J Biomol NMR (2014) 59:119–134
DOI 10.1007/s10858-014-9835-1

ARTICLE

Automated robust and accurate assignment of protein resonances for solid state NMR

Jakob Toudahl Nielsen · Natalia Kulminskaya · Morten Bjerring · Niels Chr. Nielsen

Received: 13 February 2014 / Accepted: 29 April 2014 / Published online: 10 May 2014

Abstract The process of resonance assignment represents a time-consuming and potentially error-prone bottleneck in structural studies of proteins by solid-state NMR (ssNMR). Software for the automation of this process is therefore of high interest. Procedures developed through the last decades for solution-state NMR are not directly applicable for ssNMR due to the inherently lower data quality caused by lower sensitivity and broader lines, leading to overlap between peaks. Recently, the first efforts towards procedures specifically aimed for ssNMR have been realized (Schmidt et al. in J Biomol NMR 56(3):243–254, 2013). Here we present a robust automatic method, which can accurately assign protein resonances using peak lists from a small set of simple 2D and 3D ssNMR experiments, applicable in cases with low sensitivity. The method is demonstrated on three uniformly ^{13}C, ^{15}N labeled biomolecules with different challenges on the assignments. In particular, for the immunoglobulin binding domain B1 of streptococcal protein G automatic assignment shows 100 % accuracy for the backbone resonances and 91.8 % when including all side chain carbons. It is demonstrated, by using a procedure for generating artificial spectra with increasing line widths, that our method, GAMES_ASSIGN can handle a significant amount of overlapping peaks in the assignment. The impact of including different ssNMR experiments is evaluated as well.

Keywords Software · Resonance assignments · Solid state NMR · Proteins

Introduction

Resonance assignments (RAs) is the first fundamental step in the process of structure determination of proteins based on NMR spectra and is a prerequisite in all studies of interactions and dynamics in proteins. Manual derivation of the RA requires experienced and skilled human interpretation and is often very time consuming. Nowadays, increasingly larger and more complicated systems are being studied (Griswold and Dahlquist 2002; Fiaux et al. 2002; Xu et al. 2006) in particular in the solid state (Habenstein et al. 2011; Gath et al. 2012; Kulminskaya et al. 2012), which prompts for the inclusion of more and more experimental data and data of higher dimensionality. With this increasing complexity, the data set could even ultimately be too overwhelming and cumbersome for a human to comprehend. Automatic methods were introduced decades ago to analyze liquid-state NMR data (Bartels et al. 1996, 1997; Lukin et al. 1997; Zimmerman et al. 1997; Leutner et al. 1998; Moseley and Montelione 1999; Atreya et al. 2000; Moseley et al. 2001; Coggins and Zhou 2003; Hitchens et al. 2003; Malmodin et al. 2003; Altieri and Byrd 2004; Baran et al. 2004; Jung and Zweckstetter 2004; Eghbalnia et al. 2005; Schmucki et al. 2009; Crippen et al. 2010) as reviewed in (Guerry and Herrmann 2011). For

Electronic supplementary material The online version of this article (doi:10.1007/s10858-014-9835-1) contains supplementary material, which is available to authorized users.

J. T. Nielsen (✉) · N. Kulminskaya · M. Bjerring ·
N. Chr. Nielsen
Center for Insoluble Protein Structures (inSPIN),
Interdisciplinary Nanoscience Center (iNANO), Department of
Chemistry, Aarhus University, Gustav Wieds Vej 14,
8000 Aarhus C, Denmark
e-mail: jtn@chem.au.dk

N. Chr. Nielsen
e-mail: ncn@inano.au.dk

solid-state NMR (ssNMR), however, the data is often of inferior quality due to lower sensitivity and broader lines leading to overlapping signals. Furthermore, the types of data are different for ssNMR with two notable complications. Unless you extensively deuterate samples (Chevelkov et al. 2006; Zhou et al. 2012), protons are rarely detected implying that the strategy of ^{1}H–^{15}N double-axis alignments for peak combination cannot be used making spin system generation a non-trivial task. In addition, homonuclear ^{13}C–^{13}C correlations are mostly achieved through dipole–dipole coupling interactions, i.e. trough space, thus complicating the process of side-chain RAs. Accordingly, methods for liquids are not generally transmittable to ssNMR. Despite these challenges, software has been developed revealing pioneering results for special cases of ssNMR data (Moseley et al. 2010; Tycko and Hu 2010; Hu et al. 2011). Recently, a special version of FLYA (Schmidt and Guntert 2012), called ssFLYA (Schmidt et al. 2013), which is capable of assigning the resonances of general types of ssNMR data was presented. This assignment method is based on constructing an optimal mapping between observed and expected peaks. However, in ssNMR there will often be overlapping peaks, which precludes a one-to-one mapping.

Here we present a new method, GAMES_ASSIGN (**G**enetic **A**lgorithm using **M**aximum **E**ntropy for **S**olid state NMR resonance **ASSIGN**ments of proteins), which is applicable to general types of ssNMR data of average quality (see "Methods"). We demonstrate that this method is capable of handling spectral overlap, and we analyze systematically the influence of the resonance line width on the assignment accuracy. Our method uses only unassigned peak lists and the protein sequence as input and is based on three phases: (1) Spin system generation, (2) RAs, and (3) completion of the assigned spin systems allowing a peak occasionally to have more than one assignment. All phases share the concept of item *pairing* by joining peaks, spin systems, and/or residue positions in the sequence in different combinations. Here we adopt a unified approach for the pairing in all phases using the concepts of maximum entropy and normal-variate-rank selection to make the easy decisions first in stochastic setup. GAMES_ASSIGN is available at http://nmr.au.dk/software/.

Methods

Overview of the process of resonance assignments

GAMES_ASSIGN assigns the resonances automatically for a protein using unassigned peak lists from solid state NMR experiments. Most common ssNMR experiments can be used, such as f.ex. NCACX, NCACO, NCOCX, NCOCA, CONCA and through-space experiments such as DARR-type experiments. Proton experiments are not yet implemented but a general ^{13}C/^{15}N experiment can be included by specifying the atom type (N/C′/Cα, etc. and CX for any carbon) for each axis of the experiment and the residue order (e.g. intra-residue or preceding residue) for each transfer step. The RAs are performed through three consecutive phases based on stochastic choices. The method is briefly summarized in Fig. 1. In the first phase (I), peaks are paired one-by-one to form more or less complete spin systems. If a peak is paired with another peak, which is in turn already paired with a third one, all three will become part of the new spin system. In the second phase (II), sequence specific assignments are conducted by incrementally pairing spin systems through linking or pairing a spin system with a specific position in the protein amino acid sequence. These two phases are repeated multiple times to provide a selection of candidates for RAs. The third and final phase (III) uses the statistics of the initial two phases to rebuild the spin systems. Finally, the incomplete spin systems are completed aiming to assign all missing resonances and provide assignments for all peaks. Due to the stochastic decision-making in the algorithm, the procedure will produce different outcomes when repeated multiple times, and the result with the best statistics can be chosen. All phases are summarized in Fig. 1 and described in detail below.

Stochastic pairing scheme: MENOVAR

In all stages of the algorithm, different pairings are performed. We designed a special protocol, MENOVAR (Maximum Entropy Normal-Variate Rank), in order to do the pairing stochastically, but biased towards the more probable pairings. For a given item, for example, a peak or a spin system, there is a set of possible pairings: peak-to-peak matching (alignment) or spin system to residue matching (spin system typing). To control this process, a pairing "energy", E_{pair}, (an energy, which should be minimized) is calculated by a procedure depending on the particular pairing (see below). Using this energy an un-normalized likelihood of pairing $p = \exp(-E_{pair})$, can be calculated for each pairing candidate. After normalizing, all probabilities for an item sum to unity. In our protocol, the pairing for a given item is found by roulette wheel selection (Tang et al. 1996) selecting a pairing with a statistical rate equal to its estimated probability. Furthermore, to direct the algorithm, prior to the roulette wheel selection, the first item for the pairing is selected by a "minimum entropy selection". The entropy, H, for a given item is calculated by summing the normalized probabilities, p_i, weighted by the logarithm of the probability.

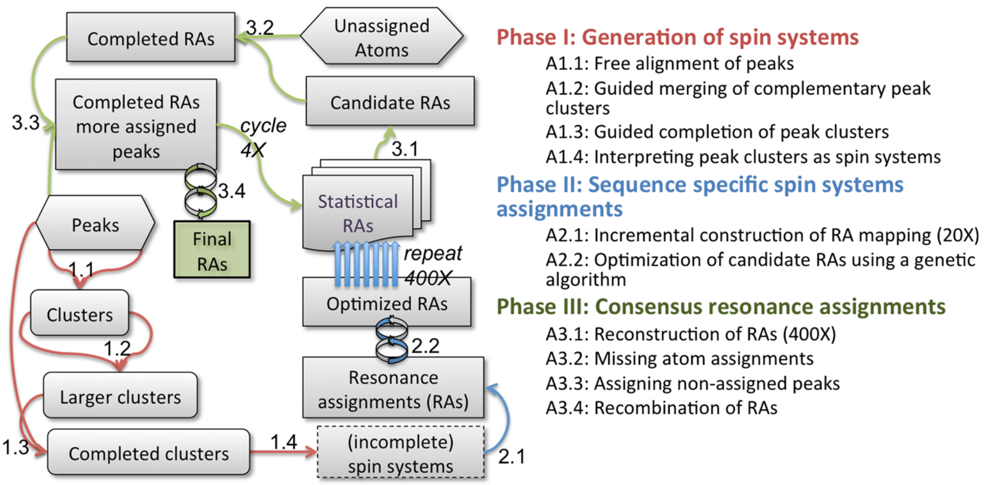
Completed RAs
3.2
Unassigned Atoms
Completed RAs more assigned peaks
cycle 4X
Candidate RAs
3.3
3.1
3.4
Statistical RAs
Peaks
Final RAs
repeat 400X
1.1
Optimized RAs
Clusters
1.2
2.2
Resonance assignments (RAs)
Larger clusters
1.3
Completed clusters
1.4
(incomplete) spin systems
2.1
Phase I: Generation of spin systems
A1.1: Free alignment of peaks
A1.2: Guided merging of complementary peak clusters
A1.3: Guided completion of peak clusters
A1.4: Interpreting peak clusters as spin systems
Phase II: Sequence specific spin systems assignments
A2.1: Incremental construction of RA mapping (20X)
A2.2: Optimization of candidate RAs using a genetic algorithm
Phase III: Consensus resonance assignments
A3.1: Reconstruction of RAs (400X)
A3.2: Missing atom assignments
A3.3: Assigning non-assigned peaks
A3.4: Recombination of RAs

Fig. 1 Overview of the GAMES_ASSIGN algorithm. The generation, starting with the unassigned peaks, of clusters, (incomplete) spin systems, and resonance assignments (*RAs*) are shown in the flow chart to the *left*. Each *arrow* represents parts of the algorithm outlined to the right. A1.1 is shorthand for Algorithm 1.1 explained in detail in the corresponding section with that heading in the text (some of the sections are shown only in the supplementary information). For further visualization of A1.3–A1.3, see Fig. 2. Two converging *arrows* from one item indicate that the operation requires two items, such as 1.1 where two peaks are combined to form a cluster. Two entangled *arrows* represented a recombination process in a GA. Algorithms 1.1 through 2.2 are run in parallel 400 times keeping the best scoring 64 RAs for the statics generation. Algorithms 3.1 through 3.3 are cycled 4 times each time producing 400 RAs and keeping the best 64. After the final cycle A3.4 is executed

$$H = -\sum_{i=1}^{N} p_i \ln(p_i) \tag{1}$$

The items with the lowest entropy will have the most unique choice of pairing compared to the ones with higher entropy. All items are first ranked according to the value of the entropy and the first item of the pairing is then chosen stochastically using a normal-variate-rank scheme. In phase I, (but not in phases II and III) a linear-combination of the entropy and the minimum pairing energy is used for the ranking. In this scheme a random number is drawn from a normal distribution N(0,σ) and the absolute value of this number is rounded to the nearest integer and the member from the ordered list is chosen as the one with this particular rank. By using standard deviations, σ, which are small compared to the total size, better pairings will be chosen more frequently.

Phase I: generation of spin systems

In the first phase, peaks are combined into clusters of peaks. In the initial step the peaks are paired one-by-one based on the agreement of peak frequencies on the shared axes. A pairing of two peaks initializes a *peak cluster*. If a peak is paired with a peak from a peak cluster, the new peak will also be appended to this peak cluster. Furthermore, if two peaks from two different clusters are paired, the two corresponding clusters will merge into one single cluster. This procedure, as summarized in Fig. 2, leads to the formation of a set of peak clusters, which are neither yet assigned to a residue type nor sequence specific. At the end of the clustering process, the peak clusters are interpreted as spin systems based on the experiment definitions.

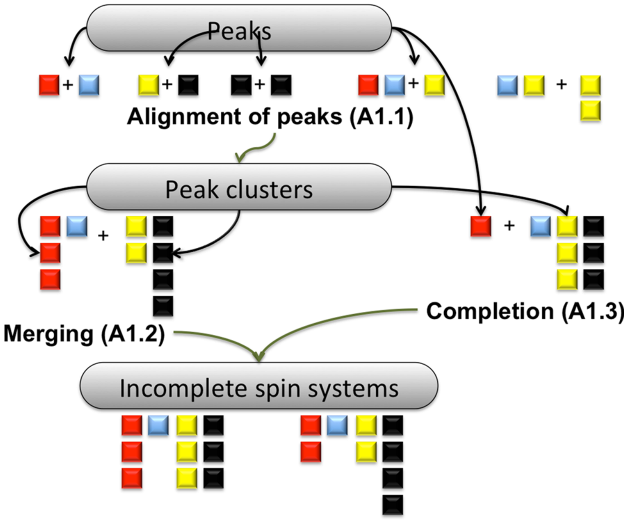

Fig. 2 Visualization of algorithms: A1.1–A1.3. A peak is shown as a *box*, with different colors for different experiments (here the *blue boxes* correspond to experiment with only one expected peak such as NCACO. Peaks from similar experiments are collected in stacked columns, a peak cluster is shown as connected columns of peaks. The individual algorithms are marked at the *arrows* for the related process with numbering corresponding to headings of the sections in the text. A1.3 is described in Supplementary Material

Algorithm 1.1: free alignment of peaks

In the first step of phase *I*, two peaks are paired (aligned) without any preference for the type of the experiment related to the pairing. In contrast, the next step considers only a subset of the pairings chosen to assist formation of typical spin systems (see below). The pairing energy, E_{pair}, as described above, controls the pairing of two peaks. The energy can be interpreted as the logarithm of the likelihood of a correct peak alignment. Here we use an energy inspired by the exponential argument for the normal distribution, leading to an expression for the coordinate match energy, E_c, based on the Euclidian distance:

$$E_c = \sum_{n=1}^{N} \frac{1}{2}\left(\left(\omega_{i,n} - \omega_{j,n}\right)/\sigma_n\right)^2 \qquad (2)$$

where $\omega_{i,n}$ and $\omega_{j,n}$ are the observed peak frequencies for peaks *i* and *j*, respectively, in the *n*'th aligned dimension. An aligned dimension is an axis sharing the same atom type, e.g., peaks from NCACB and NCACO share the two first axes corresponding to N and Cα. The difference is scaled by the average of estimated uncertainties (which must be provided by the user), σ_n, for the two experiments on the particular axis. To summarize, the probability for each peak to match with other peaks are calculated based on the energy, the entropy is calculated and the first peak in the pair is chosen using the normal-variate rank selection procedure. The partner for pairing is subsequently chosen using the Roulette wheel selection. This procedure leads to the formation of clusters of peaks.

Algorithm 1.2: guided merging of complementary peak clusters

The second step starts with the peak clusters generated from the step above. Before two peak clusters are paired into one resulting cluster, their possibility of combination is evaluated with the count expectation rule (see Algorithm 1.3 in Supplementary Material). In this case, two experiments with non-empty subsets that share the most axes (e.g., NCACX and CONCA shares two) are identified. The energy of pairing is defined as the average of the normal pairing energy between all combinations of peak pairs from the two subsets identified by the procedure described above. The pairing is performed using the MENOVAR algorithm similar to the first step.

Algorithm 1.3: guided completion of peak clusters

The third step aims at completing the peak clusters; all details can be found in the Supplementary Material.

Algorithm 1.4: interpreting peak clusters as spin systems

In the final step of phase I, the clusters of related peaks are interpreted as (incomplete) dipeptide "spin systems" henceforth referred to just as spin systems (SSYs). For each peak in the cluster, i.e. for each axis of the peak coordinates in different experiments, there is a definition of *atom type* i.e. ^{15}N or ^{13}C, *residue order* 0 or −1, denoting atoms for residues *i* and *i*−1, respectively, and in some case specific atom (atom name) such as C', Cα or Cβ. For example for a peak in a NCACX peaks list, atom type = (^{15}N, ^{13}C, ^{13}C), residue order = (0, 0, 0), atom name = (N, Cα, None), where "None" denotes a non-specific atom. In this study, per definition, for 2D/3D ^{13}C-DARR, the residue orders are set to (0, 0)/(0, 0, 0) to avoid duplicate membership in SSYs. Following these definitions, the chemical shifts from different peaks and dimensions pointing to the same combination of residue order and atom name (referred to here as atom with order) are grouped leading to sets of chemical shifts. Resulting chemical shifts sets are obtained for N(i), Cα(i)/(i−1), C'(i)/(i−1), Cβ(i)/(i−1) depending on the experimental data available for the specific protein studied and, in addition, chemical shift sets for non-specific carbons CX(i) and CX(i−1). For each chemical shift set related to a specific atom, the average chemical shift is calculated and the mapping between atoms with order and the average chemical shifts constitutes the temporary incomplete non-assigned SSYs—to be optimized in the next phases. The CX chemical shifts sets are treated as a special case because they contain convoluted chemical shifts for several different ^{13}C atoms in a residue. These sets are expected to contain an a priori unknown number of merged populations, which each are normal distributions with unknown population size and average value. The standard deviation for the sub-populations is easier estimated as a typical uncertainty of the experiment. GAMES_ASSIGN implements a heuristic stochastic algorithm optimized for speed to identify the underlying subpopulations in the CX chemical shift deconvoluting the set into a group of subsets related to still unspecific but distinct ^{13}C atoms (see Supplementary Material).

Phase II: sequence specific spin systems assignments

In phase II, the resonances are tentatively assigned by sequence specific assignments using the set of (incomplete) spin systems generated in phase I. This procedure of spin system based sequence specific RAs is a so-called quadratic assignment problem (QAP) as discussed before (Eghbalnia et al. 2005), which is known to be computationally intensive and hence cannot be solved exhaustively (Nagarajan and Sviridenko 2009; Cela 1998). The QAP here consists of assigning a sequence specific position for each of the unassigned SSYs. Candidate assignments are ranked according to two energies (scores): (1) The SSY typing energy related to how likely a certain SSY is to be due to resonances for a certain residue dipeptide step in the sequence and (2) the SSY linking energy related to the agreement between overlapping assignments of ^{13}C for SSYs assigned to consecutive positions in the sequence. The challenge here is both, to set a good definition for the typing and linking energies being robust enough to handle incomplete SSYs and possibly false peaks, but also to design an efficient global optimization algorithm for minimizing the total energy.

Definition of the energies for ranking of resonance assignment candidates

The spin system typing energy, $T = T_S + T_N$, is calculated as a sum of contributions from the energy related to specifically assigned resonances, T_S, and non-specifically assigned resonances, T_N.

$$T_S(n,\tau) = \sum_{i=0,-1,m\in S(i))} t_{mm}(i,n,\tau) \tag{3}$$

$$t_{jk}(i,n,\tau) = \min\left(e_{\max}, \frac{1}{2}\left(\left(\delta_{j,obs}(n) - \delta_{k,ref}(\tau)\right)/\sigma_k\right)^2 - e_{obs}\right) \tag{4}$$

where t_{jk} is the resonance-database value matching energy contribution and i denotes looping through residue orders (directions) in the sequence and expected specific atom, $n \in S\,(i)$, in residue i (such as backbone N, Cα and C′). $\delta_{j,obs}(n)$ is the observed chemical shift for atom j, for a certain SSY, n, and $\delta_{k,ref}(\tau)$ is the reference shift for the assigned residue position, $\tau = \tau(n)$, for SSY, n, and σ_k the reference standard deviation estimated from a database of deposited assigned chemical shifts to the specific amino acid type (Zhang et al. 2003; Nielsen et al. 2012). Furthermore, e_{max} is a correcting term, truncating the contribution from false peaks. Note that the indices n and τ corresponding to SSY and residues position numbering will be implicit in expressions henceforth for increased readability. If a certain chemical shift, $\delta_{j,obs}$, is not found in the SSY then the maximum value $e_{miss} = 1.0$ is used for the contribution, t_{jk}. This parameter is important for analyzing ssNMR data as often data is incomplete with missing expected peaks due to low sensitivity. The application of e_{max} is also vital, since false peaks can be present as noise peaks but also as false peak alignments in the first phase, which is more pronounced for samples with larger line widths. Furthermore, $e_{obs} = 0.75$ is a constant rewarding each included assignment. Some experiments, such as NCACX, use through-space transfer steps and thus produces peaks without an atom-specific assignment possibility on the corresponding axis (atom name, as defined in A1.4 is "None", corresponding to CX). The contribution to the typing energy from these non-specifically assigned side chain resonances is calculated with a modified expression (see Supplementary Material and Fig. 3).

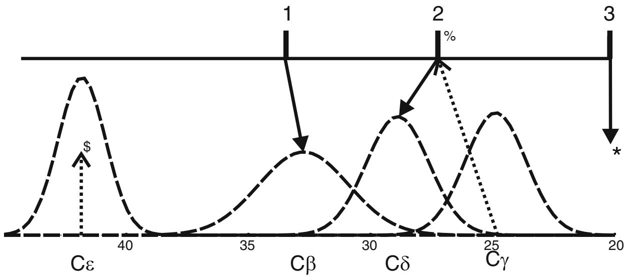

Fig. 3 Visualization of the amino acid typing of non-specifically assigned side chains atoms. The expected resonance positions are indicated by normal probability distributions centered around the average value with a standard deviation given by the sample average from the database and shown with *broken lines* with corresponding carbon atom name given below where the number indicate chemical shifts. Here the lysine side chain carbons are provided as an example. The observed resonance positions are shown at the *top* of the diagram as *black bars* and annotated with numbers above. *Solid arrows* indicate a match from a resonance position to a distribution and *dotted arrows* indicate the reverse match. Note that there is a unique match between resonance position, 1, and the distribution for the expected resonance for Cβ. The small symbols highlight special cases: *asterisk*) The resonance position 3 corresponds to a noise peak and therefore no matching distribution was found since the energy related to the difference between pos. 3 and the center for Cγ exceed the threshold, e_{max}. $) The expected peak for Cε is not matched with any observed resonance position. %) Both the Cγ and the Cδ are matched to the same resonance position, 2

The linking energy, $L = L_S + L_N$, is calculated for two SSYs assigned to be in consecutive positions, i and $i + 1$ in the sequence implying that resonances should be similar for SSYs i with order 0 and $i + 1$ with order −1 as:

$$L_S = \sum_j l_{ii}, l_{jk} = \min\left(e_{\max,k}, \frac{1}{2}\left(\left(\delta_{j,0} - \delta_{k,-1}\right)/\sigma\right)^2 - e_{obs}\right) \tag{5}$$

where the summations is over atoms with specific assignments and σ is an estimated uncertainty for agreement between observed assigned chemical shifts $\delta_{j,0}$ and $\delta_{j,-1}$ for orders 0 and −1, respectively, for the two SSYs. Again, similar to the definition of the typing, if one of the expected resonances are missing from the SSY a penalty value is used $l_{jk} = e_{miss}$ (=1.0 here). Finally, the linking energy for the non-specific atoms is defined similarly to the typing energy, $L_N = L_{N0} + L_{NR}$

$$L_{N0} = \sum_j \min_k l_{jk}, \ \textit{and}\ K = \bigcap_j \arg\min_k l_{jk}$$
$$\textit{and}\ L_{NR} = \sum_{k\notin K} \min_j l_{jk} \tag{6}$$

Since multiple candidate RAs are derived in parallel with different sets of assembled SSYs, an energy, $S(n)$, describing the quality of the SSY, n, is included:

$$S(n) = \frac{1}{N}\sum_{i}^{N} s_i(n), s_i(n) = \min\left(5.0\lambda^2, \sigma(R_i(n))^2\right) \tag{7}$$

where the index, i, runs across all atom types in the SSY, $\sigma(R_i(n))$ denotes the sample deviation among all coordinates, R_i, assigned to the same resonance and is set to $\sigma = \lambda^2$ if there is only one single coordinate in the set. By this definition the total energy, E_{tot}, for a given RA, τ, a certain SSY, n, assigned to residue position, $\tau(n)$, and linked to SSY, m, is:

$$\begin{aligned} E_{tot}(\tau, n) = {} & S(n) + L(n, m) + T(n, \tau(n)), \\ m = \tau^{-1}(\tau(n) - 1) & \end{aligned} \tag{8}$$

Following this definition the total protein RA energy, E_{RA} becomes:

$$E_{RA}(\tau, K) = \frac{1}{N}\sum_{n \in K} E_{tot}(\tau, n) + E_{viol}(\tau, K) \tag{9}$$

where N is the number of residues in the protein and E_{viol} is a violation term, which adds a fixed penalty to the total energy for each unassigned SSY in the set of SSYs, K, and each residue position which is not mapped by τ, and adds another constant violation for each peak not assigned to any SSY.

The total energy is also defined for a certain residue position, $t = \tau(n)$. For this case, the typing energy is evaluated for only one of the directions:

$$T_S(n, \tau)|_i = \sum_{m \in S(i)} t_{mm}(i, n, \tau) \tag{10}$$

i.e. the residue order, i, is fixed, and a similar expression is used for the non-specifically assigned atoms. With this definition, the position based total residue energy, $E'_{tot}(t)$, is defined as:

$$\begin{aligned} E'_{tot}(\tau, t) = {} & S\left(\tau^{-1}(t)\right) + \frac{1}{2}L\left(\tau^{-1}(t), \tau^{-1}(t-1)\right) \\ & + \frac{1}{2}L\left(\tau^{-1}(t+1), \tau^{-1}(t)\right) + T\left(\tau^{-1}(t), t\right)|_0 \\ & + \frac{1}{2}T\left(\tau^{-1}(t), t\right)|_{-1} + \frac{1}{2}T\left(\tau^{-1}(t+1), t+1\right)|_0 \end{aligned} \tag{11}$$

Algorithm for spin system based resonance assignments

The object of the RA is to find a mapping, $\tau(n) = t$, of the spin systems, n, to the position, t, in the amino acid sequence that minimizes the total protein RA energy, E_{RA} (Eq. 9). The identification of this optimal mapping is a QAP. Here we construct a large number of different solutions by performing pairing operations of typing and linking SSYs using the stochastic MENOVAR scheme and subsequently combining the solutions to form more optimal solutions using a genetic algorithm (GA).

Algorithm 2.1: incremental construction of resonance assignment mapping

The mapping is first constructed incrementally by pairwise operations of linking two SSYs or relating a SSY to a sequence position (typing). At the initialization of the algorithm all SSYs are non-linked and non-typed. An SSY can become typed, and upon linking of two SSYs a *patch* of linked SSYs are formed. Note that the energies for pairing need to be dynamically updated at each step towards the completion of the incremental assignments. The reason is *pairing implications*: once an SSY in a patch is typed all SSYs in the patch must be typed. Also linking to a typed SSY requires typing the new SSY (see Fig. 4). At the beginning of the algorithm all SSYs with only one unique possibility for linking are used to form small patches. Subsequently, all the remaining pairings are performed using the principles of minimum entropy and the normal-variate-rank selection scheme as defined above. An example of a run of the incremental algorithm is found in Fig. 5. The above procedure is repeated a number of times (20 here keeping the 16 with the lowest energy) to form a selection of candidate RAs based on the initial set of SSYs. These candidate solutions will be subject to refinement in the second stage of phase II.

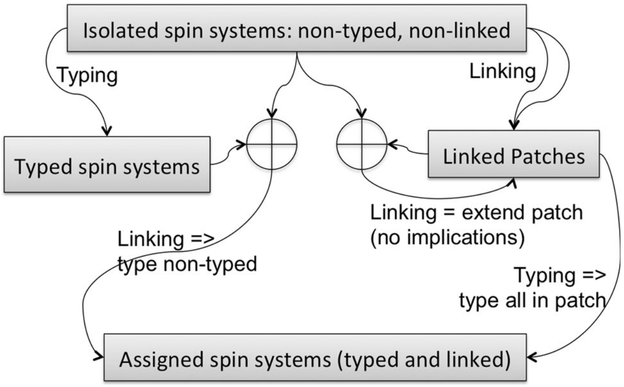

Fig. 4 Incremental RAs based on linking and typing of spin systems. Four different states for a spin system (*SSY*) are visualized by *grey boxes*: Isolated, typed (but not linked), linked (patches) and assigned (both typed and linked). The process of linking two SSYs are illustrated by two connected *arrows* and a connector point, typing is illustrated by an *arrow*. Some processes and states leads to an implicated process as indicated by a "=>": Linking a typed SSY with an isolated SSY implies the need for typing the isolated, whereas typing a patch implies the need to type all SSY member of the patch

```
------------------------------------------------------- | 012 89 de hi jkl uw xp zAB EF HJ OPQ RS X4 UV
------------------------------------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV
------------------------------------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV ab
------------------------------------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV ab MN
----5-------------------------------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV ab MN
----5-------------------------------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV ab MN st
----5--------------------st---------------------------- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 UV ab MN
----5--------------------st-------------------------UV- | 012 89 de hijkl uw xp zAB EF HJ OPQ RS X4 ab MN
----5----------hijkl-----st-------------------------UV- | 012 89 de uw xp zAB EF HJ OPQ RS X4 ab MN
----5----------hijkl-----st-------------------------UV- | 012 89 de uw xpq zAB EF HJ OPQ RS X4 ab MN
----5----------hijkl-----st-------------------------UV- | 012 789 de uw xpq zAB EF HJ OPQ RS X4 ab MN
----5----------hijkl-----st-------------------OPQ---UV- | 012 789 de uw xpq zAB EF HJ RS X4 ab MN
----5----------hijkl----rst-------------------OPQ---UV- | 012 789 de uw xpq zAB EF HJ RS X4 ab MN
----5----ab----hijkl----rst-------------------OPQ---UV- | 012 789 de uw xpq zAB EF HJ RS X4 MN
----5----ab----hijkl----rst----------EF-------OPQ---UV- | 012 789 de uw xpq zAB HJ RS X4 MN
----5----ab----hijkl----rstuw--------EF-------OPQ---UV- | 012 789 de xpq zAB HJ RS X4 MN
----5----ab----hijkl----rstuw--------EF-------OPQ---UV- | 012 789 de xpq zAB HJ RS X4 MN KL
----5----ab----hijkl----rstuw--zAB---EF-------OPQ---UV- | 012 789 de xpq HJ RS X4 MN KL
----5----ab----hijkl----rstuw-yzAB---EF-------OPQ---UV- | 012 789 de xpq HJ RS X4 MN KL
----5-789ab----hijkl----rstuw-yzAB---EF-------OPQ---UV- | 012 de xpq HJ RS X4 MN KL
----5-789ab----hijkl----rstuw-yzAB---EF---KL--OPQ---UV- | 012 de xpq HJ RS X4 MN
----5-789ab----hijkl----rstuw-yzAB---EF---KLMNOPQ---UV- | 012 de xpq HJ RS X4
----56789ab----hijkl----rstuw-yzAB---EF---KLMNOPQ---UV- | 012 de xpq HJ RS X4
----56789ab----hijkl----rstuw-yzAB---EF---KLMNOPQ---UV- | 012 de xpq HJ RS X4
----56789ab----hijkl----rstuw-yzAB---EF---KLMNOPQ---UV- | 012 cde xpq HJ RS X4
----56789ab----hijkl----rstuw-yzAB---EF---KLMNOPQRS-UV- | 012 cde xpq HJ X4
----56789ab----hijklm---rstuw-yzAB---EF---KLMNOPQRS-UV- | 012 cde xpq HJ X4
----56789ab----hijklm---rstuw-yzAB---EF---KLMNOPQRS-UV- | 012 cdef xpq HJ X4
----56789ab----hijklm---rstuw-yzAB---EF---KLMNOPQRSTUV- | 012 cdef xpq HJ X4
----56789ab----hijklm---rstuw-yzAB---EF---KLMNOPQRSTUV- | 012 cdef xpq HJ X4 Cn
----56789abcdefhijklm---rstuw-yzAB---EF---KLMNOPQRSTUV- | 012 xpq HJ X4 Cn
----56789abcdefhijklm---rstuw-yzAB---EF---KLMNOPQRSTUV- | 012 xpq HJ X4 CnD
----56789abcdefhijklm---rstuw-yzAB---EF---KLMNOPQRSTUV- | 012 xpq HJ X4 CnD vg
----56789abcdefhijklmxpqrstuw-yzAB---EF---KLMNOPQRSTUV- | 012 HJ X4 CnD vg
----56789abcdefhijklmxpqrstuw-yzAB--GEF---KLMNOPQRSTUV- | 012 HJ X4 CnD vg
----56789abcdefhijklmxpqrstuw-yzABCnDEF---KLMNOPQRSTUV- | 012 HJ X4 vg
-vg-56789abcdefhijklmxpqrstuw-yzABCnDEF---KLMNOPQRSTUV- | 012 HJ X4
012-56789abcdefhijklmxpqrstuw-yzABCnDEF---KLMNOPQRSTUV- | HJ X4
012-56789abcdefhijklmxpqrstuw-yzABCnDEF-HJKLMNOPQRSTUV- | X4
```

Fig. 5 Example of an incremental assignment run. The status of the algorithm is shown on each *new line*. The spin systems (*SSYs*) are indicated by a *single character*. The assignment progress for the protein sequence is shown to the left of the "|" showing typed SSYs with characters and unassigned residue positions with "-". Patches (linked, but not yet typed SSYs) are shown to the right separated by spaces. To ease readability and to visualize the success of the assignment run, the character used to indicate the different SSYs was chosen related to the most likely sequential SSY assignment based on the manual assignment; the ten digits, followed by *lower case letters* and *upper case letters* alphabetically was used to enumerate the SSYs. This means that "0" correspond to the SSY, which should be assigned as the first residue position. In some cases there might be more than one SSY for the same residue position, and in some cases a SSY would be missing for a certain residue position, therefore, non-consecutive letters (characters) can in some cases still correspond to a correct linking. We emphasize that the manual assignments were exclusively used for annotation purposes for this figure and were not used in any ways to derive the automatic assignments

Algorithm 2.2: optimization of candidate resonance assignments using a genetic algorithm

The set of candidates for RAs is a population of individuals. Each residue, i, represents a gene and the SSYs, s(i), assigned to a residue position is the value for the gene. The population is optimized through evolution by the processes of mutation and recombination (see below) keeping the best solutions. One of the strengths of a GA is that it can use the global knowledge of all candidate solutions to form a better solution by gene combination.

Mutation of solutions In the first part of this GA optimization, weak genes are replaced. This means that residue positions have their assigned SSY removed and subsequently the RA is reconstructed (to hopefully form a more optimal solution) using all unassigned SSYs following the MENOVAR procedure as described above. The choice of residue positions, t, for removing the assigned SSY is guided by the associated local total residue energies, $E'_{tot}(t)$ (Eq. 11), more frequently removing systems with bad corresponding energies by using the normal-variate-rank selection scheme.

Recombination of solutions The next step is to recombine pairs of candidate RAs (parents) to form a new RA (child). First the two parents, p and q, are chosen randomly from the population. Let $p_1, p_2, \ldots, p_N$ denote the assigned SSYs (value of the genes) for parent p where SSY, p_i is assigned to residue position, i, and similarly $q_1, q_2, \ldots, q_N$ the genes for parent q. After recombination the value of the gene for the child in residue position i can be either $c_i = p_i$ or $c_i = q_i$. The genes are taken from the two parents, p and q, to generate children according to an algorithm adding genes from N-term to C-term for each position switching between adding genes from a fixed parent with probability

$p_{switch} = 0.12$ used here. This procedure leads to the concatenation of consecutive stretches of genes p_j, p_{j+1}, …, p_{j+n}, and q_i, q_{i+1}, …, q_{i+m}, from the two parents to form a mixed set of genes. If the same SSY is found multiple times for the child genes, then these positions are randomly removed until each SSY is represented only once. In addition, more SSYs are randomly removed until ca. 80 % is left. Finally, the child RA is rebuilt as in the mutation step described above using the principles of linking and typing.

A large number of recombinations (2000 used here) produces a large group of new solutions (large group of children) The children compete with the parent solution keeping the best, i.e. each time a child is generated with energy, E_C, and the highest energy of the parents is E_P then the child replaces this parent in the population immediately if $E_C < E_P$ or if, using a soft Boltzmann criterion:

$$e^{(E_P - E_C)/T_{RA}} > p_{rand} \tag{12}$$

where p_{rand} is a random number drawn between 0 and 1 and T_{RA} is the "temperature" defining the scaling of the difference ($T_{RA} = 0.05$ used here throughout). Following the evolution of the population, only the top one ranked RA is kept for the next phase.

Analysis of results

After completion of the above RA each resonance in the SSY is revisited. If a non-specifically assigned resonance could still not be mapped to a specific atom, judged by whether the value, $\delta_{j,obs}$, is far from the nearest reference, $\delta_{k,ref}$, defined as $\frac{1}{2}\left(\left(\delta_{j,obs} - \delta_{k,ref}\right)/\sigma_k\right)^2 > e_{\max}$, this resonance is removed from the SSY and the peaks related to that resonance are removed from the aligned peak cluster as well. In contrast, for all resonances, which could be assigned specifically, for each peak in SSY related to the resonance the assigned residue position and assigned atom name is stored. This information serves as statistics to be used in the next phase. Since the RA was built from incomplete SSYs, not all of the protein resonances would be assigned in a given run. To remedy this problem, multiple RAs are generated in parallel, each time with a different set of generated (incomplete) SSYs from phase I, (400 repetitions were used here, the 64 RAs with lowest energy were used in the next phase). The reason for using this procedure is that all resonances would be assigned at least once.

Phase III: consensus resonance assignments

Based on the assignment data generated in phases I and II, each peak will be assigned to a certain residue position and atom in the sequence a different number of times. A histogram is generated for each peak with frequencies of residue position assignments: $f_1, f_2, \ldots, f_N$ with $\sum_{i=1}^{N} f_i = 1$ where most f_i are typically 0, and for a unique assignment one of them is 1. The position in the histogram with the highest frequency is considered the most likely assignment for the given peak. In this final phase, the RAs are reconstructed from scratch based on the individual peaks but this time using the histogram statistics to guide the assignments (consensus assignments).

Algorithm 3.1: reconstruction of resonance assignments

The reconstruction algorithm starts by for each peak trying to pair the peak with a sequence specific assignment. The choice of peak selection and pairing is guided by the MENOVAR principle as described above. The frequencies, f_i, are already normalized probabilities and are therefore used directly to calculate the chosen entropy and for the roulette wheel selection. If a peak is paired with a residue position for the first time a SSY is initialized and assigned to this position. If a SSY already exists for this position, the peak is added to the SSY provided that the shared resonances are consistent and this is quantified in the resonance-coordinate match energy: $E_c < 2.0$ (where E_c is defined as in Eq. 2). This time the chemical shift value for the related dimension of the peak and the resonance in the SSY are matched (if one of the axes for the peak is not shared with a value in the SSY, the contribution for this dimension is 0).

Next two steps aims at completing the assignments by assigning missing resonances in SSYs (A3.2 Supplementary Material) and missing atoms in the protein (A3.3 Supplementary Material).

Iterative repetition of rebuilding

Following the above steps 3.1–3.3, the assignments are rebuilt from consensus statistics producing a new set of candidate resonances assignments (400 here). By an iterative procedure, the output RAs are used as input for the statistics repeating the rebuild algorithm cycling 4 times here.

Algorithm 3.4: recombination of resonance assignments

Finally, after completing the last cycle, the best candidates RAs are combined to generate better solutions. This is done, as described in Algorithm 2.2, using a GA considering the sequence specifically assigned SSYs as genes. The scheme here is somewhat different from a standard GA. A sub-population (20 members here) is chosen from the full population by the normal-variate-rank selection scheme. This population is allowed to evolve by breeding children (1,000 steps here) where the child replaces its weakest parent in the sub-population if the energy is lower or accepted by a Boltzmann criterion (Eq. 12). By the end of the evolution, the very fittest individual is copied to a "super-population". This procedure is repeated to produce a super-population of 20 candidates. The recombination process, which is used here to breed child RAs builds the new RA sequentially. Starting from the N-terminal, a gene is inserted from one of the parents and the initialization the sequence is looped through each time evaluating if the residue based energy, $E'_{tot}(\tau,t)$ (Eq. 11), for position t, when switching to inserting the next gene from the other parent is significantly better than for the current: $E'_{tot}(\tau_{\mathrm{current}},t) - E'_{tot}(\tau_{\mathrm{other}},t) > E_{min}$ (or slightly worse using a Boltzmann soft evaluation, see Eq. 12). The threshold, E_{min}, is linearly decreased for each new step (here from 0.5 in steps of 0.1) to bias selection towards connected patches of SSYs from the same parent to improve the linking between neighboring SSYs.

Following the steps in phases I–III presented so far, the best-ranked candidate RAs, i.e. the lowest energy RAs, are kept (20 here). This multi-step procedure is run independently in parallel (here 32 calculations on multiple processors). Finally, these (20*32) solutions are recombined as in 3.4 keeping finally the best 20 RAs. The full algorithm including all three phases is summarized in Fig. 1.

Statistics on results

The final 20 RAs represent an ensemble of solutions. A final single value for a given resonance in the protein is obtained by binning the ensemble solution into a histogram and selecting the bin with the highest count. This value is referred to here as the *median assignment* or, in brief, just as *the assignment*. The standard deviation from the median assignment is a measure of the precision and the reliability of the assignment (the smaller the better). A phenomenological figure of merit, Q_a, for the validation of the assignment is calculated based on this number and other quantities (see Supplementary Information).

Results

Our method, GAMES_ASSIGN, for automatic assignment of solid state NMR resonances was evaluated on three different biologically relevant protein systems. These systems represent proteins of different sizes and varying quality of data, which are realistic for biologically relevant systems and standard solid state NMR instrumentation.

Performance on three proteins

The first is the immunoglobulin binding domain B1 of streptococcal protein G (GB1), which has been studied intensively and frequently used as a benchmarking protein in structural biology (Bouvignies et al. 2006; Gallagher et al. 1994). GB1 consists of 56 amino acids with a globular fold containing both an alpha helix and four beta-sheets. We analyze a microcrystalline sample of high quality with a homogenous line width of ca. 0.5 ppm measured as the full width at half maximum intensity (FWHM). The second protein is the 76 residue Ubiquitin (Zech et al. 2005; Igumenova et al. 2004; Vijaykumar et al. 1987), which is a regulatory protein, performing its myriad functions through conjugation to a large range of target proteins. Ubiquitin has a mixed secondary structural with an alpha-helix, a short piece of 3(10)-helix, a mixed beta-sheet that contains five strands, and seven reverse turns. This sample has a larger line width of ca. 0.7 ppm on average. The third sample is a more challenging heterogeneous system containing the 59 amino acid protein CsmA in a biological mixed matrix formed with both lipids and the small molecule cofactors, carotenoids and BChl *a* (Kulminskaya et al. 2012). The protein resonances have an average line width of ca. 0.8 ppm. CsmA is a small, pigment-binding antenna protein from phototrophic green sulfur bacteria, which has been found to play a fundamental role in energy transfer of those organisms (Frigaard et al. 2005; Pedersen et al. 2008). Due to the small relative amount of the protein (CsmA) of interest in this sample, the sensitivity is quite much lower. Published complete assignments exist for the two microcrystalline GB1 (Franks et al. 2005) and Ubiquitin (Igumenova et al. 2004) samples whereas almost complete assignments exist for CsmA (Kulminskaya et al. 2012). The resonances for all three systems were also assigned manually using standard procedures to compare with the automatic assignments. Standard ssNMR experiments were conducted on an Advance 700 MHz Bruker spectrometer (Rheistetten, Germany) equipped with standard triple resonance 2.5 mm and 4 mm probes. These include 2*D* and 3*D* homonuclear ^{13}C DARR experiments and 3D heteronuclear NCACX, NCOCX and CONCA experiments, all details can be found in supplementary information. All peaks were picked manually for all three targets and by an automatic procedure for GB1 as well. For GB1, most of the peak lists are relatively complete and most peaks can be assigned, the automatically picked data is of lesser quality and CsmA and Ubiquitin have even less optimal peak lists. The quality of the all input peak lists is summarized in Table 1. Further details related to the experiments and the peak picking are described in the experimental procedures section in the supplementary material. Representative

Table 1 Quality of experimental peak picked data

	Completeness	Assignable	rms (ppm)
GB1			
NCOCX	81.3 % (63.2 %)[a]	98.2 % (91.8 %)	0.24 (0.25)
CONCA	92.6 % (96.3 %)	100.0 % (85.0 %)	0.17 (0.19)
DARR 2D[b]	90.7 % (90.7 %)	75.5 % (62.1 %)	0.16 (0.16)
NCACX	92.6 % (82.6 %)	98.6 % (79.2 %)	0.16 (0.16)
Average	89.3 % (83.2 %)	93.1 % (79.5 %)	0.18 (0.19)
CsmA			
NCOCX	63.6 %	87.5 %	0.19
CONCA	78.0 %	83.3 %	0.34
NCACX	93.3 %	93.8 %	0.16
Average	78.3 %	88.2 %	0.23
Ubiquitin			
NCOCX	73.6 %	82.5 %	0.29
CONCA	43.5 %	50.0 %	0.58
CAN(CO)CX	28.1 %	15.0 %	0.34
DARR 3D	5.3 %	26.8 %	0.53
DARR 2D	67.3 %	79.3 %	0.33
NCACX	79.4 %	84.4 %	0.16
Average	49.5 %	56.3 %	0.37

[a] Values in brackets are for automatically picked peaks, and all other values correspond to manually picked peaks

[b] DARR 2D and 3D produces cross peaks between ^{13}C atoms close in space

Completeness: The fraction of the theoretical peaks that have a matched observed peak. The theoretical peaks are defined as all possible peaks when including all intra-residue ^{13}C–^{13}C transfers (to CX) within a distance cut-off provided there exists a manual RA for this carbon. Except for the DARR spectra where only aliphatic carbons were considered in the indirect dimensions, and excluding all aromatic carbons in all spectra except for 2D DARR. Cut-offs were set, according to distances measured in corresponding X-ray structures, to 3.6 and 5.5 Å, respectively, for first and last transfers in 3D DARR and 4.5 Å for 2D DARR, NCACX, NCOCX and CAN(CO)CX. A matched observed peak is defined as a peak with a related theoretical peak with a chemical shift deviation less than 0.55 ppm in each dimension

Assignable: The fraction of observed peaks that can be matched to at least one theoretical peak as defined above

rms: Root mean square deviation on average for all matched peaks

examples of the spectra with picked peaks are shown in Figure S1 in the supplementary material and in Kulminskaya et al. (2012).

The peak list data were used together with the primary sequence as input for the automatic assignment. C-terminal parts of CsmA (residues 53–59) and Ubiquitin (residues 71–76) that are known to be flexible and hence not observable with ssNMR were excluded from the calculations. Furthermore, the flexible loop in residues 7–10 in Ubiquitin was also removed from the target input sequence. Often information about disordered regions is available from other biophysical techniques (Konermann et al. 2011; Vilar et al. 2012; Alexandrescu 2001), or can be predicted with high confidence from the sequence (He et al. 2009; Coeytaux and Poupon 2005), or can be derived by in an iterative approach by running the automatic assignments and identifying regions with low estimated precision.

The automatic assignments were compared with the manual ones. An automatic assignment was considered correct if the error was less than 0.55 ppm. Assignments with higher estimated precision, Q_a, are considered more reliable (see "Methods" section). We categorize an assignment as *validated,* when $Q_a > 80$, other assignments should be considered as tentative, but are often correct. Almost all assignments (98.8 %) were validated for GB1 (numbers are for the manually picked peaks where nothing else is stated), whereas fewer are validated for Ubiquitin (63.1 %) and CsmA (48.8 %). Generally, we find that assignments for backbone atoms, N, Cα and C′, are frequently correct compared with the side chain atoms (and are more often validated). We get a fraction of correct assignments of 100–85.5 % for backbone and 91.8–74.2 % for all atoms. The results are summarized in Table 2 and visualized in Fig. 6 and described in more detail below.

For GB1 all the automatic assignments for the backbone atoms are correct and validated. This is a remarkable result, given that only a few standard experiments were used for the assignments. For the side chain atoms, 21 (17.9 %) are assigned with an error > 0.55 ppm, most of these atoms are aromatic carbons and only four of them are Cβ. For the

Table 2 Performance of GAMES_ASSIGN for automatic assignments

	GB1		CsmA		Ubiquitin	
	All	Validated[a]	All	Validated	All	Validated
Backbone[b]	100.0 % (98.8 %)[c]	100.0 % (98.7 %)	80.6 %	92.0 %	85.5 %	93.8 %
Side chain[d]	82.1 % (74.0 %)	84.6 % (75.0 %)	89.5 %	87.0 %	60.2 %	81.6 %
All	91.8 % (87.0 %)	93.0 % (87.5 %)	83.1 %	90.8 %	74.2 %	88.8 %

Results for backbone and side chain, validated and tentative

[a] $Q_a > 80$, see Supplementary section

[b] N, Cα and C′ atoms

[c] Values in brackets are for automatically picked peaks, other values are for manually picked peaks

[d] Including Cβ atoms

automatically picked peaks data the performance is also very good with 98.8 % correct backbone assignments and 87.0 % correct for all resonances, demonstrating that GAMES_ASSIGN is robust towards data with a higher content of missing expected peaks and noise peaks. For CsmA, the lines widths are larger and the sensitivity is lower. Furthermore, due to overlap with resonances from the non-protein cofactors in the sample, it was not possible to obtain good data for standard ^{13}C–^{13}C DARR homonuclear correlation spectra (and other similar techniques), except for the Cα's, which resonate in disjoint regions. Due to this challenge, some of the side chain carbons, in particular non-aliphatic, could not be assigned neither by manual nor automatic methods. However, even for this challenging system with a limited set of experiments, we still get a reasonable success for the automatic assignments yielding 83.1 % correct assignments. This number increases to 90.8 % if considering only the validated assignments. Ubiquitin is the largest system considered here and shows a larger line width of ca. 0.7 ppm compared to GB1 (ca. 0.5 ppm). Still 85.5 % of the backbone assignments get assigned correctly whereas 74.2 % are correct when considering all atoms. These numbers increases to 93.8 % and 88.8 % correct for backbone and all atoms, respectively, for the validated assignments only.

To summarize, GAMES_ASSIGN derives the automatic assignments with near perfect success (with 100 % correct backbone assignments) for the micro-crystalline sample of the small protein GB1 with narrow lines, and with considerable success for the more challenging systems having larger line widths and/or more residues.

It is not possible to deconvolute the effects of the sample resonance line width, sensitivity, and the protein size. However, we demonstrate below that GAMES_ASSIGN can handle significant signal overlap by generating different systematically varied sets of simulated peak lists for GB1 using our program VirtualSpectrum (Nielsen and Nielsen 2014) and evaluating the assignment performance. By this analysis we address quantitatively how the line width influences the accuracy on the assignment and how adding more (advanced) experiments also would improve the accuracy. A side benefit from using simulated peaks lists based on published assignments is that the "true assignments" are known completely free of human interpretation.

Performance dependence with increasing resonance line widths and number of experiments

The program, VirtualSpectrum, was used to simulate ssNMR spectra for GB1 for NCACX, NCACO, NCOCX, CONCA and $2D$ ^{13}C–^{13}C DARR using different uniform ^{13}C and ^{15}N line widths. The peaks were simulated with Gaussian signal shape, s:

$$s(x, \mu, \sigma) = A \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{13}$$

where A is the intensity of the peak, μ is the location parameter, and the scaling parameter is varied by $\sigma = 0.3$, 0.4, 0.5, 0.7, and 1.0 ppm for both ^{13}C and ^{15}N in all dimensions corresponding to FWHM $= 2.355\sigma$. All other parameters are as described in (Nielsen and Nielsen 2014). An excerpt from the DARR spectrum is shown in Figure S2 (Supplementary Information). It is observed that an increasingly amount of peaks overlap and as a result increasingly fewer peaks are found in the simulated peak lists for larger line widths (see Fig. 7a, b). In the peak lists derived from a line width of $\sigma = 0.3$ ppm the total number of peaks is 760 and only two of the peaks from the backbone experiments are overlapped, whereas in contrast, only a total of 385 peaks are observed in the peak lists derived with a line width of $\sigma = 1.0$ ppm. Therefore, almost half of the peaks cannot be assigned uniquely and have distorted positions due to signal merging. Along with the increasing line width, the number of incorrectly assigned resonances also increases from 2 and 26 (out of 165 and 159 total) for backbone and side chain, respectively, to 75 and 56, corresponding to line widths of $\sigma = 0.3$ ppm and 1.0 (Fig. 7c). Here an incorrect

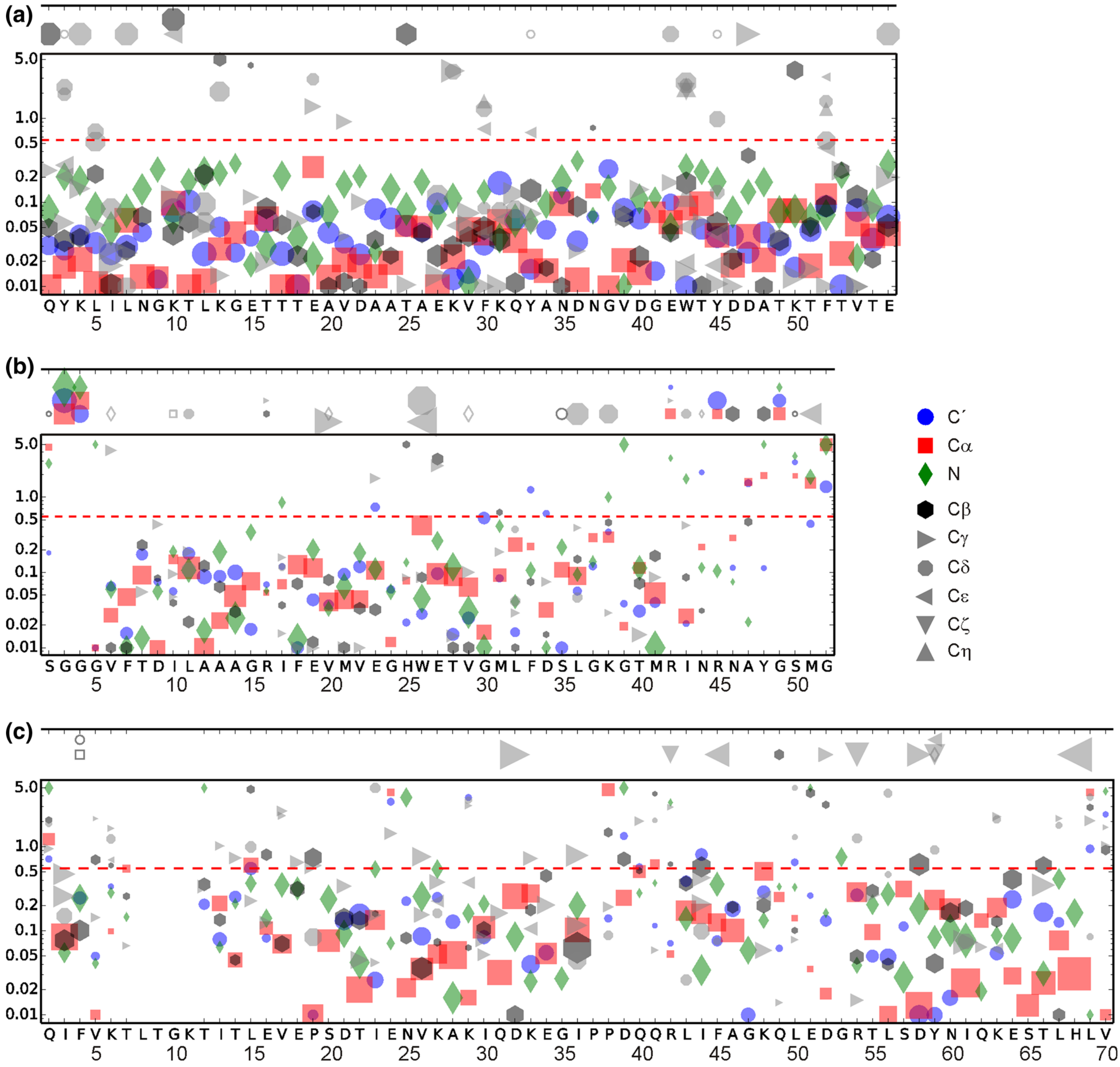

Fig. 6 Visualization of the assignment error as a function of residue position in the sequence. **a–c** shows results for manually peak picked data for GB1, CsmA and Ubiquitin, respectively. The absolute value of the difference between the manual and automatic assignments is shown on a logarithmic scale. The primary sequence is shown below the chart for reference. Different atom types are shown with markers with different *symbol* and *shape* (see symbolic legend to the *right*), side chain carbon atoms are shown in *gray* and *black* in case of Cβ. The size of the marker is proportional to estimated confidence of the assignment. A reference maximum acceptable error is highlighted with a *red broken line*. The errors are clipped to a maximum of 5.0 ppm and a minimum of 0.01 ppm for increased readability. Missing assignments are shown above the chart, assignment present in the automatic but missing in the manual assignments shown with *filled markers*, whereas those present in the manual assignments but missing in the automatic are shown with *open markers*. The missing *x*-ticks in panel c for residues 7–10, 71–76 are due to the absence of manually assigned resonances for this region

assignment is defined to have an error larger than the line width (i.e. more generous definition with increasing line width). In between these two extremes, we find that when increasing the line width to $\sigma = 0.5$ ppm, corresponding to FWHM = 1.18 ppm, leads to 19 % fewer peaks observed due to overlapping, and GAMES_ASSIGN still performs very well by assigning 97 % of the backbone atoms (within 0.5 ppm) and 78 % of all atoms correctly.

To analyze whether the addition of more (advanced) experiments increases the accuracy of the assignments for a challenging case with a large line width, virtual peak lists were generated for a line width of $\sigma = 1.0$ ppm,

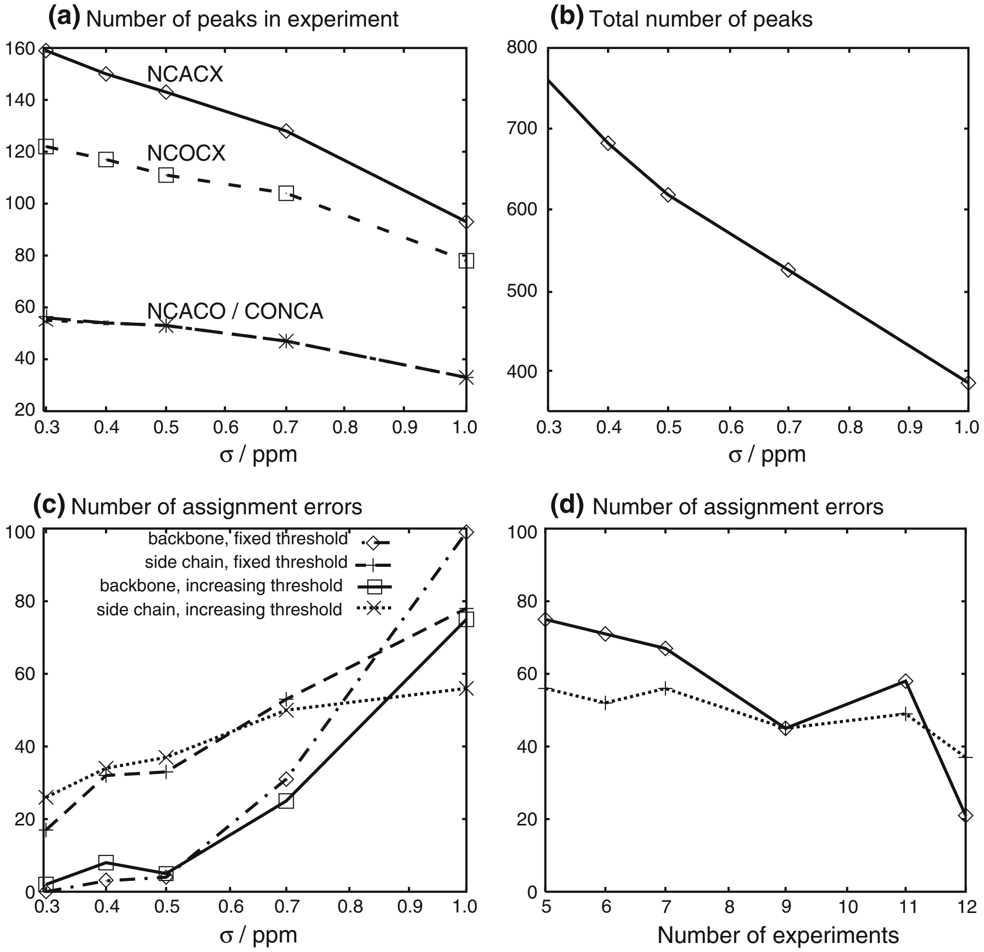

Fig. 7 Number of peaks and performance using simulated data with various line widths and data sets. **a**, **b** number of peaks as a function of line width, σ, corresponding to FWHM $= 2.355\sigma$ in different experiments (**a**) and total (**b**). **c** Number of errors (N_{err}) as a function of line width; judged by a fixed threshold of 0.5 ppm shown with *long dashes* and *dash-dot line* for side chain and backbone atoms, respectively. N_{err} using a threshold equal to the line width is shown with a *full* and *dotted line* for backbone and side chain atoms, respectively. **d** N_{err} as a function of the number of experiments (N_{exp}) with a fixed threshold for an error set to 1.0 ppm. All and backbone errors are shown with *dotted* and *full lines*, respectively, as in **c**. The experiments are appended to the extended set of experiments in the order of 3*D* DARR (CCC) ($N_{exp} = 6$), NCACB ($N_{exp} = 7$), CAN(CO)CX + N(CO)CACX ($N_{exp} = 9$), CON(CA)CX + N(CA)COCX ($N_{exp} = 11$), 4*D* CANCOCA ($N_{exp} = 12$)

corresponding to FWHM = 2.355 ppm for different types of spectra. Each experiment with its peak list was appended to the assignment data one by one, or two at the same time, and the performance of the assignments were then evaluated for the cases of 5, 6, 7, 9, 11 and 12 experiments. Not surprisingly, the accuracy of the assignment increases with the number of experiments; the number of errors and the type of experiments added are displayed in Fig. 7d. Adding the CCC (3D DARR) in the first experiment addition increases the dimensionality of the carbon side chain correlations and therefore the number of assignment errors decreases. Adding NCACB has less effect, whereas again adding two experiments with new observations of the Cα resonances, the N(CO)CACX and CAN(CO)CX experiments with relayed transfer through C′, leads to a substantial decrease in the number of assignment errors both for backbone and side chain. In contrast, adding the "opposite" experiments with observation of C′ and relayed transfer through Cα, N(CA)COCX and CON(CA)CX, actually leads to a small increase in the number of assignment errors. This shows that generally the Cα chemical shift is more useful for the RA process, and we argue that this is due to the better dispersion and the more unique amino acid typing information for Cα (Ulrich et al. 2008; Yao et al. 1997). Finally, adding a 4D CANCOCA spectrum, again leads to a dramatic improvement of the assignments arriving at 21 and 37 errors for backbone and side chain resonances, respectively, when using all 12 experiments compared to 75 and 56 errors when using only 5 standard experiments. In this 4D CANCOCA spectrum only 5 out of the 51 peaks in the peak list represent overlapping peaks compared to 16 out of 33 for 3D CONCA, suggesting that increasing the dimensionality of the spectra to resolve the overlap is very helpful for assisting RAs in cases with large line widths.

Discussion

We have demonstrated here that GAMES_ASSIGN can automatically assign the resonances for proteins with average quality ssNMR experimental data using standard ssNMR experiments. By this simplistic approach, the assignments can be derived fast without the need for applying special selective labeling techniques, implementation of advanced pulse sequences and without relying on a very high field. GAMES_ASSIGN showed an excellent performance for GB1 assigning all backbone resonance correctly and 91.8 % when including also side chains, the accuracy only drops slightly when using automatically picked peaks (see Table 2). The data for GB1 is of excellent quality, which is difficult to obtain for most cases, however, we have shown that GAMES_ASSIGN still performs reasonable for more challenging systems, such as the heterogeneous baseplate systems, CsmA and the larger protein Ubiquitin (see Table 2). A more reliable assignment can be derived by trusting the assignments considered as validated by GAMES_ASSIGN, while doing a manually inspection of the parts of spectra corresponding to the other assignments not considered as validated. Although the performance of GAMES_ASSIGN in the present form is very good in most cases, we imagine that a few things could potentially improve the assignments. By using predictions for the secondary structure, it would be possible to derive more specialized probability distributions for the chemical shifts, which would probably assist in the spin system typing process. GAMES_ASSIGN uses peak lists as input and, in particular, information about the raw data such as line width, peak shape, and noise level is neglected. Including this information could also possibly improve the performance.

The analysis presented here shows that GAMES_ASSIGN can handle a significant amount of overlapping peaks. Increasing the line width up to 1.18 ppm at FWHM did no significantly affect the assignment accuracy with 97 % backbone peaks still correctly assigned. This is a very promising result since in biological samples, such as fibrils or in heterogeneous systems, line widths are often large due to local heterogeneity of the structure. When further increasing line width the assignment process is more prone to errors, but by adding 3D spectra with complementary combinations of resonance observations and/or adding data with higher dimensionality, GAMES_ASSIGN is also capable of deriving an almost error-free assignment. Unfortunately, the systems with relatively large line widths often also have lower sensitivity and thus acquiring data with higher dimensionally may be impractical. High dimensional data could be acquired faster by applying techniques for high-throughput acquisition where fewer sampling points are needed (Rovnyak et al. 2004; Kupce and Freeman 2003; Kim and Szyperski 2003). Another solution could be to use proton detection to enhance sensitivity, which would, however, typically require special ^{2}H labeling procedures (Chevelkov et al. 2006; Zhou et al. 2012) and/or ultra-fast MAS spinning (Holland et al. 2010). It is possible, through the flexible implementation of GAMES_ASSIGN, to include data from experiments derived with the two above techniques.

Conclusion

We have presented the method, GAMES_ASSIGN that assigns the resonances for proteins automatically using peak list from various ssNMR experiments. Our method is sufficiently robust to handle average quality ssNMR data, which often suffer from overlapping and missing signals. A stochastic approach is applied to derive the assignments using concepts of variable pairing (paring peaks, spin systems, and residues) based on the MENOVAR scheme, which relies on the ideas of minimum entropy choice and normal-variate rank selection. Candidate solutions are combined to form optimized solutions using a GA.

Our method was tested on three different proteins and shows very impressive result with 100 % correct backbone assignments and 91.8 % correct assignment for all atoms for GB1 and the performance is 80.6/83.1 % and 85.5/74.2 %, respectively for backbone/all resonances for the more challenging CsmA baseplate and Ubiquitin systems.

We demonstrated that GAMES_ASSIGN could handle a significant amount of overlapping peaks by analyzing simulated spectra with increasing line widths. For very large line widths, the accuracy of the RAs is lower, but it was shown here also, that these problems due to overlap can be resolved by adding more dimensions to the spectra, if possible, and/or by acquiring more experiments with complementary data—preferentially providing chemical shift information for Cα rather than for C′.

We envisage that GAMES_ASSIGN will be applied in the analysis of ssNMR data to save valuable human resource time and be used as a tool in the pipeline to solve challenging protein structures.

Acknowledgments We acknowledge Prof. Dr. Chad Rienstra for kindly providing a microcrystalline sample of uniformly ^{13}C, ^{15}N labeled GB1.

References

Alexandrescu AT (2001) An NMR-based quenched hydrogen exchange investigation of model amyloid fibrils formed by cold shock protein A. Pac Symp Biocomput 6:67–78

Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. Curr Opin Struct Biol 14(5):547–553

Atreya HS, Sahu SC, Chary KVR, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17(2):125–136

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104(8):3541–3555

Bartels C, Billeter M, Guntert P, Wuthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. J Biomol NMR 7(3):207–213

Bartels C, Guntert P, Billeter M, Wuthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18(1):139–149

Bouvignies G, Meier S, Grzesiek S, Blackledge M (2006) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. Angew Chem Int Ed 45(48):8166–8169

Cela E (1998) The quadratic assignment problem. Theory and Algorithms. Kluwer Academic Publishers, Dordrecht

Chevelkov V, Rehbein K, Diehl A, Reif B (2006) Ultrahigh resolution in proton solid-state NMR spectroscopy at high levels of deuteration. Angew Chem Int Ed 45(23):3878–3881

Coeytaux K, Poupon A (2005) Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 21(9):1891–1900

Coggins BE, Zhou P (2003) PACES: protein sequential assignment by computer-assisted exhaustive search. J Biomol NMR 26(2):93–111

Crippen GM, Rousaki A, Revington M, Zhang YB, Zuiderweg ERP (2010) SAGA: rapid automatic mainchain NMR assignment for large proteins. J Biomol NMR 46(4):281–298

Eghbalnia HR, Bahrami A, Wang LY, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). J Biomol NMR 32(3):219–233

Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) NMR analysis of a 900 K GroEL-GroES complex. Nature 418(6894):207–211

Franks WT, Zhou DH, Wylie BJ, Money BG, Graesser DT, Frericks HL, Sahota G, Rienstra CM (2005) Magic-angle spinning solid-state NMR spectroscopy of the beta 1 immunoglobulin binding domain of protein G (GB1): N-15 and C-13 chemical shift assignments and conformational analysis. J Am Chem Soc 127(35):12291–12305

Frigaard NU, Li H, Martinsson P, Das SK, Frank HA, Aartsma TJ, Bryant DA (2005) Isolation and characterization of carotenosomes from a bacteriochlorophyll c-less mutant of *Chlorobium tepidum*. Photosynth Res 86(1–2):101–111

Gallagher T, Alexander P, Bryan P, Gilliland GL (1994) 2 crystal-structures of the B1 immunoglobulin-binding domain of streptococcal protein-G and comparison with nmr. Biochemistry 33(15):4721–4729

Gath J, Habenstein B, Bousset L, Melki R, Meier BH, Boeckmann A (2012) Solid-state NMR sequential assignments of alpha-synuclein. Biomol NMR Assigm 6(1):51–55

Griswold IJ, Dahlquist FW (2002) Bigger is better: megadalton protein NMR in solution. Nat Struct Biol 9(8):567–568

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Quart Rev Biophys 44(3):257–309

Habenstein B, Wasmer C, Bousset L, Sourigues Y, Schuetz A, Loquet A, Meier BH, Melki R, Boeckmann A (2011) Extensive de novo solid-state NMR assignments of the 33 kDa C-terminal domain of the Ure2 prion. J Biomol NMR 51(3):235–243

He B, Wang KJ, Liu YL, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19(8):929–949

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):1–9

Holland GP, Cherry BR, Jenkins JE, Yarger JL (2010) Proton-detected heteronuclear single quantum correlation NMR spectroscopy in rigid solids with ultra-fast MAS. J Magn Reson 202(1):64–71

Hu K-N, 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(3):267–276

Igumenova TI, McDermott AE, Zilm KW, Martin RW, Paulson EK, Wand AJ (2004) Assignments of carbon NMR resonances for microcrystalline ubiquitin. J Am Chem Soc 126(21):6720–6727

Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30(1):11–23

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125(5):1385–1393

Konermann L, Pan JX, Liu YH (2011) Hydrogen exchange mass spectrometry for studying protein structure and dynamics. Chem Soc Rev 40(3):1224–1234

Kulminskaya NV, Pedersen MO, Bjerring M, Underhaug J, Miller M, Frigaard N-U, Nielsen JT, Nielsen NC (2012) In situ solid-state NMR spectroscopy of protein in heterogeneous membranes: the baseplate antenna complex of Chlorobaculum tepidum. Angew Chem Int Ed 51(28):6891–6895

Kupce E, Freeman R (2003) Fast multi-dimensional NMR of proteins. J Biomol NMR 25(4):349–354

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11(1):31–43

Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (C-13, N-15)-labeled proteins. J Biomol NMR 9(2):151–166

Malmodin D, Papavoine CHM, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J Biomol NMR 27(1):69–79

Moseley HNB, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9(5):635–642

Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Nucl Magn Reson Biol Macromol Pt B 339:91–108

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(3):123–128

Nagarajan V, Sviridenko M (2009) On the maximum quadratic assignment problem. Math Oper Res 34(4):859–868

Nielsen JT, Nielsen NC (2014) VirtualSpectrum, a tool for simulating realistic peak list for multi-dimensional NMR spectra. Submitted

Nielsen JT, Eghbalnia HR, Nielsen NC (2012) Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. Progr Nuc Magn Reson Spectrosc 60:1–28

Pedersen MØ, Underhaug J, Dittmer J, Miller M, Nielsen NC (2008) The three-dimensional structure of CsmA: a small antenna protein from the green sulfur bacterium *Chlorobium tepidum*. FEBS Lett 582(19):2869–2874

Rovnyak D, Frueh DP, Sastry M, Sun ZYJ, 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(1):15–21

Schmidt E, Guntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134(30):12817–12829

Schmidt E, Gath J, Habenstein B, Ravotti F, Szekely K, Huber M, Buchner L, Boeckmann A, Meier BH, Guentert P (2013) Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. J Biomol NMR 56(3):243–254

Schmucki R, Yokoyama S, Guentert P (2009) Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. J Biomol NMR 43(2):97–109

Tang KS, Man KF, Kwong S, He Q (1996) Genetic algorithms and their applications. IEEE Signal Process Mag 13(6):22–37

Tycko R, Hu K-N (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(2):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. Nucl Acids Res 36:D402–D408

Vijaykumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 A resolution. J Mol Biol 194(3):531–544

Vilar M, Wang L, Riek R (2012) Structural Studies of Amyloids by Quenched Hydrogen-Deuterium Exchange by NMR. In: Sigurdsson EM, Calero M, Gasset M (eds) Amyloid Proteins: Methods and Protocols, 2ed. 849. Methods Mol Biol 1:pp 185–198

Xu Y, Zheng Y, Fan J-S, Yang D (2006) A new strategy for structure determination of large proteins in solution without deuteration. Nat Methods 3(11):931–937

Yao J, Dyson HJ, Wright PE (1997) Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. FEBS Lett 419(2–3):285–289

Zech SG, Wand AJ, McDermott AE (2005) Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. J Am Chem Soc 127(24):8618–8626

Zhang HY, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biom NMR 25(3):173–195

Zhou DHH, Nieuwkoop AJ, Berthold DA, Comellas G, Sperling LJ, Tang M, Shah GJ, Brea EJ, Lemkau LR, Rienstra CM (2012) Solid-state NMR analysis of membrane proteins and protein aggregates by proton detected spectroscopy. J Biomol NMR 54:291

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(4):592–610