaks points out one weakness of the BATCH strategy which relies on the assumption of minimal overlapping in the ^{1}H–^{15}N HSQC spectra. Nevertheless, almost complete, and accurate backbone resonance assignment was obtained for HMR from a single run of the BATCH assignment using a data set of HADAMAC and H–N–CA data acquired in about 2 h. The data analysis was carried out with default cutoff parameters in a stepwise manner after each completed experiment. The final assignment of HMR was achieved in ~2 h 30 including data collection.

Ubiquitin: a favorable case for BATCH assignment using a minimal NMR data set

In order to evaluate the performance of the BATCH strategy for a highly concentrated, well behaving (in terms of frequency dispersion and intensity distribution) protein sample, a 1.9 mM ubiquitin sample was used. First, a ^{1}H–^{15}N BEST-HSQC spectrum was recorded in an experimental time of 2 min. All expected 70 cross peaks (76 residues minus 3 prolines, the N-terminal residue, and residues E24 and G53) were detected in this spectrum in agreement with results reported in the literature. Automatic BATCH analysis of the HADAMAC spectrum, recorded in 20 min, yielded amino-acid-type assignment for 64 (out of 70) correlation peaks. The six remaining cross peaks were then assigned to one of the amino-acid groups by subsequent manual inspection of the HADAMAC spectrum. Finally, a pair of sequential and intra-residue H–N–CA experiments was performed to complete the BATCH data set. For each data set, 10 complex points were collected in the ^{13}C$^{\alpha}$ time domain from $t = 0$ to 3 ms, resulting in experimental times of 11 and 13 min, respectively.

Based on this minimal data set, a single BATCH assignment run using the default phase and intensity cutoff values only resulted in 3% of the assignments. Therefore we tried a different, iterative strategy. Several runs of BATCH resonance assignment were performed using different parameter settings for the COBRA analysis. In a first step, the COBRA phase and intensity cutoff parameters were set to 1° and 40%, respectively. This resulted in a COBRA map with a single correlation peak for most residues characterized by high SNR signals, while no COBRA correlation peak was detected for residues giving rise to lower SNR peaks in the NMR spectra. With these cutoff values, a total of 51 residues could be unambiguously assigned by the automated procedure. This partial assignment was then fixed, and a second BATCH assignment run, using a slightly increased phase cutoff of 3° and the same intensity cutoff (40%), lead to unambiguous assignment of 17 additional residues. At this stage 68 (out of 70) ^{1}H–^{15}N correlations were (correctly) assigned by the automatic BATCH assignment protocol. The unassigned cross-peaks corresponds to T9 and N25 that belong both to the less discriminative amino-acid-type group *Rest* (they follow residues L8 and E24, respectively), and no sequential connectivity was detected in the COBRA map for the chosen phase (3°) and intensity (40%) cutoffs. A third BATCH run, using a further increased cutoff phase ($\phi_0 = 10°$), allowed retrieving the missing correlation peaks in the COBRA map, and yielded correct and unambiguous assignment for all of the 70 correlation peaks observed in the ^{1}H–^{15}N HSQC spectrum.

The example of ubiquitin demonstrates that in favorable cases (high SNR, homogeneous cross peak intensities, and good ^{1}H–^{15}N peak dispersion), the minimal NMR data set required for BATCH is sufficient to obtain complete assignment. This data set can be recorded in only about 50 min of overall data acquisition time, making this approach very attractive for protein samples of limited time stability. An iterative run of the BATCH assignment protocol, using different COBRA phase cutoffs, makes optimal use of the high SNR contained in the NMR data for frequency discrimination, and chemical shift matching based on a single nuclear spin species (^{13}C$^{\alpha}$).

Hyl1 and VSV-P$_{CTD}$: examples of small proteins with heterogeneous peak intensity distribution

For the 77-residue Hyl1 and the 73-residue VSV-P$_{CTD}$ proteins, the same initial data sets (^{1}H–^{15}N HSQC, HADAMAC, and H–N–CA pair) were recorded as described above for ubiquitin. 66 (out of 75), and 64 (out of 70) correlation peaks were detected in the HSQC and HADAMAC spectra of Hyl1 and VSV-P$_{CTD}$, with a sufficiently high SNR. The remaining peaks were either undetectable or of very low intensity, suggesting chemical or conformational exchange induced line broadening in some parts of the protein.

Using this minimal data set, only about 30–40% of the detected ^{1}H–^{15}N peaks could be unambiguously assigned with default or further optimized cutoff values. To reduce the ambiguities in the COBRA map, an additional H–N–CA pair of experiments was collected with time-shifted data sampling in the $^{13}C^{\alpha}$ dimension ($t = 25$–28 ms). The default phase cutoff values of $\phi_0 = 15°$ and $45°$ were chosen for the COBRA analysis of the standard, and time-shifted HNCA experiments to account for the intrinsically lower SNR in the time-shifted data set. For the Hyl1 protein this extended data set, recorded in $\sim$2 h, proved sufficient for unambiguous (and correct) assignment of all 66 ^{1}H–^{15}N correlations detected in the HADAMAC spectra. The remaining unassigned residues are mainly located at the N terminus (H2–C9) of the protein fragment and undergo slow time-scale (ms) motion in solution. For VSV-P_{CTD}, only 42 (out of 64) ^{1}H–^{15}N correlations could be unambiguously and correctly assigned by BATCH using this extended data set. This score increased to 54 (out of 64) assignments when an additional H–N–CB pair of experiments ($t = 0$–3 ms), recorded in $\sim$1 h 20, was added to solve remaining ambiguities in the COBRA map due to degenerate $^{13}C^{\alpha}$ chemical shifts. Note that although no complete assignment was obtained, all the proposed assignments are correct due to the very conservative nature of the BATCH assignment protocol.

The different outcome of the automated BATCH assignment strategy for the two proteins of similar size is mainly explained by the location of the missing ^{1}H–^{15}N correlation peaks. While for Hyl1 they mainly cluster in a single part (N terminus) of the peptide chain, various small peptide stretches are undetected for VSV-P_{CTD} because of extensive line broadening in the NMR spectra. As a consequence, smaller peptide fragments are identified from the COBRA analysis, and the HADAMAC-derived amino-acid-type information may not always be sufficient to place a segment unambiguously onto the peptide sequence. For a more complete resonance assignment the experimental time has to be significantly increased in order to detect correlation peaks for the missing residues in the HADAMAC and H–N–C correlation experiments that are of sufficient SNR for the BATCH analysis.

Conclusions and perspectives

Speeding up backbone resonance assignment of proteins is an active field of current research within the NMR community. In this manuscript, we introduce and demonstrate the performance of a new strategy together with an original software platform for iterative time-optimized resonance assignment of small proteins. The BATCH protocol is based on a limited set of NMR data recorded in short experimental time using a combination of various fast NMR data acquisition tools. The BATCH software has been designed for rapid, user-friendly data processing, extraction of the relevant spectral information, and automated sequential resonance assignment. All calculations required during the different BATCH steps are completed within seconds on a single CPU laptop computer, making BATCH a valuable tool for real-time analysis of NMR data sets available at different well-defined intermediate steps of the BATCH data acquisition protocol. The result of the BATCH analysis provides a measure of the completeness of the information contained in the data recorded so far, and a guideline of what NMR data need to be recorded for the next step. The BATCH strategy is thus ideally suited for target-driven NMR data acquisition where NMR data are recorded in small "pieces" that are analyzed while the next piece of data is recorded, and data acquisition is stopped once all (or most) of the information expected from a particular data set has been obtained.

A particularity of the BATCH strategy, that distinguishes BATCH from other existing methods for backbone resonance assignment, is that the amino-acid-type and sequential connectivity information is gathered from different NMR data sets. The HADAMAC experiment provides highly discriminative amino-acid-type identification, while H–N–C correlation experiments are used for the identification of sequentially connected ^{1}H–^{15}N frequency pairs (residues). As a side effect, the latter experiments also provide the ^{13}C chemical shift information. The "decoupling" of these two types of information allows individual adjustment of the corresponding NMR experiments for a specific goal. In particular, because of the powerful HADAMAC amino-acid-type discrimination, only little additional connectivity information may be required for unambiguous sequential resonance assignment, as demonstrated here for the protein ubiquitin. This is in contrast to standard assignment strategies, where a complete set of H–N–CA, H–N–CB, and H–N–CO-type experiments needs to be recorded, because the correlation information in these spectra not only serves for sequential connectivity identification, but also for amino-acid-type discrimination (through the $^{13}C^{\alpha}$, $^{13}C^{\beta}$, and ^{13}CO chemical shifts).

The BATCH strategy has some intrinsic limitations with respect to the proteins that are amenable to BATCH assignment. BATCH is based on a single 2D ^{1}H–^{15}N peaklist, with each peak supposed to correspond to a single residue in the peptide sequence. Therefore well-dispersed ^{1}H–^{15}N correlation spectra with limited peak overlap are a prerequisite for successful BATCH assignment. Note, that partial peak overlap can often be solved by close visual inspection of the HADAMAC spectrum (Fig. 3). In addition, the sensitivity of the HADAMAC, and some of the H–N–C correlation experiments, e.g. the H–N–CB experiments optimized for the detection of a single peak per residue, rapidly decreases with increasing protein size, because of ^{13}C-relaxation-induced signal loss during the long transfer delays required for these experiments.

Finally, the HADAMAC experiment exploits $^{1}H–^{13}C$ spin-coupling topologies requiring fully protonated samples, thus excluding perdeuterated or partially deuterated proteins from BATCH assignment. For these reasons, BATCH is ideally suited for backbone resonance assignment of small globular proteins. The results shown here for several proteins (70–80 residues) make us believe that BATCH will prove useful for application to monomeric globular proteins and protein fragments in the size range of up to ~120–150 residues where deuteration is generally not required for an NMR investigation.

In principle, the BATCH strategy may also be attractive for intrinsically unfolded or denatured proteins, where spin relaxation is not a limiting factor for the various correlation experiments required for BATCH assignment. However, the low $^{1}H–^{15}N$ spectral dispersion observed for proteins lacking a stable tertiary fold renders the COBRA and HADAMAC analysis difficult. Although not yet implemented in the current version of the BATCH software, this overlap problem may be overcome by additional ^{13}CO frequency editing to disperse the overlapping $^{1}H–^{15}N$ peaks along a third (^{13}CO) dimension. ^{13}CO frequency editing is compatible with the HADAMAC and all sequential and intra-residue H–N–C correlation experiments, and requires only minor modifications in the corresponding pulse sequences. As demonstrated recently, experimental times can still be kept short by simultaneous ASCOM optimization of the spectral widths in the ^{15}N- and ^{13}CO-dimensions (Lescop et al. 2007b) and the use of BEST principle. A 3D $^{1}H–^{15}N–^{13}CO$ peaklist then replaces the 2D $^{1}H–^{15}N$ peaklist for the subsequent BATCH analysis and assignment. Implementation of this 3D-based approach within the BATCH software is currently under way in our laboratories.

Acknowledgments This work was supported by the Commissariat à l'Energie Atomique, the Centre National de la Recherche Scientifique, the University Grenoble 1, the French Research Agency (ANR-JCJC-05-0077), and the European commission (I3, EU-NMR, Contract No. 026145). We thank I. Ayala for preparation of the ubiquitin sample, P. Gans for the calmodulin sample, and B. Bersch, K.-M. Derfoufi, G. Vandenbussche, R. Rasia, J. Boisbouvier, and M. Jamin, for making samples of HMR, Hyl1 and VSV-P_{CTD} available for this study.

References

Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc Natl Acad Sci U S A 101:9642–9647

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

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

Bruschweiler R (2004) Theory of covariance nuclear magnetic resonance spectroscopy. J Chem Phys 121:409–414

Brutscher B (2004) Combined frequency- and time-domain NMR spectroscopy. Application to fast protein resonance assignment. J Biomol NMR 29:57–64

Brutscher B, Simorre JP, Caffrey MS, Marion D (1994) Design of a complete set of two-dimensional triple-resonance experiments for assigning labeled proteins. J Magn Reson B 105:77–82

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

Cornilescu G, Bahrami A, Tonelli M, Markley JL, Eghbalnia HR (2007) HIFI-C: a robust and fast method for determining NMR couplings from adaptive 3D to 2D projections. J Biomol NMR 38:341–351

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Deschamps M, Campbell ID (2006) Cooling overall spin temperature: protein NMR experiments optimized for longitudinal relaxation effects. J Magn Reson 178:206–211

Friedrichs MS, 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

Frydman L, Scherf T, Lupulescu A (2002) The acquisition of multidimensional NMR spectra within a single scan. Proc Natl Acad Sci U S A 99:15858–15862

Frydman L, Lupulescu A, Scherf T (2003) Principles and features of single-scan two-dimensional NMR spectroscopy. J Am Chem Soc 125:9204–9217

Gal M, Schanda P, Brutscher B, Frydman L (2007) UltraSOFAST HMQC NMR and the repetitive acquisition of 2D protein spectra at Hz rates. J Am Chem Soc 129:1372–1377

Hiller S, Fiorito F, Wuthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci U S A 102:10876–10881

Jaravine VA, Orekhov VY (2006) Targeted acquisition for real-time NMR spectroscopy. J Am Chem Soc 128:13421–13426

Jaravine V, Ibraghimov I, Orekhov VY (2006) Removal of a time barrier for high-resolution multidimensional NMR spectroscopy. Nat Methods 3:605–607

Jaravine VA, Zhuravleva AV, Permi P, Ibraghimov I, Orekhov VY (2008) Hyperdimensional NMR spectroscopy with nonlinear sampling. J Am Chem Soc 130:3927–3936

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

Kazimierczuk K, Kozminski W, Zhukov I (2006) Two-dimensional Fourier transform of arbitrarily sampled NMR data sets. J Magn Reson 179:323–328

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393

Kupce E, Freeman R (2003a) Frequency-domain Hadamard spectroscopy. J Magn Reson 162:158–165

Kupce E, Freeman R (2003b) Projection-reconstruction of three-dimensional NMR spectra. J Am Chem Soc 125:13958–13959

Lescop E, Brutscher B (2007) Hyperdimensional protein NMR spectroscopy in peptide-sequence space. J Am Chem Soc 129: 11916–11917

Lescop E, Schanda P, Brutscher B (2007a) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson 187:163–169

Lescop E, Schanda P, Rasia R, Brutscher B (2007b) Automated spectral compression for fast multidimensional NMR and increased time resolution in real-time NMR spectroscopy. J Am Chem Soc 129:2756–2757

Lescop E, Rasia R, Brutscher B (2008) Hadamard amino-acid-type edited NMR experiment for fast protein resonance assignment. J Am Chem Soc 130:5014–5015

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

Lin G, Xu D, Chen ZZ, Jiang T, Wen J, Xu Y (2002) An efficient branch-and-bound algorithm for the assignment of protein backbone NMR peaks. Proc IEEE Comput Soc Bioinform Conf 1:165–174

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

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

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

Marion D (2005) Fast acquisition of NMR spectra using Fourier transform of non-equispaced data. J Biomol NMR 32:141–150

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

Monleon D, Colson K, Moseley HN, Anklin C, Oswald R, Szyperski T, Montelione GT (2002) Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture, and an integrated set of spectral analysis tools. J Struct Funct Genomics 2:93–101

Olson JB Jr, 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 IV, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Orekhov VY, Ibraghimov I, Billeter M (2003) Optimizing resolution in multidimensional NMR by three-way decomposition. J Biomol NMR 27:165–173

Pantoja-Uceda D, Santoro J (2008) Amino acid type identification in NMR spectra of proteins via beta- and gamma-carbon edited experiments. J Magn Reson 195:187–195

Pervushin K, Vogeli B, Eletsky A (2002) Longitudinal (1)H relaxation optimization in TROSY NMR spectroscopy. J Am Chem Soc 124:12898–12902

Ribeiro EA Jr, Favier A, Gerard FC, Leyrat C, Brutscher B, Blondel D, Ruigrok RW, Blackledge M, Jamin M (2008) Solution structure of the C-terminal nucleoprotein-RNA binding domain of the vesicular stomatitis virus phosphoprotein. J Mol Biol 382:525–538

Rovnyak D, Frueh DP, Sastry M, Sun ZY, Stern AS, Hoch JC, Wagner G (2004) Accelerated acquisition of high resolution triple-resonance spectra using non-uniform sampling and maximum entropy reconstruction. J Magn Reson 170:15–21

Schanda P, Van Melckebeke H, Brutscher B (2006) Speeding up three-dimensional protein NMR experiments to a few minutes. J Am Chem Soc 128:9042–9043

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

Vitek O, Vitek J, Craig B, Bailey-Kellogg C (2004) Model-based assignment and inference of protein backbone nuclear magnetic resonances. Stat Appl Genet Mol Biol 3:6

Vitek O, Bailey-Kellogg C, Craig B, Kuliniewicz P, Vitek J (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. Bioinformatics 21(Suppl 2):230–236

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

Volk J, Herrmann T, Wuthrich K (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J Biomol NMR 41:127–138

Wan X, Lin G (2006) A graph-based automated NMR backbone resonance sequential assignment. Comput Syst Bioinformatics Conf 5:5–66

Wan X, Lin G (2007) GASA: a graph-based automated NMR backbone resonance sequential assignment program. J Bioinform Comput Biol 5:313–333

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

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