J Biomol NMR (2008) 42:77–86
DOI 10.1007/s10858-008-9269-8

ARTICLE

Automatic assignment of protein backbone resonances by direct spectrum inspection in targeted acquisition of NMR data

Leo E. Wong · James E. Masse · Victor Jaravine · Vladislav Orekhov · Konstantin Pervushin

Received: 5 August 2008 / Accepted: 7 August 2008 / Published online: 11 September 2008

Abstract The necessity to acquire large multidimensional datasets, a basis for assignment of NMR resonances, results in long data acquisition times during which substantial degradation of a protein sample might occur. Here we propose a method applicable for such a protein for automatic assignment of backbone resonances by direct inspection of multidimensional NMR spectra. In order to establish an optimal balance between completeness of resonance assignment and losses of cross-peaks due to dynamic processes/degradation of protein, assignment of backbone resonances is set as a stirring criterion for dynamically controlled targeted nonlinear NMR data acquisition. The result is demonstrated with the 12 kDa ^{13}C,^{15} N-labeled apo-form of heme chaperone protein CcmE, where hydrolytic cleavage of 29 C-terminal amino acids is detected. For this protein, 90 and 98% of manually assignable resonances are automatically assigned within 10 and 40 h of nonlinear sampling of five 3D NMR spectra, respectively, instead of 600 h needed to complete the full time domain grid. In addition, resonances stemming from degradation products are identified. This study indicates that automatic resonance assignment might serve as a guiding criterion for optimal run-time allocation of NMR resources in applications to proteins prone to degradation.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-008-9269-8) contains supplementary material, which is available to authorized users.

L. E. Wong · K. Pervushin (✉)
School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore
e-mail: kpervushin@ntu.edu.sg;
konstantin.pervushin@unibas.ch

J. E. Masse
Laboratorium für Physikalische Chemie, ETH-Hönggerberg, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland

Present Address:
J. E. Masse
National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892-0520, USA

V. Jaravine · V. Orekhov
Swedish NMR Centre, Gothenburg University, Box 465, Gothenburg 40530, Sweden

Present Address:
V. Jaravine
Institute of Biophysical Chemistry, J. W. Goethe-University Frankfurt, Max-von-Laue-Street 9, 60438 Frankfurt am Main, Germany

K. Pervushin
Biozentrum of University Basel, Klinlegberg-Str. 70, 4056 Basel, Switzerland

Keywords MDD · Automatic resonance assignment · Nonlinear data sampling · Targeted NMR data acquisition

Abbreviations

DSS	2,2-Dimethyl-2-silapentane-5-sulfonate, sodium salt
NMR	Nuclear magnetic resonance
RHP	Relative hypothesis prioritization
MDD	Multidimensional decomposition
NLS	Nonlinear sampling
TA	Targeted acquisition

Introduction

Resonance assignment is typically recognized as an essential intermediate benchmark in NMR analysis of biomolecules with many downstream steps in 3D structure reconstructions being automated (Nilges et al. 1997; Herrmann et al. 2002a, b). The use of this benchmark as a target for NMR data acquisition (Jaravine and Orekhov 2006) would result in a streamlined process greatly contributing to the efficiency of NMR structure determination. In the last few years several attempts at automating resonance assignment demonstrated considerable progress (Atreya et al. 2000, 2002; Tian et al. 2001; Pristovsek et al. 2002; Coggins and Zhou 2003; Hitchens et al. 2003; Jung and Zweckstetter 2004; Langmead and Donald 2004; Eghbalnia et al. 2005a, b; Lin et al. 2005; Masse and Keller 2005; Masse et al. 2006; Wu et al. 2005). In the most favorable cases, sequence specific resonance assignment and structural NOE assignment can be done in parallel with structure calculations (Grishaev and Llinas 2004; Grishaev et al. 2005; Takeda et al. 2007). However, despite the impressive efforts directed to this problem, there is often still enough complexity left to require at least some human assistance. This situation is further aggravated when resonance assignment is set as a target for spectral acquisition with less stable proteins that are prone to degradation during measurement. In this case, not only the spectrum analysis should be well advanced to deal with a host of problems such as missing and spurious cross-peaks, wide amplitude variation of the detectable resonances (e.g., due to dynamic line broadening), raise of signals from degradation products, but fast data acquisition methods might be essential.

Recently several accelerated acquisition schemes of multidimensional NMR spectra were developed, which can be used to control real-time data acquisition targeted to as complete as possible assignment of NMR resonances. For concentrated solutions of small- and medium-sized proteins (data sampling limited cases), “projection reconstruction” was introduced (Kupce and Freeman 2004). Adaptive selection of the tilt-angles was proposed helping to optimize the time usage of spectrometer (Kupce and Freeman 2004; Eghbalnia et al. 2005a, b; Hiller et al. 2005, 2007). A robust scheme can be created by combining the basic features of projection reconstruction and “fast” data acquisition approaches, e.g., a spatially encoded (Frydman et al. 2003) and relaxation-optimized approach (Pervushin et al. 2002), fast pulsing techniques (Kupce and Freeman 2007), combined for “ultrafast” spectroscopy (Schanda et al. 2005; Gal et al. 2007; Mishkovsky et al. 2007).

The other reconstruction methods include the use of G-matrix Fourier transform-NMR (Kim and Szyperski 2003) and J couplings networks (Atreya et al. 2007), covariance spectroscopy of higher dimensions (Zhang and Brüschweiler 2004; Snyder et al. 2007a, b), fast Fourier transforms of non-equispaced data (Marion 2005), 2D Fourier transformations of arbitrarily sampled NMR data sets (Kazimierczuk et al. 2006a, b), filter diagonalization methods (Mandelshtam et al. 1998, 2000; Armstrong et al. 2005), maximum entropy reconstructions (Rovnyak et al. 2004; Frueh et al. 2006), and multidimensional decomposition algorithm (MDD) (Orekhov et al. 2001; Luan et al. 2005). Feedback looping of data acquisition and analysis was conceptualized by introducing a targeted acquisition (TA) scheme for real-time NMR spectroscopy (Jaravine and Orekhov 2006). TA/MDD offers a possibility to combine adaptive, non-regular data sampling in *n*-dimensions with full (or partial) reconstruction of incompletely sampled 3D spectra or hyper-dimensional spectra (Jaravine et al. 2006, 2008), resulting in concurrent data accumulation, processing, and monitoring of spectral quality.

In the current paper, novel methodological development in the TA approach addresses problems of a protein system featuring sample degradation and missing spin systems by combining automated assignment and TA. We for the fist time systematically explore limits of applicability and robustness of such a combination. In the previous works on automatic assignment (Masse and Keller 2005; Masse et al. 2006), TA (Jaravine and Orekhov 2006) and hyper dimensionality (HD) (Jaravine et al. 2008) all studied proteins were chemically homogeneous and not compromised by degradation and presence of many minor peaks, so that TA reached nearly 100% of the expected number of peaks and assignment level (Jaravine et al. 2008), besides they used a peak-list based assignment procedure (Bohm et al. 2005). In the case of chemically degrading proteins the expected scores are significantly lower, so that information in abstracted 1D shapes obtained in HD (Jaravine et al. 2008) might not be sufficient to resolve ensuing ambiguities. In this paper for the first time full 3D spectra reconstructed at each step of TA were automatically interpreted, which might be a more robust approach in difficult cases. This task required developing an entirely new algorithm containing its own logic inference engine and tools for direct spectrum analysis in full 3D. Critical novelty is to use a hypothesis driven recursive construction of spin systems (chemical shifts of connected spins) by direct evaluation of higher order hypotheses (spin system identification or assignment) against spectral intensities as opposed to simple cataloging of spectral maxima for later processing opening a way to dynamically reschedule the data sampling scheme, which was not achieved in our previous work on HD (Jaravine et al. 2008). As a test system, we selected the medium-sized 15 N- and 13C-labeled protein Cytochrom-c maturation protein E, a heme chaperone in its diamagnetic apo form (Enggist et al. 2002) rapidly degrading on time scales needed to perform 3D NMR experiments.

Materials and methods

NMR sample preparation

Apo-CcmE-H6 (residues 30–159) was expressed and purified as described (Enggist et al. 2002). The sample was dialyzed against 300 mM NaCl, 50 mM sodium phosphate, pH 7.2, and was subsequently concentrated to 0.5 mM. After 1-week incubation at 45°C, MALDI-TOF spectra showed essentially complete removal of the structurally flexible C-terminal fragment (residues 131–159) and presence of short peptides as degradation products manifested as additional small cross-peaks in the [1H–15 N] HSQC spectrum. Thus, in the course of time typical for NMR measurements, the protein exhibited (1) small deviation of chemical shifts in apo-CcmE (L30–H130) lacking C-terminus in comparison to the full length construct, (2) broadening beyond detection of a set of resonances located primarily in loop regions, and (3) emergence of additional cross-peaks due to degradation. This sample was judged as a suitable model for automatic resonance assignment of a protein on the background of emerging artifacts (e.g., degradation products and missing cross-peaks) typically encountered in real protein samples.

NMR measurements and MDD reconstruction of 3D spectra

NMR experiments with full length apo-CcmE were performed at 20°C on a Bruker Avance 600 MHz spectrometer equipped with a cryogenic probe, while experiments with apo-CcmE (L30–H130) were performed on a Varian Inova 600 MHz spectrometer using the TA approach featuring incremental non-uniform sampling (INUS) (Jaravine and Orekhov 2006) (see details in Table 1). A quarter (for HNCACB, HN(CA)CO, and ct-HNCA) or 11% (for HNCO and HN(CO)CA) of the points in the regular grid were sampled using the INUS schedule (Jaravine and Orekhov 2006). The time domain data was converted into the *nmrPipe* (Delaglio et al. 1995) format and the directly detected dimension was processed in the conventional way. In the indirectly acquired dimensions (t_1 and t_2), NMR data were processed using R-MDD (Jaravine et al. 2006) as is implemented in the *MDDnmr* (Tugarinov et al. 2005) program and inspected using *CARA* (www.nmr.ch). The ^{1}H chemical shifts were referenced to the DSS (sodium 2,2-dimethyl-2-silapentane-5-sulfonate) signal at 0 ppm and the ^{13}C and ^{15} N chemical shifts were referenced indirectly using the ^{15} N/^{1}H and ^{13}C/^{1}H gyromagnetic ratios.

Psyte and *AutoLink II* analysis

We introduce *Psyte*, a new module of the program *AutoLink II*, to work with the semi-automated resonance assignment program *CARA*. A schematic diagram and description of the algorithm are provided in Supporting Information. In brief, *Psyte* is developed for the recognition of spin systems within multidimensional NMR spectra. The module combines expert knowledge, systematic rules, and competition-based (non-monotonic) decision-making processes in order to group chemical shifts extracted from spectra into spin systems. As with our previous programs [*AutoLink* (Masse and Keller 2005) and *SideLink* (Masse et al. 2006)], *Psyte*'s algorithm is designed to do human-like reasoning in order to achieve its goal, as illustrated in the schematic diagram of the algorithm (Supplementary Figure 11). *Psyte* resolves spectral overlap by deconvolution of peak models derived directly from the NMR spectra, but does not require any well-resolved peak to derive its models as is implemented in *XEASY* (Bartels et al. 1995). Additionally, *Psyte* cross-compares spectra in order to validate decisions made on an individual spectrum. The generated spin system grouping hypotheses are verified by establishing recursion loops to the spectral intensities. *Psyte* successfully removes most spectral artifacts. The program's artifact detectors can be divided into two main categories, specific and non-specific. Specific artifact recognizers are designed to identify artifacts of a known type (e.g., due to truncation of time-domain signal, presence of strong solvent resonance etc.). The non-specific artifact detectors recognize artifacts by comparing the

Table 1 Multidimensional decomposition algorithm reconstruction of a set of 3D spectra with non-uniform sampling of the time-domain signal

Spectrum	Acq. time, h	Spectral width in ω_1(^{13}C), Hz	Spectral width in ω_2(^{15} N), Hz	Number of points in regular grid	
	Total: 40			t_1(^{13}C)	t_2(^{15} N)
HNCO	5 (11%)[a]	2,100	2,500	120	60
HN(CO)CA	5 (11%)	4,527	2,500	90	60
ct-HNCA	10 (25%)	4,527	2,500	120	60
HN(CA)CO	10 (25%)	2,100	2,500	120	60
HNCACB	10 (25%)	12,071	2,500	120	60

[a] In brackets is percent of the full acquisition time needed to sample the complete regular grid

observed peaks against the program's "expert knowledge" of what the spectra are supposed to look like. These non-specific artifact recognizers rate NMR spectra according to *Psyte*'s cognitive understanding of what their data and artifact content are and adjust the spectrum interpretation accordingly.

Due to the fact that the results of its analysis are spin systems rather than purely ungrouped peaks, *Psyte*'s output is served as input for the downstream resonance assignment programs like *AutoLink II* and *SideLink*. For the most difficult cases, *Psyte* has been designed such that it can work iteratively with a user and account for user modifications to the *CARA* repository, so that a spectroscopist can help the program in its analysis if cases are found where the program is prone to error.

Results

Targeted time domain data acquisition and MDD reconstruction

An apo form of medium-sized protein apo-CcmE (L30–H130) (Enggist et al. 2002) was used in course of the analysis. Figure 1 shows a fingerprint 2D [1H–15 N]-projection of 3D HNCO spectrum, where assigned cross-peaks as well as missing and spurious cross-peaks are indicated. Five triple resonance spectra typically employed for backbone resonance assignment were measured using INUS schemes (Jaravine and Orekhov 2006) with maximum numbers of sampled t_1 and t_2 hyper-complex time domain points in 3D spectra reaching 11 or 25% of the regular grid (see Table 1). In order to simulate run-time progression in acquisition of NMR signal, MDD reconstructions were performed using selected quartets of 1D FIDs representing hyper-complex data points in t_1 and t_2 extracted in accordance with an INUS schedule (Jaravine and Orekhov 2006) from the prerecorded 25 and 11% sampled INUS datasets. Therefore, we obtained 16 time-snapshots of the complete dataset available at 1.5, 3, 4.5 h, etc. after the start of acquisition. Figure 2 shows the numbers of cross-peaks identified in the individual MDD-reconstructed spectra in the set as a function of the progressively sampled grid. Direct inspection shows that HNCO and HN(CO)CA experiments have reached the targeted numbers of cross-peaks after sampling of ca. 8% of the grid, and hence, can be dropped from the acquisition schedule, dedicating the spectrometer resources to other more demanding experiments.

At a closer look into the resolution and sensitivity of the resonance peaks in the 13C-dimension of the MDD-reconstructed spectra, resonances of residue E105 exhibiting typical signal amplitudes in HNCACB, HN(CA)CO, and HNCO spectra are shown in Fig. 3. As is expected in MDD reconstruction, the peak position and the line-width remained constant with the progression of data sampling, while gradual gain of spectral sensitivity occurred as the sampling level increased; however, a certain minimal percentage of data sampling is required for a given resonance peak to appear in the spectra.

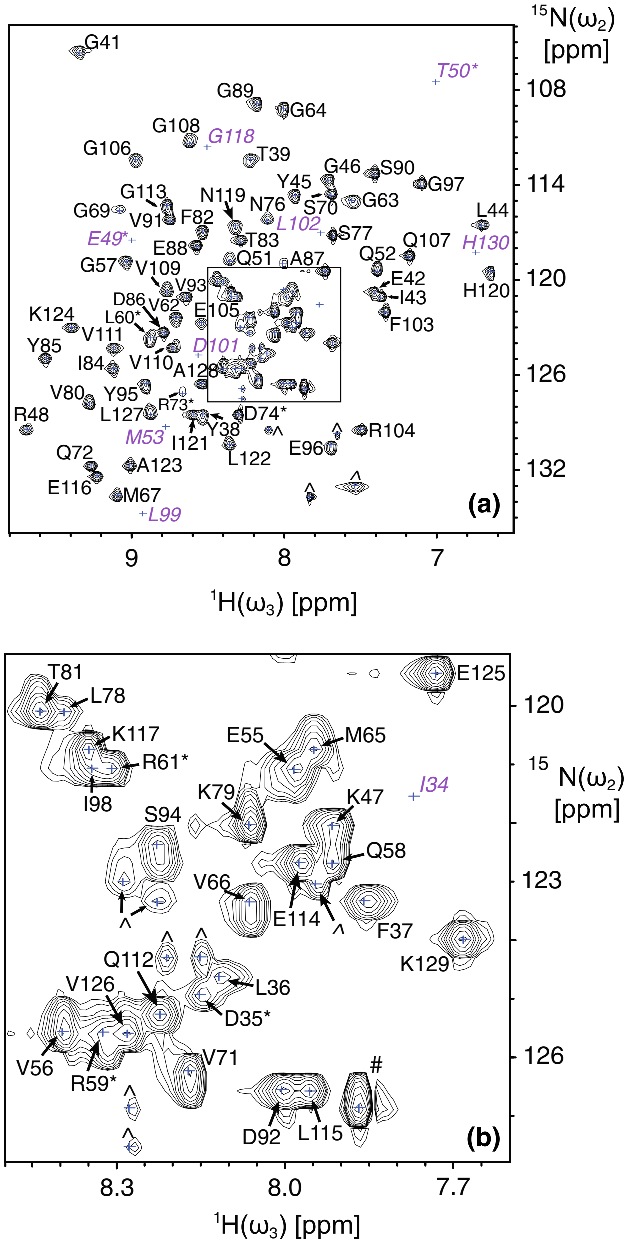

Fig. 1 (**a**) [1H–15 N]-projection of the 3D HNCO spectrum reconstructed from 11% of time domain data (see Table 1). (**b**) Inset of **a**. All cross-peaks are marked by *blue crosses*, while those that are present in the reference HSQC spectrum but are not detected in the MDD-reconstructed spectra are indicated by *magenta italics*. Residues for which ambiguous assignment is found are indicated by *asterisk*. Cross-peaks due to degradation products are marked by "*caret symbol*"

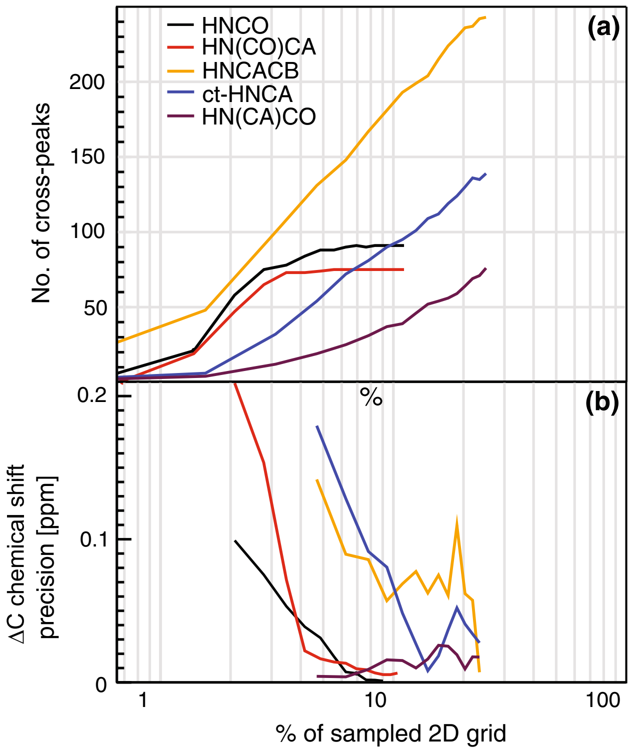

Fig. 2 (**a**) The numbers of cross-peaks identified in the MDD-reconstructed 3D spectra versus fraction of sampled time domain in two indirectly detected dimensions using INUS scheme. In this representation, the 100% corresponds to the number of sampled points spanning the full regular grid (Table 1). (**b**) The precision of identified cross-peak positions versus fraction of sampled time domain

Automatic determination of spin systems

In the first step, *Psyte* groups resonances detected in a set of 3D heteronuclear spectra into spin systems. Due to complexity inherent to multidimensional spectra reconstructed from sparse data, the spin systems obtained do not always correspond exactly to those typically identified by a spectroscopist. In the case of apo-CcmE (L30–H130), some extra spin systems were created from ambiguous regions in the spectra that contain only artifact resonances, and they were generally deficient in the relevant spins. Nonetheless, these spurious spin systems had rarely been assigned to the amino acid sequence by downstream logic in *AutoLink II*, owing to their low link hypothesis scores and competition with link hypotheses involving only real spin systems.

The amount of information directed to *AutoLink II* for residue specific assignment in the form of created spin systems from different combinations of MDD-reconstructed spectra is shown in Fig. 4. Progressively sampled spectra, according to the corresponding level of time domain data sampling shown in Fig. 3, were assembled in 13 sets forming three groups. Inspection of the identified spin systems in the first group (sets 1–6) indicates that for all reasonably high levels of data sampling, the numbers of created spin systems remain approximately constant (between 90 and 110) with the tendency of abstracting more spin systems from better sampled spectra. The fact that more noisy spectra do not result in higher numbers of created spin systems indicates high tolerance of *Psyte* algorithm to the presence of noise and spectral artifacts. It should be noted that at the very low level of spectral sampling represented in set 6 (i.e., 3.1% for HNCACB, HN(CA)CO, and ct-HNCA; 1.4% for HNCO and HN(CO)CA), significantly higher numbers of spin systems

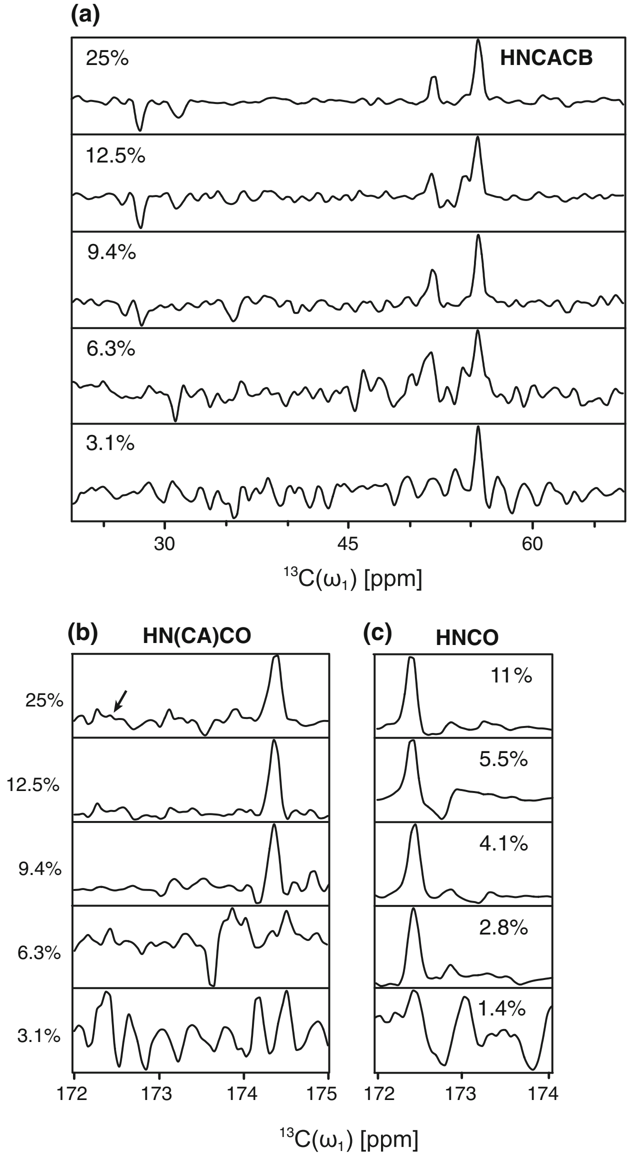

Fig. 3 1D traces along the 13C-dimension of (**a**) HNCACB, (**b**) HN(CA)CO, and (**c**) HNCO spectra taken at δ^{15} N = 122.69 ppm and δ^{1}H = 8.54 ppm (residue E105) at variable percentage of sampled time domain relative to the full grid. The *arrow* in (**b**) indicates the position of inter-residual cross-peak still absent in HN(CA)CO spectrum sampled at 25% of the full grid

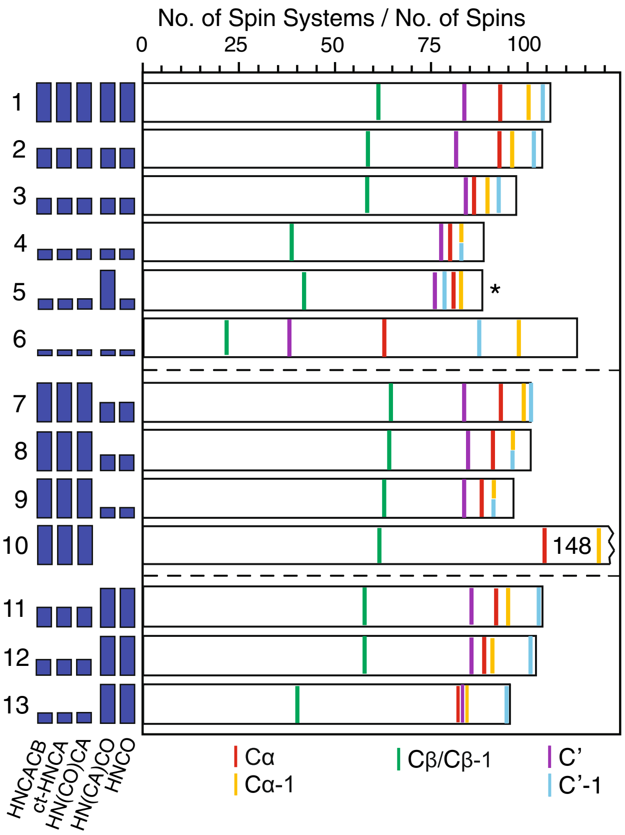

Fig. 4 The numbers of spin systems (*horizontal bars*) and the numbers of the respective spins (*colored lines*) identified in 13 sets of MDD-reconstructed 3D spectra. The percentage of data sampling of each individual spectrum in the set is indicated by *vertical full bars*. In set 1, the *full bars* correspond to HNCACB (25%), ct-HNCA (25%), HN(CO)CA (11%), HN(CA)CO (25%), and HNCO (11%). In set 10, 148 spins systems are determined. Set 5 marked by *asterisk* contains HNCO sampled at 2.8% and HN(CA)CO sampled at 25% of the full grid

can be picked with reduced amount of Cα/Cα-1, and C′ spins. We identified the generally reduced S/N ratio in HN(CA)CO spectrum as the cause for this behavior of the algorithm. A selective increase in the sampling level of only this spectrum (e.g., in set 5) drives the numbers of identified spins and spin systems closer to optimum found in sets 1–3. Therefore, besides reaching a minimal sampling level for all spectra (ca. above 6%), some individual spectra with low sensitivity must be preferentially sampled for larger proportion of time domain data, in order to remove or alleviate spin system abstraction "bottlenecks." Based on this observation, a readjustment of the general sampling scheme can be performed (see Sect. "Discussion").

Residue specific resonances assignment

Residue specific assignment was achieved by *AutoLink II* using the amino acid sequence of apo-CcmE (L30–H130) and the spin systems identified by *Psyte* as the input to the automatic assignment module, as described earlier (Masse and Keller 2005). Due to the critical dependence of the spin system identification algorithm implemented in *Psyte* on variation of the local spectral noise (appearing in clusters and bands in MDD reconstructions, see Supplementary Figure 7) as well as the order of generated hypothesis on the spin system groupings, slightly different lists of spin systems were generated even for small variation of control parameters. This variability of spin system determination is subsequently propagated to variations in residue specific assignment (Fig. 5), which are typically found for residues with low spectral S/N ratio such as residues L36 and R73. For instance, different chemical shifts were assigned to the same spin, to variable extent between different spins. Therefore, several lists of spin systems were generated using the same set of spectra followed by the residue specific assignment attempt. In Fig. 5, two of such assignment attempts designated as SL1 and SL2 are shown. The discrepancies in the assignment between the two (or more) spin systems lists helped to isolate problematic residues and regions in the spectra, thus allowing us to either spot the wrong assignments (e.g., R59), or assign more residues after manually inspecting the relevant spin systems and cross-peaks in the spectra (e.g., D35).

AutoLink II provides a graphical output of assigned residues as well as statistics associated with the quality of assignment as described previously (Masse and Keller 2005). Figure 5 shows the sequence fit score of the automatically achieved assignment. Almost all assigned chemical shifts score higher than 0.80 due to generally good spectral S/N ratio. Even though the wrong assignments in this instance could be identified by their low sequence fit scores, such measure was not employed as much more ambiguity arose for the sets of spectra with lower levels of data sampling. Apparently, the chemical shift of Cα-1 of G108 assigned by *Psyte* in SL1 was slightly off-center, and hence, it did not match well with that of Cα of Q107. Nonetheless, the assignment was judged correctly by *AutoLink II* as it did not violate the overall assignment result. Similarly, owing to inaccurate chemical shifts, R59 was wrongly assigned to a spin system containing the cross-peak marked by "#" in Fig. 1. In general, assignment of spectral features due to MDD reconstruction artifacts to the amino acid sequence was never observed. Furthermore, resonances due to degradation products (Fig. 1) were also not assigned by *AutoLink II*. Although the assignment of D35, R73, and D74 was problematic as the resonances of L36 and R73 had been broadened to great extent (Fig. 1), yet they were correctly assigned.

The capability of *AutoLink II* to assign R73 and D74 from set 1 illustrates the robustness of the algorithm as well as the implementation of TA on multiple spectra, in dealing with seriously broadened resonances and degenerated

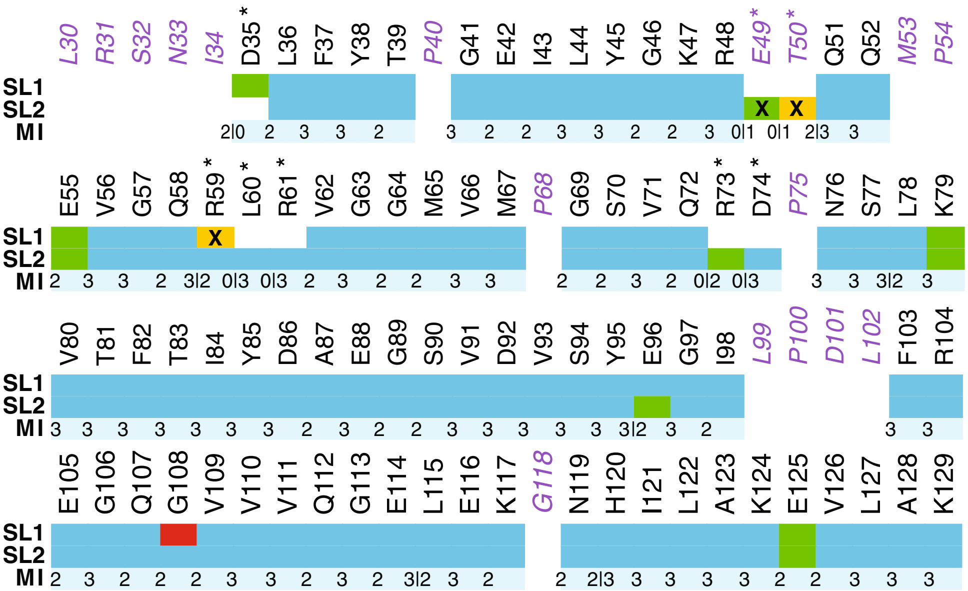

Fig. 5 Overview of residue specific assignment of backbone resonances based on two spin systems lists (SL) determined in separate executions of *Psyte* using 25 and 11% INUS sampled spectra of Table 1. The *AutoLink* sequence fit score (adapted from Masse and Keller 2005) of individual residue is represented by the following color scheme: *blue* (0.81–1.00), *green* (0.71–0.80), *yellow* (0.61–0.70), and *red* (0.60 and below). Cross (*X*) represents wrong assignment as compared to the reference (manual) assignment (adapted from Enggist et al. 2002). The *asterisk* indicates discrepancy in resonance assignment between SL1 and SL2 (e.g., R59 and D35). After manual inspection of identified spin systems in *Psyte* output, these discrepancies can be resolved resulting in the final accurate assignment colored turquoise. The numbers of inter-residue matches of Cα-1/Cα, Cβ-1/Cβ, and C′ − 1/C′ chemical shifts are listed below the amino acid sequence representing the availability of connectivity information within the assigned spin systems in SL1 and SL2. Residues in *magenta italics* are designated as "not-assignable", due to fact that they are either proline residues or the corresponding cross-peaks are broadened beyond detection in all MDD-reconstructed 3D spectra

chemical shifts. The identity of the degenerated Cα and Cα − 1 resonances of R73 could be deduced from the HNCACB and HN(CO)CA spectra, even though the resonances had disappeared completely in the ct-HNCA spectrum (Supplementary Figure 8a, b). Similarly, the C′ and C′ − 1 resonances of R73 were completely lost in the HN(CA)CO spectrum (Supplementary Figure 8c). However, the assignment of both R73 and D74 was still possible in SL2 even without the Cα connectivity and the C′ chemical shift of R73, as it did not violate the overall assignment result.

Discussion

We developed a method of automatic resonance assignment process with our new software *Psyte/AutoLink II* to handle specific problems arising in real-time automatic analysis of NMR data (Fiorito et al. 2006). The complexity of the problem has been effectively reduced by the introduction of two levels of analysis. First, *Psyte* sorts out variations of chemical shifts among the 3D spectra and determines a single representative spin system list. At the second level, *Autolink II* attempts assignment of backbone resonances simultaneously identifying spin systems that may arise from artifacts and degradation products. In order to systematically explore limits of applicability and robustness of such a combination as well as for the purpose of establishing an optimal schedule of data acquisition, this method was applied to the apo form of medium-sized protein apo-CcmE (L30–H130), of which exhaustively analyzed multidimensional NMR spectra and 3D structure are available (Enggist et al. 2002). We find this protein a relevant model since the NMR spectra of this protein, in addition to a set of resonances stemming from the structurally defined core, feature signals from identifiable degradation products as well as resonances broadened by conformational exchange located in loops and flanking regions. Overall, 83 ^{1}H–^{15}N cross-peaks were identified as assignable (among which five exhibiting significantly reduced intensity), 11 were assigned to polypeptidic degradation products showing reduced intensity in all connected spin systems and 13 backbone resonances were broadened beyond detection (Fig. 1). We demonstrated that in the absence of human intervention and with the use of the optimal spectrum sampling scheme, 95% of the assignable resonances can be stably assigned.

Previously the number of cross-peaks picked from reconstructed spectra was set as target for data acquisition (Jaravine and Orekhov 2006). This acquisition termination

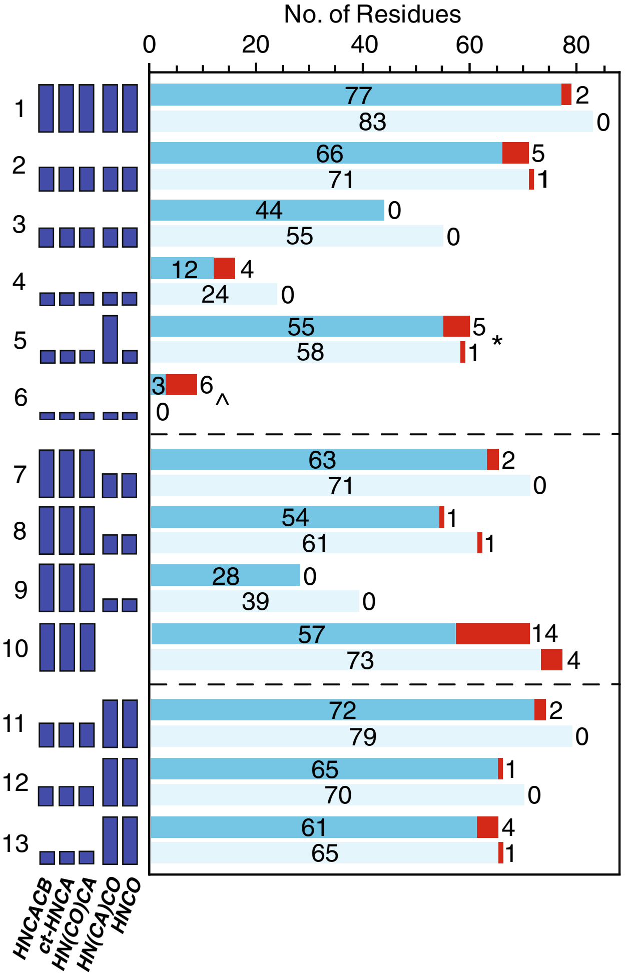

Fig. 6 The numbers of consistently assigned residues (*blue*), the maximum numbers of wrong assignments (*red*), and the total numbers of assigned residues after manual inspection of identified spin systems in *Psyte* output (*turquoise*) achieved in 13 sets of MDD-reconstructed spectra at various levels of data sampling (as in Fig. 4). Set 5 marked by *asterisk* contains HNCO spectrum sampled at 2.8% and HN(CA)CO spectrum sampled at 25% of the full grid. For set 6 marked by "*caret symbol*", no correct assignment is achieved

criterion is shown to be sufficient for chemically stable proteins with homogenous distribution of cross-peak intensities throughout the amino acid sequence. However, the application of a spectrum-wide global noise threshold (as it is implemented in most standard peak-pickers) on spectra from degrading proteins might result in missing assignments, especially at lower sampling levels. This is due to the presence of low intensity resonances stemming from dynamic regions of protein, which could be missed by threshold filtering, as well as the presence of low molecular weight degradation products giving rise to spurious cross-peaks. In addition, the noise in the MDD-reconstructed spectra is not anymore uniformly distributed throughout the spectrum, but rather appears in bands centered at ^{1}H and ^{15} N resonances along indirectly detected dimensions (Supplementary Figure 7). On this basis, an appropriate estimate of the spectrum-wide threshold value becomes difficult to obtain (see Supplementary Figure 9).

The critical novelty in our current approach is to replace linear threshold-based peak picking with hypothesis-driven recursive construction of spin systems. We note that the use of previous implementations of *AutoLink* required manual grouping of peaks from several peak-lists into a single list of spin systems, which is deemed impossible in a real-time data acquisition. In essence, reliable and complete identification of cross-peaks requires a priori knowledge of expected signals together with extensive noise filtering procedures implemented in the module *Psyte*. This observation led us to conclude that simple abstraction of complex spectrum to a list of cross-peaks might suffer from inherent inability to correctly decompose peak clusters in the absence of higher order analysis information (e.g., recursion from assignment/structure levels) (Supplementary Figure 10).

Feasibility of real-time TA scheme applicable to proteins prone to degradation was tested off-line by attempting automatic resonance assignment using pre-recorded and MDD-reconstructed spectra at various percentages of sampled time domain grid (sets 1–13 in Figs. 4, 6). This method was used to explore robustness of the process (Bootstrap approach) and to define how much data is needed to reach the target as it allows to consider many scenarios, which otherwise would require real-time recording of all of them, which is technically hard taking into account fast degradation of the protein sample. After a suitable schedule of acquisition and assignment has been established, we simulated real-time coupling. The results show that the real-time control is computationally feasible with metric time needed to reconstruct the corresponding 3D spectra using a Linux station equipped with four CPUs always not exceeding about half of NMR spectrometer acquisition time needed to partially and simultaneously sample the five 3D spectra. Every assignment attempt takes between 1 and 2 h depending on the convergence properties of the current problem, and typically runs in parallel with MDD reconstructions, each on a separate CPU node.

Contribution of individual spectrum to the overall assignment result was investigated by fixing the level of data sampling for HNCACB, ct-HNCA, and HN(CO)CA spectra, while reducing that for HNCO and HN(CA)CO spectra progressively, and vice versa. The group with fixed levels of data sampling for HNCO and HN(CA)CO spectra clearly outperformed the other group (i.e., set 7, 8, and 9). This indicates that the percentage of data from which HNCO and HN(CA)CO spectra were reconstructed could be restricting the completeness of assignment, and hence, representing the assignment "bottleneck." In fact, the most significant limiting factor in our results was the quality of the HN(CA)CO spectrum involved – a case which can be justified by comparing set 4 and 5 (Fig. 6). On the other hand, the low sensitivity of ct-HNCA spectrum can be complemented by connectivities supplied by HNCACB spectrum.

Conclusion

Here we tested computational feasibility of a real time assignment of backbone resonances of a protein prone to degradation in a time scale of 3D NMR experiments. We propose a possibility of real-time TA with automated resonance assignment devoid of spectroscopist intervention (or at best very minor) set as the target based on the off-line analysis of prerecorded data. This method is used to explore robustness of the process (a bootstrap approach) and to define how much data is needed to reach the target as it allows considering many scenarios. The approach of automatic resonance assignment based on fast nonlinearly sampled and MDD-reconstructed set of simultaneously acquired spectra might represent a viable strategy in structural NMR studies of rapidly degrading biological molecules. Our method provides a flexible criterion to optimally allocate NMR resources, completeness and accuracy of achievable assignment as well as the level of protein degradation. Since degradation of protein material might limit the propagation of spectral S/N ratio with the overall NMR time invested, devising optimal strategies to guide acquisition represents a new and important avenue in the field of automation of NMR structure determination.

The setup of INUS sampling of 3D triple resonance spectra is available on a web-based platform for dynamic generation and share of NMR experiments at http://www.trosy.com/nex. The program *AutoLink II* is accessible at http://www.autolink.nmr-software.org. The program *MDDnmr* is available by a direct request to the authors (Vladislav Orekhov. and Victor Jaravine).

Acknowledgement This work was funded by the Academic Research Fund Tier-2 grant of Ministry of Education of Singapore to Konstantin Pervushin and Bruker BioSpin Group AG, Germany, grant to James E. Masse.

References

Armstrong GS, Mandelshtam VA, Shaka AJ, Bendiak B (2005) Rapid high-resolution four-dimensional NMR spectroscopy using the filter diagonalization method and its advantages for detailed structural elucidation of oligosaccharides. J Magn Reson 173:160–168

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

Atreya HS, Chary KVR, Govil G (2002) Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. Curr Sci 83:1372–1376

Atreya HS, Garcia E, Shen Y, Szyperski T (2007) J-GFT NMR for precise measurement of mutually correlated nuclear spin-spin couplings. J Am Chem Soc 129:680–692

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K (1995) The Program Xeasy for computer-supported nmr spectral-analysis of biological macromolecules. J Biomol NMR 6:1–10

Bohm M, Stadlthanner K, Tome AM, Gruber P, Teixeira AR, Theis FJ, Puntonet CG, Lang EW (2005) AutoAssign—an automatic assignment tool for independent components. In: Proceedings of pattern recognition and image analysis, Pt. 2, vol. 3523. Springer, Berlin, pp 75–82

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

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

Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005a) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127:12528–12536

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

Enggist E, Thony-Meyer L, Güntert P, Pervushin K (2002) NMR structure of the heme chaperone CcmE reveals a novel functional motif. Structure 10:1551–1557

Fiorito F, Hiller S, Wider G, Wüthrich K (2006) Automated resonance assignment of proteins: 6D APSY-NMR. J Biomol NMR 35:27–37

Frueh DP, Sun ZY, Vosburg DA, Walsh CT, Hoch JC, Wagner G (2006) Non-uniformly sampled double-TROSY hNcaNH experiments for NMR sequential assignments of large proteins. J Am Chem Soc 128:5757–5763

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

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

Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinas M (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins 61:36–43

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

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

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

Hiller S, Wasmer C, Wider G, Wüthrich K (2007) Sequence-specific resonance assignment of soluble nonglobular proteins by 7D APSY-NMR spectroscopy. J Am Chem Soc 129:10823–10828

Hitchens TK, 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

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

Jaravine VA, 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 (2006a) Two-dimensional Fourier transform of arbitrarily sampled NMR data sets. J Magn Reson 179:323–328

Kazimierczuk K, Zawadzka A, Kozminski W, Zhukov I (2006b) Random sampling of evolution time space and Fourier transform processing. J Biomol NMR 36:157–168

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 (2004) Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. J Am Chem Soc 126:6429–6440

Kupce E, Freeman R (2007) Fast multidimensional NMR by polarization sharing. Magn Reson Chem 45:2–4

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

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

Luan T, Jaravine V, Yee A, Arrowsmith CH, Orekhov VY (2005) Optimization of resolution and sensitivity of 4D NOESY using multi-dimensional decomposition. J Biomol NMR 33:1–14

Mandelshtam VA (2000) The multidimensional filter diagonalization method. J Magn Reson 144:343–356

Mandelshtam VA, Taylor HS, Shaka AJ (1998) Application of the filter diagonalization method to one- and two-dimensional NMR spectra. J Magn Reson 133:304–312

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

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

Mishkovsky M, Kupce E, Frydman L (2007) Ultrafast-based projection-reconstruction three-dimensional nuclear magnetic resonance spectroscopy. J Chem Phys 127:034507

Nilges M, Macias MJ, Odonoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 269:408–422

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

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

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

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, Kupce E, Brutscher B (2005) SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. J Biomol NMR 33:199–211

Snyder DA, Xu Y, Yang D, Brüschweiler R (2007a) Resolution-enhanced 4D 15 N/13C NOESY protein NMR spectroscopy by application of the covariance transform. J Am Chem Soc 129:14126–14127

Snyder DA, Zhang F, Brüschweiler R (2007b) Covariance NMR in higher dimensions: application to 4D NOESY spectroscopy of proteins. J Biomol NMR 39:165–175

Takeda M, Ikeya T, Güntert P, Kainosho M (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. Nat Protocol 2:2896–2902

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

Tugarinov V, Kay LE, Ibraghimov I, Orekhov VY (2005) High-resolution four-dimensional H-1-C-13 NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. J Am Chem Soc 127:2767–2775

Wu KP, Chang JM, Chen JB, Chang CF, Wu WJ, Huang TH, Sung TY, Hsu WL (2005) RIBRA—an error-tolerant algorithm for the NMR backbone assignment problem. In: Proceedings of research in computational molecular biology, vol. 3500. Springer, Berlin, pp 103–117

Zhang F, Brüschweiler R (2004) Indirect covariance NMR spectroscopy. J Am Chem Soc 126:13180–13181