ARTICLE

Automated solid-state NMR resonance assignment of protein microcrystals and amyloids

Elena Schmidt · Julia Gath · Birgit Habenstein · Francesco Ravotti · Kathrin Székely · Matthias Huber · Lena Buchner · Anja Böckmann · Beat H. Meier · Peter Güntert

Received: 5 March 2013 / Accepted: 6 May 2013 / Published online: 21 May 2013

Abstract Solid-state NMR is an emerging structure determination technique for crystalline and non-crystalline protein assemblies, e.g., amyloids. Resonance assignment constitutes the first and often very time-consuming step to a structure. We present ssFLYA, a generally applicable algorithm for automatic assignment of protein solid-state NMR spectra. Application to microcrystals of ubiquitin and the Ure2 prion C-terminal domain, as well as amyloids of HET-s(218–289) and α-synuclein yielded 88–97 % correctness for the backbone and side-chain assignments that are classified as self-consistent by the algorithm, and 77–90 % correctness if also assignments classified as tentative by the algorithm are included.

Keywords Automated assignment · Sequence-specific assignment · Amyloid · CYANA · FLYA

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9742-x) contains supplementary material, which is available to authorized users.

E. Schmidt · L. Buchner · P. Güntert (✉)
Center for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Goethe University Frankfurt am Main, Frankfurt am Main, Germany
e-mail: guentert@em.uni-frankfurt.de

E. Schmidt · L. Buchner · P. Güntert
Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany

J. Gath · F. Ravotti · K. Székely · M. Huber · B. H. Meier (✉)
Physical Chemistry, ETH Zurich, Wolfgang-Pauli-Strasse 10, 8093 Zurich, Switzerland
e-mail: beme@ethz.ch

B. Habenstein · A. Böckmann (✉)
Institut de Biologie et Chimie des Protéines, UMR 5086 CNRS/Université de Lyon 1, 7 passage du Vercors, 69367 Lyon, France
e-mail: a.bockmann@ibcp.fr

P. Güntert
Graduate School of Science and Engineering, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji, Tokyo 192-0397, Japan

Introduction

Sequence-specific resonance assignments are a prerequisite for protein structure determination and the study of protein interactions and dynamics. The manual or semi-automated determination of resonance assignments is cumbersome but remains the standard in solution NMR, and the virtually exclusive approach in solid-state NMR. This unsatisfactory situation is due to features or imperfections of experimental NMR spectra, which may be even more pronounced in solid-state spectra, i.e., linewidth, signal overlap, low signal-to-noise ratio, and spectral artifacts.

The automated assignment of solid-state NMR spectra is still very challenging. Assignments are typically done on the basis of ^{13}C and ^{15}N resonances only. In addition, the resonance lines can be broader than in solution. Also, different rules must be applied for the analysis of the spectra because most solid-state NMR experiments transfer polarization through space rather than through covalent bonds, which leads to more complex peak patterns that can produce additional peak overlap, but also contain valuable information.

Many algorithms have been developed for automated resonance assignment in solution NMR (Guerry and Herrmann 2011). However, many approaches have limitations such as assigning only backbone and C$^{\beta}$ resonances, or need additional input information, e.g., collecting signals in spin systems, which requires an analysis of the spectra besides peak picking. Only very few automated algorithms have been reported for solid-state NMR resonance assignment. A first algorithm (Moseley et al. 2010) developed on the basis of the AutoAssign (Zimmerman et al. 1997) package for solution NMR has been applied to assign the backbone resonances of GB1 using peak lists from 3D NCACX, CAN(CO)CA, and 4D CANCOCX experiments as input. A second approach (Hu et al. 2011; Tycko and Hu 2010) can in principle assign backbone and side-chain signals by analyzing arbitrary combinations of spectra with arbitrary dimensions. In the examples shown its input consists of lists of signals from NCACX, NCOCA, and CONCA spectra. The signals of a residue must be grouped together and assigned to atom types (e.g., N, C^{α}, C^{β}, C^{γ}), and possible residue-type assignments must be specified before running the algorithm.

Recently, we introduced the FLYA automated resonance assignment algorithm for solution NMR spectra and showed that it is more general and yields more accurate results than other automated assignment methods for all chemical shifts (Schmidt and Güntert 2012). Here, we present the ssFLYA resonance assignment algorithm for solid-state NMR data, apply it to four proteins for which resonance assignments based on essentially the same spectra have been obtained earlier, and evaluate its performance with peak lists obtained from automated or manual peak picking in the experimental spectra.

Materials and methods

The ssFLYA algorithm

NMR resonance assignment is based on experiments that correlate nuclear spins such that they give rise to cross peaks in multidimensional spectra. Assignment experiments are chosen to complement each other in such a way that the connectivity of the atoms in a protein can be represented by a network of peaks that are expected to be observed. Mapping this network of expected peaks with unknown positions to the unassigned measured peaks with known positions provides an assignment of the frequencies to the spins (Bartels et al. 1996, 1997). The ssFLYA algorithm for automated backbone and side-chain resonance assignment uses this general approach to assign solid-state NMR spectra. It is based on the recently introduced FLYA automated resonance assignment algorithm for solution NMR (Schmidt and Güntert 2012), implemented in the software package CYANA (Güntert 2009; Güntert et al. 1997). As input, ssFLYA uses exclusively the sequence of the protein and unassigned peak lists from any combination of multidimensional solid-state NMR spectra.

All experimental data is used simultaneously in order to exploit optimally the redundancy present in the input peak lists and to avoid potential pitfalls of assignment strategies in which results obtained in a given step remain fixed input data for subsequent steps. Instead of prescribing a specific assignment strategy, the ssFLYA resonance assignment algorithm generates the peaks expected in a given spectrum by applying a set of rules for through-bond or through-space polarization transfer, and determines the resonance assignment by constructing an optimal mapping between the expected peaks, assigned by definition but having unknown positions, and the measured peaks, initially unassigned but with known positions in the spectrum (Bartels et al. 1996, 1997; Schmidt and Güntert 2012; Schmucki et al. 2009).

The main difference to solution NMR lies in the rules for generating expected peaks, which have been implemented for many different solid-state NMR experiments (Table 1). Expected peaks for experiments like DARR, which give signals between atoms that are close in space, are obtained using random structures of the respective proteins. An expected peak is generated for each atom pair up to a given cutoff on the maximal distance between the two atoms in the ensemble of random structures. This will generate expected peaks only if the atoms are close together in the primary structure, e.g., for intraresidual and sequential distances. It corresponds to the generation of expected peaks for NOE-based experiments in solution NMR (Schmidt and Güntert 2012). Expected peaks for all other experiments are obtained based on the covalent connections between atoms. For each experiment the covalent bond patterns that hold this information are provided to the algorithm in the CYANA library file. It is straightforward to add new experiments or to modify the rules for existing experiments. Since most of the solid state NMR experiments that are based on covalent bond patterns include a relatively unspecific ^{13}C–^{13}C transfer, some peak lists include additional signals resulting from neighboring carbons. The probability to observe these signals is highest for directly bound neighbors. This effect has been taken into account by adding covalent bond patterns with lower observation probability in cases in which additional signals are expected. On the other hand, in the CCC experiment, which is in general a combination of DARR/PDSD and DREAM transfers, a large number of combinations of carbon atoms not only within an amino acid but, due to the DARR step, also to spatially adjacent amino acid may theoretically give rise to a cross peak. In practice only the more intense ones can be observed. To avoid generating too many expected CCC peaks, they are defined by polarization transfer rules (instead of short distances in random structures as for DARR), which are restricted to generate only the most probably observed intraresidue peaks.

The best mapping of expected peaks to measured peaks is obtained using an evolutionary optimization routine that

Table 1 Polarization transfer pathways used for the generation of expected peaks

```
SPECTRUM NCACB N CA C
 0.98  N:N_AMI CA:C_ALI C_BYL C_ALI C:C_*
 0.30  N:N_AMI CA:C_ALI C_BYL C_ALI C_ALI C:C_*

SPECTRUM NCACBCX N CB C
 0.98  N:N_AMI C_ALI C_BYL C_ALI CB:C_ALI C:C_*
 0.30  N:N_AMI C_ALI C_BYL C_ALI CB:C_ALI C_ALI C:C_*

SPECTRUM NCOCACB C1 C2 N
 0.98  C1:C_ALI C2:C_ALI C_BYL N_AMI C_ALI N:N_AMI
 0.50  C1:C_ALI C2:C_ALI C_ALI C_BYL N_AMI C_ALI N:N_AMI

SPECTRUM CANCOCA CA N C
 0.98  CA:C_ALI N:N_AMI C_BYL C:C_ALI
 0.30  CA:C_ALI N:N_AMI C:C_BYL
 0.30  CA:C_ALI N:N_AMI C_BYL C_ALI C:C_ALI
 0.10  CA:C_ALI N:N_AMI C_BYL C_ALI C_ALI C:C_ALI

SPECTRUM CANCO C N CA
 0.98  C:C_BYL N:N_AMI C_ALI C_BYL CA:C_ALI

SPECTRUM NCACO N CA C
 0.98  N:N_AMI CA:C_ALI C:C_*
 0.30  N:N_AMI CA:C_ALI C_ALI C:C_*

SPECTRUM CCC C1 C2 C3
 0.80  C1:C_BYL C3:C_ALI C2:C_ALI
 0.80  C1:C_BYL C2:C_ALI C3:C_ALI
 0.70  C_BYL C3:C_ALI C2:C_ALI C1:C_ALI
 0.70  C_BYL C2:C_ALI C3:C_ALI C1:C_ALI
 0.60  C_BYL C3:C_ALI C2:C_ALI C_ALI C1:C_ALI
 0.60  C_BYL C2:C_ALI C3:C_ALI C_ALI C1:C_ALI

SPECTRUM NCACX N CA C
 0.98  N:N_AMI CA:C_ALI C_BYL C_ALI C:C_ALI
 0.80  N:N_AMI CA:C_ALI C_BYL C_ALI C_ALI C:C_ALI
 0.60  N:N_AMI CA:C_ALI C_BYL C_ALI C_ALI C_ALI C:C_ALI
 0.30  N:N_AMI CA:C_ALI C_BYL C_ALI C_ALI C_ALI C_ALI C:C_ALI
 0.98  N:N_AMI CA:C_ALI C:C_BYL
 0.80  N:N_AMI CA:C_ALI C_ALI C:C_*
 0.60  N:N_AMI CA:C_ALI C_ALI C_ALI C:C_*
 0.30  N:N_AMI CA:C_ALI C_ALI C_ALI C_ALI C:C_*

SPECTRUM NCOCA N CO C
 0.98  N:N_AMI CO:C_BYL C_ALI N_AMI C_ALI C_BYL C:C_ALI
 0.80  N:N_AMI CO:C_BYL C_ALI N_AMI C_ALI C_BYL C_ALI C:C_ALI
 0.60  N:N_AMI CO:C_BYL C_ALI N_AMI C_ALI C_BYL C_ALI C_ALI C:C_ALI

SPECTRUM NCOCX N CO C
 0.98  N:N_AMI CO:C_BYL C:C_ALI
 0.80  N:N_AMI CO:C_BYL C_ALI C:C_ALI
 0.60  N:N_AMI CO:C_BYL C_ALI C_ALI C:C_ALI
 0.30  N:N_AMI CO:C_BYL C_ALI C_ALI C_ALI C:C_ALI

SPECTRUM NCO N C
 0.98  N:N_AMI C:C_BYL

SPECTRUM NCA N C
 0.98  N:N_AMI C:C_ALI
```

For each spectrum, the first line gives the spectrum name and the atom labels that will be used to identify the respective columns in the peaks lists. The number of atom labels defines the dimensionality of the spectrum. Each of the following lines specifies a (formal) polarization transfer pathway, characterized by the probability of the resulting expected peak followed by a sequence of atom types (N_AMI, amide nitrogen, C_ALI, aliphatic carbon, C_BYL, carbonyl carbon, etc., as used in the CYANA residue library; '*' matches anything) that defines a molecular pattern of atoms linked by direct covalent bonds. In each pathway the atoms whose shifts will determine the position of the resulting peak are identified by their corresponding atom labels, followed by ':'

works with a population of individuals, each representing an assignment solution for the protein. This evolutionary optimization is complemented by local optimization. Solutions that are produced during the optimization are generated such that the search space of an expected peak for a mapping is defined by a chemical shift statistics [by default from the BMRB (Ulrich et al. 2008), or user defined], the deviations of the measured frequencies of measured peaks that are assigned to the same atom remain within a given tolerance, and an expected peak can be mapped to only one measured peak. The first generation of solutions is generated randomly, but subject to these conditions. In each generation a local optimization algorithm takes small parts of a mapping back and reassigns the expected peaks for a defined number of iterations, 15,000 is default. Afterwards the different solutions of one generation are recombined into a new generation. The individuals and the specific parts of an individual that contribute to a new individual are selected via a scoring function. The solution that maximizes this function is given as the final assignment at the end of the calculation.

The scoring and optimization of assignments are performed in ssFLYA as described for solution NMR data (Schmidt and Güntert 2012). The global score for complete assignment solutions evaluates four attributes of an assignment solution, the distribution of chemical shift values with respect to the given shift statistics, the alignment of peaks assigned to the same atom, the completeness of the assignment, and a penalty for chemical shift degeneracy. The global score G is defined by

$$G = \frac{\sum_{a \epsilon A} \left[w_1(a) Q_1(a) + \sum_{n \epsilon N'_a} w_2(a,n) Q_2(a,n)/b(n) \right]}{\sum_{a \epsilon A_0} \left[w_1(a) + \sum_{n \epsilon N_a} w_2(a,n) \right]}.$$

A_0 denotes the set of all atoms for which expected peaks exist, $A \subseteq A_0$ the set of assigned atoms, N_a the set of expected peaks for atom a, and $N'_a \subseteq N_a$ the subset of expected peaks that are mapped to a measured peak. $b(n)$ refers to the ambiguity of the assignment and equals the number of expected peaks that are assigned to the same measured peak as expected peak n. Unassigned atoms and unmapped peaks contribute through the normalization by the denominator. The weighting factors were set to $w_1(a) = 4$ and $w_2(a,\ n) = 1$ for all calculations in this paper. The quality measure $Q_1(a)$ represents the agreement of the average chemical shift $\bar{\omega}(a)$ in the chemical shift list of atom a with the corresponding general chemical shift statistics. Similarly, $Q_2(a,\ n)$ measures the agreement between the chemical shift of atom a obtained from the measured peak to which the expected peak n is mapped and the average frequency of the atom in the assigned peaks of the corresponding spectrum (Schmidt and Güntert 2012). The quality measures Q are designed such that a perfect match corresponds to $Q = 1$, $Q < 1$ in all other cases, a deviation that is considered "as bad as no assignment" yields $Q = 0$, and an infinitely large deviation $Q = -\infty$. Consequently, the global score G is normalized such that $G = 1$ for a (hypothetical) perfect assignment, and $G < 1$ in all other cases.

The main difference to solution NMR lies in the rules for generating expected peaks, which have been implemented for many different solid-state NMR experiments (Table 1; see above).

To improve and assess the accuracy of the assignment, m independent runs of the algorithm, 20 for all calculations in this paper, are performed with different random seeds. For each atom a consensus chemical shift is computed from the values obtained in the individual runs (López-Méndez and Güntert 2006; Malmodin et al. 2003; Schmidt and Güntert 2012). The consensus chemical shift $\tilde{\omega}(a)$ for an atom a is the value that maximizes the function

$$\mu(\omega) = \frac{1}{m} \sum_{j=1}^{m} \exp\left(-\frac{1}{2} \left(\frac{\omega - \bar{\omega}_j(a)}{\varepsilon(a)} \right)^2 \right),$$

where $\bar{\omega}_j(a)$ is the chemical shift value obtained for atom a in run j, and $\varepsilon(a)$ is the chemical shift tolerance, which was set to 0.55 ppm for all calculations in this paper. The maximum value of this function, $\mu(\tilde{\omega}(a))$, is a measure of the self-consistency of the chemical shift values obtained in the individual runs of the algorithm, since it approximately equals the fraction of runs that yielded a chemical shift value within the tolerance $\varepsilon(a)$ from the consensus value $\tilde{\omega}(a)$. This quantity can be calculated without knowledge of reference assignments. If all chemical shift values are identical, then $\mu(\tilde{\omega}(a)) = 1$. In this paper we consider assignments with $\mu(\tilde{\omega}(a)) \geq 0.8$ as "strong" or self-consistent, all others as "weak". Weak assignments should be considered as tentative, although they are correct in many cases.

Proteins and experimental data

Automated chemical shift assignment was performed with solid-state NMR data sets of four proteins for which the assignments had been determined earlier by conventional techniques, i.e., microcrystalline ubiquitin, the C-terminal domain of the Ure2 prion (Ure2p) (Habenstein et al. 2011), and two proteins that form amyloid fibrils, HET-s(218–289) (Siemer et al. 2006; Wasmer et al. 2008) and α-synuclein (Gath et al. 2012). Flexible termini and His-tags were omitted from the input sequences if they could not be observed in the solid-state NMR spectra. In detail, calculations were performed on the following sequences. Ubiquitin: 76 residues; calculations performed for residues 1–70, excluding 6 invisible C-terminal residues. HET-s(218–289) (Siemer et al. 2006; Wasmer et al. 2008): 72 residues; calculations performed for residues 222–289, omitting 4 N-terminal residues. α-synuclein (Gath et al. 2012): 140 residues; calculations performed for residues 1–100; the 40 C-terminal residues are known to be very flexible and therefore invisible in the spectra. C-terminal domain of the Ure2 prion (Ure2p) (Habenstein et al. 2011): 242 residues numbered 113–354. Automated assignments were based for ubiquitin on 11 peak lists from DARR, NCA, NCO, CANCO, CAN(CO)CA, N(CO)CACB, NCACB, NCACBCX, NCACX, NCOCX and CCC solid-state NMR spectra, for HET-s(218–289) on seven peak lists from DARR, CANCO, CCC, NCA, NCACB, NCACX, and NCOCX spectra, for α-synuclein on six peak lists from CANCO, NCACB, NCACO, NCOCA, NCA, and CCC spectra, and for Ure2p on four peak lists from CAN(CO)CA, NCACX, N(CO)CACB, and CCC spectra. The NCACX list for Ure2p included peaks obtained from two spectra, NCACX and NCACB. Details of the NMR measurements have been reported previously (Gath et al. 2012; Habenstein et al. 2011; Siemer et al. 2006; Wasmer et al. 2008). The manually determined chemical shift assignments (Supplementary Tables S1–S7) were used as reference assignments to evaluate the correctness of the assignments from automated procedures.

Peak lists

Peak lists were obtained either with the automated peak-picking algorithm of the program CcpNmr Analysis (Stevens et al. 2011) without manual corrections, or manually during and with partial knowledge of the manual assignment. Only peak positions are relevant for ssFLYA; peak intensities are not used. Peak list statistics are given in Table 2. A measured peak was considered as assignable if, based on the reference assignment, there was at least one expected peak within a tolerance of 0.55 ppm. The remaining measured peaks are likely to be artifacts. Since for some spectra, e.g., CCC and DARR, only expected peaks with a high probability to be observed in the measurement were generated (Table 1), the number of artifact peaks could be overestimated for these spectra. The completeness, defined as the percentage of expected peaks that can be mapped to a measured peak based on the reference assignment provides a measure of how many real peaks have been picked. The remaining expected peaks correspond to missing peaks in the measured peak list. The peak lists and reference assignments are available for download from http://www.cyana.org/ssflyalists.tgz.

Assignment calculations

The ssFLYA resonance assignment algorithm was used in the same way and with the same parameters for the four proteins. The tolerance for chemical shift matching was 0.55 ppm for ^{13}C and ^{15}N for all calculations. The same tolerances were used for the determination of the assignments and their evaluation by comparison with the manually determined reference assignments. While all experimental polarization-transfer schemes applied here employ the dipolar interaction, the selective cross-polarization steps as well as the DREAM mixing periods employed have characteristics that can be well described by a through-bond scheme. Expected peaks for these spectra were generated according to the polarization transfer rules of the CYANA library (Table 1). Expected peaks for the DARR spectra were generated on the basis of 20 conformers calculated with CYANA that fulfill the steric restraints but are otherwise random. Expected DARR peaks with probabilities 0.9, 0.8, 0.7, 0.6, and 0.5 were generated for the ^{13}C–^{13}C distances that were shorter than 4.0, 4.5, 5.0, 5.5, and 6.0 Å, respectively, in all 20 random conformers. The population size for the evolutionary algorithm was 50.

Chemical shift assignments were consolidated from 20 independent runs with different random number generator seeds (López-Méndez and Güntert 2006; Schmidt and Güntert 2012). The assignment of an atom was classified as "strong" if 80 % or more of its 20 chemical shift values deviated by less than the tolerance of 0.55 ppm from the consensus value.

Results and discussion

Automated assignments were determined for the proteins ubiquitin, HET-s(218–289), α-synuclein, and the Ure2p C-terminal domain; using 11, 7, 6, and 4 different types of spectra, respectively (see section Methods). Terminal parts of the protein sequences that are known to be flexible and not observable were not considered in the calculations, leading to input sequences of 68–242 residues. All data sets used for this paper consist exclusively of experimentally measured spectra. Peak lists were obtained by purely automatic (except for Ure2p) and manual peak picking (Table 2). They are realistic and far from ideal, as indicated in Table 2 by the percentages of assigned measured peaks and the completeness of measured peaks with respect to the known reference assignments, which shows that the individual experimental peak lists lack 13–82 % of the expected peaks and contain 3–91 % artifacts. Automatically prepared peak lists have in general more missing peaks (expected peaks not found in the peak list), and more artifact peaks (with no corresponding expected peak) than manually picked ones. These differences are small for ubiquitin and HET-s(218–289), and more pronounced for α-synuclein. The experimental input to the ssFLYA algorithm consisted exclusively in the positions of the peaks in the 2D or 3D spectra. Preparations that imply partial assignments, for example grouping the chemical shifts by spin system, assigning them to positions (N, C$^{\alpha}$, C$^{\beta}$, C$^{\gamma}$, etc.) within a residue, or restricting the possible residue types for spin systems (Hu et al. 2011), were not done.

The ssFLYA assignments were compared to the manually determined ^{13}C and ^{15}N reference assignments (Gath et al. 2012; Habenstein et al. 2011; Siemer et al. 2006; Wasmer et al. 2008). The latter are complete to 86 % for ubiquitin, 80 % for HET-s(218–289), 75 % for α-

Table 2 Experimental peak lists for ubiquitin, HET-s(218–289), α-synuclein, and Ure2p

Spectrum	Expected peaks	Automatic peak picking				Manual peak picking			
		Measured Peaks	Assigned (%)	Complete (%)	Deviation (ppm)	Measured Peaks	Assigned (%)	Complete (%)	Deviation (ppm)
Ubiquitin									
DARR	2,952	1,324	80	62	0.31	929	83	43	0.31
NCA	75	60	85	77	0.26	65	80	79	0.27
NCO	77	71	63	69	0.30	54	80	71	0.30
CANCO	69	90	61	83	0.36	73	79	87	0.35
CAN(CO)CA	203	94	59	28	0.40	68	96	32	0.39
N(CO)CACB	189	108	51	30	0.36	77	69	29	0.37
NCACB	274	184	48	32	0.40	108	81	32	0.38
NCACBCX	315	170	49	28	0.33	72	92	23	0.31
NCACX	351	218	69	45	0.31	178	81	43	0.32
NCOCX	248	201	53	44	0.37	120	86	43	0.35
CCC	212/382[a]	345	20	37	0.32	790	16	36	0.31
Total	4,965/5,135	2,865	66	53	0.32	2,534	64	41	0.32
Average	–	–	58	49	0.34	–	77	47	0.33
HET-s(218–289)									
DARR	2,530	453	79	26	0.32	495	97	36	0.32
CANCO	67	116	34	61	0.35	57	79	67	0.32
CCC	300	633	17	40	0.31	382	36	47	0.31
NCA	74	51	59	51	0.37	54	81	68	0.34
NCACB	246	118	42	22	0.26	59	95	23	0.25
NCACX	306	232	41	33	0.30	159	70	37	0.28
NCOCX	210	242	48	58	0.31	143	84	58	0.29
Total	3,733	1,845	47	31	0.31	1,349	78	9	0.31
Average	–	–	46	42	0.32	–	77	48	0.30
α-synuclein									
CANCO	99	76	68	53	0.34	94	86	83	0.27
NCACB	358	271	58	48	0.26	224	71	45	0.20
NCACO	358	289	62	54	0.32	246	66	47	0.26
NCOCA	262	396	38	60	0.27	294	65	75	0.22
NCA	100	103	56	68	0.34	103	57	69	0.33
CCC	426	1,488	9	34	0.26	370	42	40	0.22
Total	1,603	2,623	30	49	0.29	1,331	66	53	0.24
Average	–	–	49	53	0.30	–	65	60	0.25
Ure2p									
CAN(CO)CA	907					346	94	37	0.22
NCACX	1,169					446	81	32	0.21
N(CO)CACB	613					142	76	18	0.22
CCC	1,094					815	32	27	0.25
Total	3,783					1,749	64	30	0.23
Average	–					–	71	29	0.23

Expected peaks Number of expected peaks by ssFLYA based on the polarization transfer rules of Table 1, or, for the DARR spectrum, consistently short distances in a bundle of randomized conformers (see section Methods). *Measured peaks* Number of measured peaks

Assigned Percentage of measured peaks that can be assigned, within a tolerance of 0.55 ppm, based on the reference chemical shift assignments. The theoretical maximum of 100 % corresponds to having all measured peaks assigned. Note that several expected peaks can be mapped to the same measured peak, i.e., assignments of measured peaks can be unambiguous or ambiguous. Remaining unassigned measured peaks are likely to be artifacts. *Complete* Percentage of expected peaks that can be mapped to a measured peak based on the reference chemical shift assignments. The theoretical maximum of 100 % corresponds to the situation that the measured peak list contains all expected peaks. Each expected peak can be mapped to at most one measured peak. Remaining expected peaks correspond to missing peaks in the measured peak list. *Deviation* Root-mean-square deviation between the chemical shift position coordinates of the measured peaks to which an expected peak can be mapped and the corresponding reference chemical shift value

[a] The first number applies to the calculation with automatically picked peaks, the second number to the calculation with manually picked peaks. The difference is due to the fact that the spectral region in the CCC spectrum containing the backbone C′ atoms, C^{γ} of Asn and Asp, and C^{δ} of Gln and Glu was excluded from automatic peak picking. Consequently, no expected peaks involving these atoms were generated for the calculation with automatically picked peaks

synuclein, and 59 % for Ure2p for the residues included in the calculations (see section Methods). Some stretches of residues are missing in the manual assignments because they correspond to dynamic residues that are invisible in the experiments. This information was not used for the ssFLYA calculations, leading to assignments also of these regions (blue in Fig. 1). Since the reference assignments have been obtained by thorough, exhaustive manual analysis, one must assume that the current spectra do not contain sufficient information to make additional assignments not yet present among the reference assignments. In the following, percentages of assignments are therefore given relative to the number of reference assignments. Overall, 77–90 % of all ssFLYA assignments and 79–94 % of the backbone assignments are correct (Table 3; green in Figs. 1, 2). The ssFLYA algorithm reports an assignment for every atom that is assigned to at least one peak but distinguishes between "strong" assignments that are self-consistent over at least 80 % of the individual runs of the algorithm, and "weak", merely tentative assignments (Schmidt and Güntert 2012) (see section Methods). The strong assignments agree with the reference assignments in 88–97 % of all cases and 91–98 % of the cases concerning the backbone (Table 3; dark green in Figs. 1, 2). Ideally, the strong ssFLYA assignments should include all correct assignments but no others, i.e., no assignments that are wrong according to the reference and no assignments for which no reference is present, because the spectra lack sufficient data for the atoms without reference assignment. The algorithm achieved this to a high degree. There remain some exceptions of manually unassigned atoms for which the algorithm reported a strong assignment (dark blue in Figs. 1, 2), whose correctness cannot be ascertained. The individual assignments are visualized in Figs. 1 and 2, and listed in Supplementary Tables S1–S7.

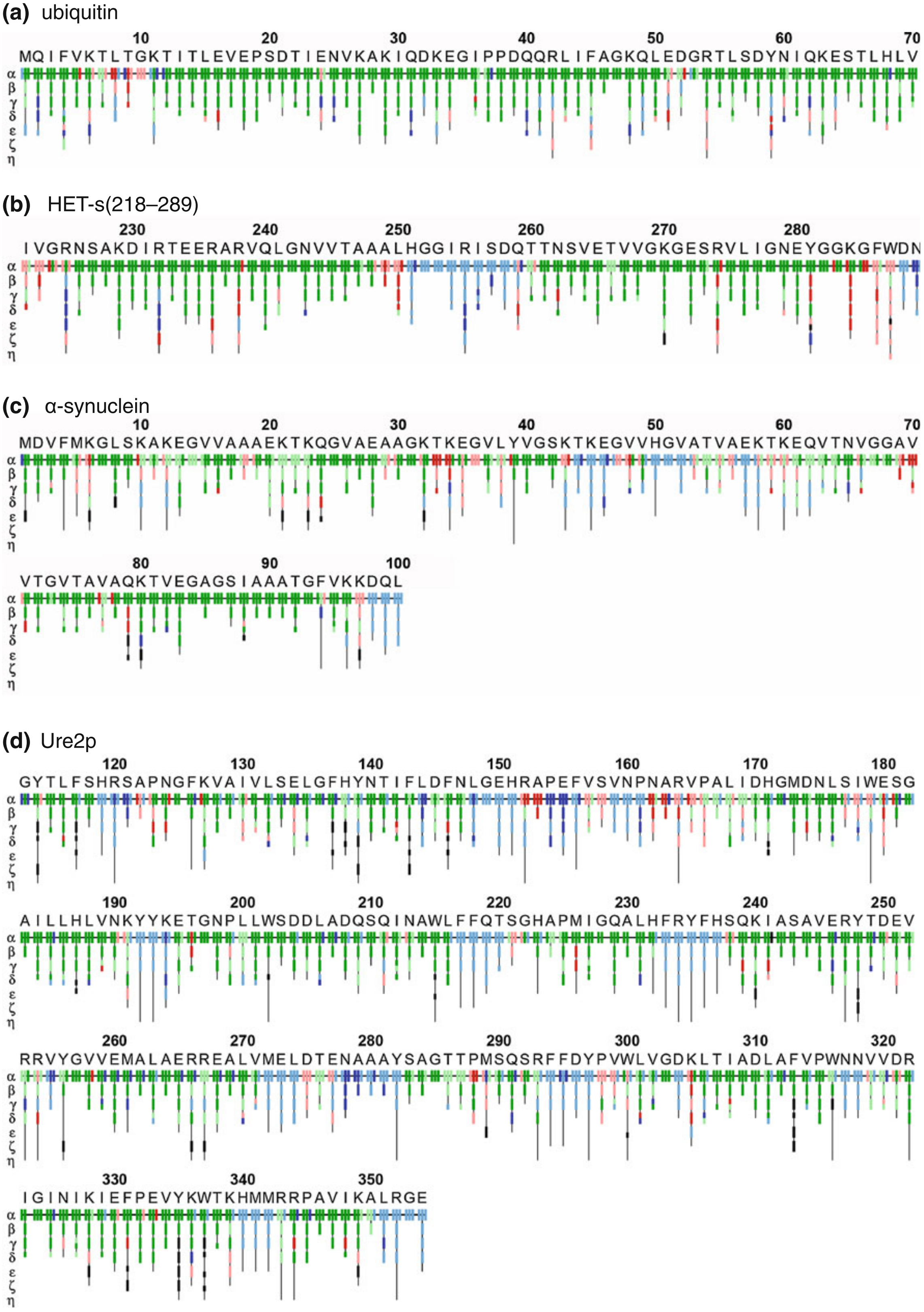
Ubiquitin

The automated assignment was most correct and complete for ubiquitin for which the largest set of different peak lists was available. Incorrect assignments occur for residues 7–11, which are highly dynamic and yield very weak signals, for a few isolated backbone atoms, and for remote side-chain atoms. The automatically picked lists yielded 4 % more correct assignments than the manually picked peak lists, presumably because slightly more expected peaks are missing in the manual peak lists, which overcompensates the presence of more artifacts in the automatic peak lists (Table 2). This is consistent with the earlier empirical finding that for FLYA a missing peak is about 12 times more severe than an additional artifact peak (Schmidt and Güntert 2012).

To investigate the robustness of the ssFLYA algorithm, automated assignment calculations were performed also with selected subsets of 2–8 out of the 11 automatically picked peak lists that were available for ubiquitin. The input always included complementary lists with which it is possible to reveal the sequential connectivity. Table 4 summarizes the results in comparison to the 90/94 % correct assignments for all/backbone atoms obtained using all 11 spectra. Dropping the 2D spectra (DARR, NCA, NCO) does not significantly affect the results (88/95 % correct assignments with eight 3D spectra). Lower but still considerable degrees of correct assignments can be obtained with small numbers of spectra, e.g., up to 81/92 % correct assignments for all/backbone atoms from four spectra (CANCO, CANCOCA, NCACX, NCOCX), 79/88 % from three spectra (CANCO, NCOCACB, NCACX), and 69/75 % from two spectra (NCOCACB, NCACX). As expected, the maximum assignment correctness for a specific number of peak lists decreases with the number of peak lists. For the assignment of backbone atoms, one can obtain results with four peak lists that are just 2 percentage points below the results that were obtained using all available peak lists. Nevertheless, the correctness varies up to nearly 40 % for different selections with a fixed number of peak lists. Consequently, in the present case the combination and the quality of the peak lists are more indicative for the assignment correctness than the number of peak lists. Using NCACBCX and/or NCACX peak lists was a prerequisite for high correctness. Throughout all calculations with reduced data sets that yielded more than 40 % strong assignments, the strong assignments remained to 85–99 % correct, except for one case with 60 % strong assignments (79 % correct), and the extent of strong assignments is a good indicator for the data quality, e.g., the percentages of all strong assignments and all correct assignments are linearly correlated with a correlation coefficient of 0.96.

HET-s(218–289) amyloid fibrils

For HET-s(218–289), manual and automatic peak picking yielded similar results with a correctness of 77 % for all and 88–89 % for the strong assignments. No reference assignments are available for the region of residues 251–259. Incorrect assignments occur mostly adjacent to this region, at the beginning and end of the assigned sequence, and for single side-chain atoms. In the assignment based on automatic peak lists almost all of the assignments for residues 251–259 were (correctly) classified as weak (light blue in Fig. 1), whereas in the case of manual peak picking several (probably erroneous) assignments in this region are classified as strong (dark blue in Fig. 2).


(a) ubiquitin
10 20 30 40 50 60 70
MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLV
α β γ δ ε ζ η
(b) HET-s(218–289)
230 240 250 260 270 280
IVGRNSAKDIRTEERARVQLGNVVTAAALHGGIRISDQTTNSVETVVGKGESRVLIGNEYGGKGFWDN
α β γ δ ε ζ η
(c) α-synuclein
10 20 30 40 50 60 70
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAV
α β γ δ ε ζ η
80 90 100
VTGVTAVAQKTVEGAGSIAAATGFVKKDQL
α β γ δ ε ζ η
(d) Ure2p
120 130 140 150 160 170 180
GYTLFSHRSAPNGFKVAIVLSELGFHYNTIFLDFNLGEHRAPEFVSVNPNARVPALIDHGMDNLSIWESG
α β γ δ ε ζ η
190 200 210 220 230 240 250
AILLHLVNKYYKETGNPLLWSDDLADQSQINAWLFFQTSGHAPMIGQALHFRYFHSQKIASAVERYTDEV
α β γ δ ε ζ η
260 270 280 290 300 310 320
RRVYGVVEMALAERREALVMELDTENAAAYSAGTTPMSQSRFFDYPVWLVGDKLTIADLAFVPWNNVVDR
α β γ δ ε ζ η
330 340 350
IGINIKIEFPEVYKWTKHMMRRPAVIKALRGE
α β γ δ ε ζ η

◀ **Fig. 1** Extent, correctness, and reliability of individual assignments obtained with the ssFLYA automated resonance assignment algorithm using automatically picked peak lists for **a** ubiquitin, **b** HET-s(218–289), **c** α-synuclein, and manually picked peak lists for **d** Ure2p. Each assignment for an atom is represented by a *colored rectangle*. *Green*, assignment by ssFLYA agrees with the manually determined reference assignment within a tolerance of 0.55 ppm; *red*, assignment differs from reference; blue, assigned by ssFLYA but no reference available; *black*, with reference assignment but not assigned by ssFLYA. *Respective light colors* indicate assignments classified as weak by the chemical shift consolidation. The α-row shows for each residue the N, C^{α}, and C′ assignments from *left* to *right*. The rows β-η show the side-chain assignments for the heavy atoms. In the case of branched side-chains, the corresponding row is split into an upper part for one branch and a lower part for the other branch

α-synuclein amyloid fibrils

For α-synuclein, ssFLYA achieved a correctness of 77 % for all and 89 % for the strong assignments with automatically picked peak lists, and 89 % for all and 94 % for the strong assignments with manually prepared peak lists. Reference assignments are lacking for residues 44–57 [the few reference assignments in this region are tentative and were not included in the BMRB deposition (Gath et al. 2012)], and three residues at the C-terminal end of the ordered region. Most of these residues are highly dynamic and do not lead to peaks in the dipolar-transfer-based

Table 3 Statistics of resonance assignments for ubiquitin, HET-s(218–289), α-synuclein, and Ure2p

Class	Ubiquitin		HET-s(218–289)		α-Synuclein		Ure2p
	Automatic	Manual	Automatic	Manual	Automatic	Manual	Manual
Backbone and side-chains							
Reference assignments	381	381	326	326	422	422	942
ssFLYA assignments							
All (strong & weak)	381 (100 %)	381 (100 %)	323 (99 %)	324 (99 %)	409 (97 %)	409 (97 %)	886 (94 %)
All, correct	341 (90 %)	329 (86 %)	248 (77 %)	251 (77 %)	315 (77 %)	364 (89 %)	735 (83 %)
All, incorrect	40 (10 %)	52 (14 %)	75 (23 %)	73 (23 %)	94 (23 %)	45 (11 %)	151 (17 %)
Strong	322 (85 %)	321 (84 %)	254 (79 %)	267 (82 %)	256 (63 %)	365 (89 %)	623 (70 %)
Strong, correct	311 (97 %)	300 (93 %)	225 (89 %)	235 (88 %)	228 (89 %)	343 (94 %)	575 (92 %)
Strong, incorrect	11 (3 %)	21 (7 %)	29 (11 %)	32 (12 %)	28 (11 %)	22 (6 %)	48 (8 %)
Weak	59 (15 %)	60 (16 %)	69 (21 %)	57 (18 %)	153 (37 %)	44 (11 %)	263 (30 %)
Weak, correct	30 (51 %)	29 (48 %)	23 (33 %)	16 (28 %)	87 (57 %)	21 (48 %)	160 (61 %)
Weak, incorrect	29 (49 %)	31 (52 %)	46 (67 %)	41 (72 %)	66 (43 %)	23 (52 %)	103 (39 %)
Backbone (N, C^{α}, C′, C^{β})							
Reference assignments	266	266	222	222	338	338	672
ssFLYA assignments							
All (strong & weak)	266 (100 %)	266 (100 %)	220 (99 %)	220 (99 %)	338 (100 %)	338 (100 %)	671 (100 %)
All, correct	250 (94 %)	246 (92 %)	183 (83 %)	186 (85 %)	266 (79 %)	305 (90 %)	576 (86 %)
All, incorrect	16 (6 %)	20 (8 %)	37 (17 %)	34 (15 %)	72 (21 %)	33 (10 %)	95 (14 %)
Strong	243 (91 %)	247 (93 %)	186 (85 %)	199 (90 %)	213 (63 %)	309 (91 %)	495 (74 %)
Strong, correct	237 (98 %)	237 (96 %)	171 (92 %)	181 (91 %)	194 (91 %)	292 (94 %)	466 (94 %)
Strong, incorrect	6 (2 %)	10 (4 %)	15 (8 %)	18 (9 %)	19 (9 %)	17 (6 %)	29 (6 %)
Weak	23 (9 %)	19 (7 %)	34 (15 %)	21 (10 %)	125 (37 %)	29 (9 %)	176 (26 %)
Weak, correct	13 (57 %)	9 (47 %)	12 (35 %)	5 (24 %)	72 (58 %)	13 (45 %)	110 (63 %)
Weak, incorrect	10 (43 %)	10 (53 %)	22 (65 %)	16 (76 %)	53 (42 %)	16 (55 %)	66 (38 %)

Input peak lists for the ssFLYA algorithm were obtained by either automatic or manual peak picking (columns ‘Automatic’ and ‘Manual’, respectively). Assignments are considered as correct if they agree with the manually determined reference assignment within the chemical shift tolerance of 0.55 ppm for ^{13}C and ^{15}N. Strong assignments are those of atoms for which 80 % or more of the individual chemical shift values from 20 independent runs of the assignment algorithm deviated by less than 0.55 ppm from the consensus value. Other assignments by ssFLYA are classified as weak. Percentages in the rows “all (strong & weak)” are relative to the corresponding number of reference assignments. Percentages in the rows “strong” and “weak” are relative to the corresponding number of all (strong & weak) ssFLYA assignments. Percentages of correct or incorrect assignments are relative to the corresponding number of strong + weak assignments

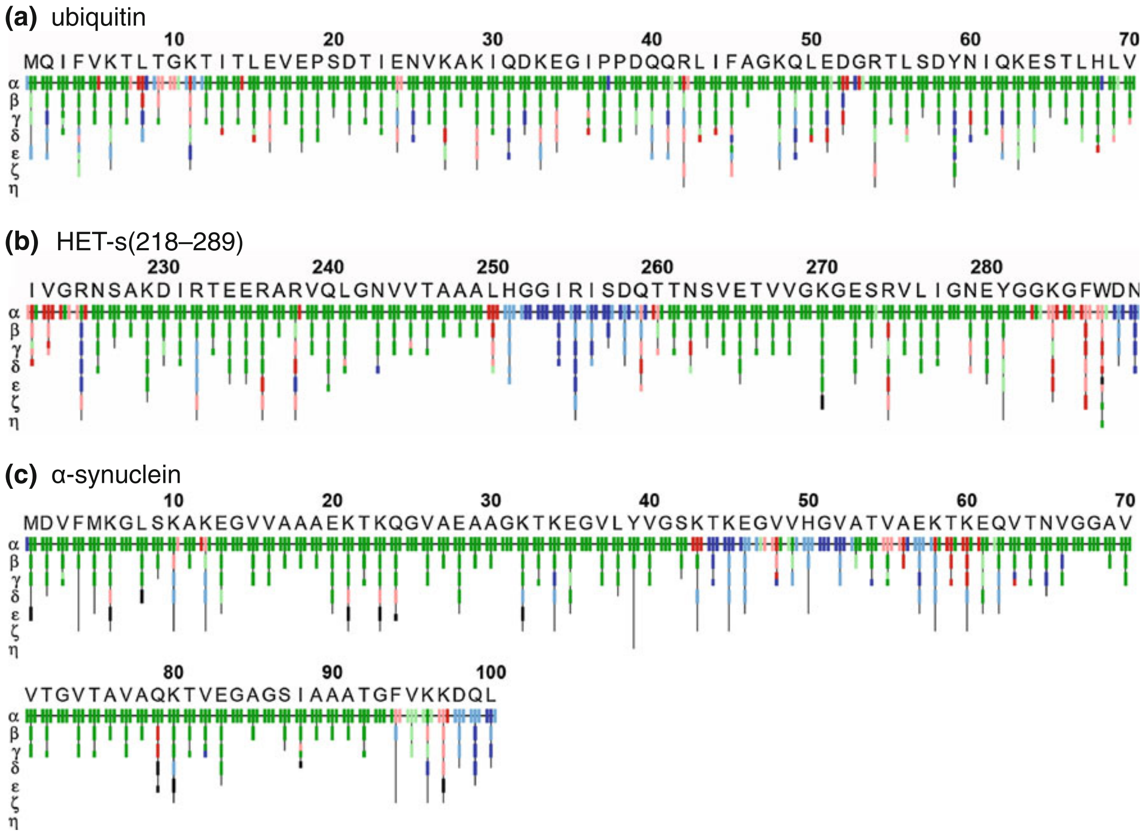


Fig. 2 Extent, correctness, and reliability of individual assignments obtained with the ssFLYA automated resonance assignment algorithm using manually picked peak lists for **a** ubiquitin, **b** HET-s(218–289), and **c** α-synuclein. See Fig. 1 for details

spectra used. With manual peak lists, incorrect assignments occur mainly around these residues. With automatic peak lists, additional incorrect assignments occur in the region of residues 29–39, and at isolated other positions. Resonances of residues 32–34 showed peak doubling and consequently two manual assignments, the first one being represented by the reference assignment, while ssFLYA with automatic peak lists found the second one. Overall, manual peak picking yielded 12 % more correct assignments. This may be due to the fact that two of the manually picked lists, CANCO and NCOCA, which are crucial for backbone assignment and complementing each other, have an above-average quality with 83 and 75 % completeness (percentage of expected peaks that can be mapped to a measured peak based on the reference chemical shifts; Table 2). In addition, automatic peak picking produced many more artifact peaks (on average only 49 % of the measured peaks are compatible with the reference chemical shifts) than manual picking (65 %).

Ure2p C-terminal domain

The Ure2p C-terminal domain is to date the biggest protein assigned by solid-state NMR. About 60 % of all ^{13}C and ^{15}N nuclei could be assigned by manual methods (Habenstein et al. 2011). Out of these 942 reference assignments ssFLYA could assign 735 (83 %) correctly. Out of 623 strong assignments, 575 (92 %) were correct. There are several regions of up to 6 residues without reference assignments. Incorrect assignments by ssFLYA cluster mainly around these regions. Only very few shifts within the manually unassigned regions were classified as strong by ssFLYA.

Conclusions

The results of the automated resonance assignment calculations with four different proteins in Table 3 allow drawing several conclusions that may provide general guidelines for future applications of the ssFLYA algorithm to other proteins in the solid state. (1) ssFLYA yields a (strong or weak) assignment for almost all (94–100 %) atoms, for which a reference assignment could be determined manually. (2) The percentage of strong assignments varies between 63 and 89 %, depending on the quality of the input peak lists. (3) Strong assignments are 89–97 % correct. The reliability of the strong assignments is not significantly affected by the

Table 4 Statistics of ssFLYA resonance assignments for ubiquitin obtained with different sets of automatically picked peak lists

CANCO	CANCOCA	NCOCACB	NCACB	NCACBCX	NCACX	NCOCX	CCC	Backbone and side-chains			Backbone (N, C^{α}, C′, C^{β})		
								Strong	All & correct	Strong & correct	Strong	All & correct	Strong & correct
All 11 peak lists								322 (85 %)	341 (90 %)	311 (97 %)	243 (91 %)	250 (94 %)	237 (98 %)
8 Peak lists													
x	x	x	x	x	x	x	x	355 (93 %)	334 (88 %)	332 (94 %)	264 (99 %)	254 (95 %)	254 (96 %)
4 Peak lists													
x	x				x	x		310 (81 %)	309 (81 %)	293 (95 %)	242 (91 %)	246 (92 %)	236 (98 %)
x	x	x			x			297 (78 %)	303 (80 %)	284 (96 %)	231 (87 %)	237 (89 %)	228 (99 %)
	x	x		x		x		277 (73 %)	298 (78 %)	263 (95 %)	228 (86 %)	244 (92 %)	222 (97 %)
x		x	x				x	213 (56 %)	251 (66 %)	188 (88 %)	157 (59 %)	190 (71 %)	140 (89 %)
x	x	x				x		120 (31 %)	160 (42 %)	90 (75 %)	107 (40 %)	141 (53 %)	85 (79 %)
3 Peak lists													
x		x			x			285 (75 %)	301 (79 %)	270 (95 %)	216 (81 %)	233 (88 %)	213 (99 %)
x				x		x		205 (54 %)	277 (73 %)	198 (97 %)	169 (64 %)	229 (86 %)	168 (99 %)
	x	x			x			317 (83 %)	278 (73 %)	268 (85 %)	243 (91 %)	212 (80 %)	207 (85 %)
	x				x		x	230 (60 %)	241 (63 %)	182 (79 %)	169 (64 %)	174 (65 %)	135 (80 %)
x		x	x					138 (36 %)	182 (48 %)	120 (87 %)	110 (41 %)	152 (57 %)	96 (87 %)
x	x					x		79 (21 %)	121 (32 %)	62 (78 %)	68 (26 %)	107 (40 %)	57 (84 %)
2 Peak lists													
		x			x			269 (71 %)	263 (69 %)	236 (88 %)	200 (75 %)	199 (75 %)	181 (91 %)
	x				x			208 (55 %)	237 (62 %)	180 (87 %)	153 (58 %)	177 (67 %)	134 (88 %)
	x	x						112 (29 %)	120 (32 %)	72 (64 %)	109 (41 %)	110 (41 %)	70 (64 %)

The ssFLYA calculations were performed with all 11 automatically picked peak lists for ubiquitin (Table 2), and for various subsets of 2–8 peak lists thereof. Peak lists that were used in the respective calculation are marked by an x. The columns “strong” list strong assignments, i.e., those for which 80 % or more of the individual chemical shift values from 20 independent runs of the assignment algorithm deviated by less than 0.55 ppm from the consensus value. The columns “all & correct” list all correct assignments. Assignments are considered as correct if they agree with the manually determined reference assignment within the chemical shift tolerance of 0.55 ppm for ^{13}C and ^{15}N. The columns “strong & correct” list assignments that are simultaneously strong and correct. Percentages in the rows “strong” and “all & correct” are relative to the corresponding number of reference assignments, i.e., 381 for all backbone and side-chain atoms, or 266 for the backbone (N, C^{α}, C′, C^{β}) atoms (Table 3). Percentages in the rows “strong & correct” are relative to the number of all (correct + incorrect) strong ssFLYA assignments in the corresponding column “strong” quality of the input data. However, lower quality input data results in fewer strong assignments. The percentage of strong assignments, which can be determined without knowledge of reference assignments, can serve as a measure for the automated assignment "difficulty" of a given protein and its available peak lists. (4) The percentage of all (strong + weak) correct assignments is in general larger than the percentage of strong assignments. (5) Weak assignments are 28–61 % correct, and should only be accepted after further verification. The classification of assignments as strong or weak is thus a valuable tool to distinguish reliable from merely tentative assignments. (6) Incorrect assignments occur predominantly near chain ends and around regions with lacking signals (Figs. 1, 2).

The ssFLYA results show that the algorithm is capable of correctly assigning experimental solid-state NMR spectra almost completely for small microcrystalline proteins, and to an extent that is comparable to what can be achieved by extensive manual analysis for amyloids and challenging larger proteins. Careful manual peak picking can improve the results especially for difficult systems but is not an absolute prerequisite for the algorithm, which can yield similarly correct assignments also with purely automatic peak picking. The ssFLYA algorithm thus introduces automated assignment into protein solid-state NMR and facilitates structural studies of protein amyloids that are currently inaccessible to other techniques.

Acknowledgments We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation, by the Deutsche Forschungsgemeinschaft (DFG), the Swiss National Science Foundation (Grant 200020_124611), the Agence Nationale de la Recherche (SIMI8 2012), and the European Commission (FP7, contract Bio-NMR 261863).

References

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

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

Gath J, Habenstein B, Bousset L, Melki R, Meier BH, Böckmann A (2012) Solid-state NMR sequential assignments of α-synuclein. Biomol NMR Assign 6:51–55

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44:257–309

Güntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298

Habenstein B, Wasmer C, Bousset L, Sourigues Y, Schütz A, Loquet A, Meier BH, Melki R, Böckmann A (2011) Extensive de novo solid-state NMR assignments of the 33 kDa C-terminal domain of the Ure2 prion. J Biomol NMR 51:235–243

Hu KN, Qiang W, Tycko R (2011) A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. J Biomol NMR 50:267–276

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128: 13112–13122

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

Moseley HNB, Sperling LJ, Rienstra CM (2010) Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of beta 1 immunoglobulin binding domain of protein G (GB1). J Biomol NMR 48:123–128

Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134: 12817–12829

Schmucki R, Yokoyama S, Güntert P (2009) Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. J Biomol NMR 43:97–109

Siemer AB, Ritter C, Steinmetz MO, Ernst M, Riek R, Meier BH (2006) ^{13}C, ^{15}N resonance assignment of parts of the HET-s prion protein in its amyloid form. J Biomol NMR 34:75–87

Stevens TJ, Fogh RH, Boucher W, Higman VA, Eisenmenger F, Bardiaux B, van Rossum BJ, Oschkinat H, Laue ED (2011) A software framework for analysing solid-state MAS NMR data. J Biomol NMR 51:437–447

Tycko R, Hu KN (2010) A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. J Magn Reson 205:304–314

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Wasmer C, Lange A, Van Melckebeke H, Siemer AB, Riek R, Meier BH (2008) Amyloid fibrils of the HET-s(218–289) prion form a β solenoid with a triangular hydrophobic core. Science 319: 1523–1526

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