J Biomol NMR (2015) 62:143–156
DOI 10.1007/s10858-015-9926-7

ARTICLE

# Exploiting image registration for automated resonance assignment in NMR

Madeleine Strickland[1] · Thomas Stephens[1] · Jian Liu[1] · Nico Tjandra[1]

Received: 2 February 2015 / Accepted: 24 March 2015 / Published online: 1 April 2015
© Springer Science+Business Media Dordrecht (outside the USA) 2015

**Abstract** Analysis of protein NMR data involves the assignment of resonance peaks in a number of multidimensional data sets. To establish resonance assignment a three-dimensional search is used to match a pair of common variables, such as chemical shifts of the same spin system, in different NMR spectra. We show that by displaying the variables to be compared in two-dimensional plots the process can be simplified. Moreover, by utilizing a fast Fourier transform cross-correlation algorithm, more common to the field of image registration or pattern matching, we can automate this process. Here, we use sequential NMR backbone assignment as an example to show that the combination of correlation plots and segmented pattern matching establishes fast backbone assignment in fifteen proteins of varying sizes. For example, the 265-residue RalBP1 protein was 95.4 % correctly assigned in 10 s. The same concept can be applied to any multidimensional NMR data set where analysis comprises the comparison of two variables. This modular and robust approach offers high efficiency with excellent computational scalability and could be easily incorporated into existing assignment software.

**Keywords** Automation · Backbone assignment · Fast Fourier transform · Cross-correlation · FFT

## Introduction

Solving the structure of a protein by NMR spectroscopy is a complicated task that requires analysis of multidimensional spectra, many of which contain hundreds, if not thousands, of peaks. A great deal of this process involves recognizing patterns in the data that are sometimes not obvious, for example, linking residues in backbone resonance assignment, side-chain resonance assignment, and analysis of NOESY data for generation of distance restraints. The relevant information is usually extracted from a combination of multiple NMR data sets that might be acquired at different times and on different spectrometers. Therefore, this process is sensitive to experimental variations. For instance, backbone assignment typically involves matching a pair of NMR resonances within one-dimensional strips of at least two 3D spectra. Variations along this dimension in the different experiments will make it difficult to create perfect matches. In addition, resonance matching is completed for each pair of residues in a sequential manner. Consequently when the protein size increases, this process becomes extremely slow due to the number of possible matches that need to be considered. For these reasons, the overall process of NMR resonance identification is a bottleneck in NMR structural studies.

Substantial efforts are underway to automate some or all stages of NMR resonance assignment for use in high-throughput structure analysis (Dutta et al. 2014; Linge et al. 2003; López-Méndez and Güntert 2006; Rieping et al. 2007; Volk et al. 2008, for a review see Güntert 2009; Moseley and Montelione 1999; Williamson and Craven

Madeleine Strickland and Thomas Stephens have contributed equally to this work.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-015-9926-7) contains supplementary material, which is available to authorized users.

✉ Nico Tjandra
tjandran@nhlbi.nih.gov

[1] Laboratory of Molecular Biophysics, National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (NIH), Building 50, Room 3503, Bethesda, MD 20892, USA

2009). For protein backbone assignment, there are a collection of strategies based on optimization of fit between measured data and external or predicted information, for example the programs Mars (Jung and Zweckstetter 2004), AutoAssign (Zimmerman et al. 1993), PINE (Bahrami et al. 2009), GARANT (Bartels et al. 1997), and the algorithm by Lukin (1997). Each of these methods offers different advantages and scalability. In general, as the protein gets larger, the performance of these methods is severely reduced. Approaches that aim to reduce the search space of this problem have included the exploitation of geometric relationships that exist within correlated NMR data sets (Borkar et al. 2011; Chen et al. 2010; Hiller et al. 2005) or connectivity matrices that can rapidly exclude poor linkage candidates (Xu et al. 2006). It is important to note that improvements in data quality can increase the chances of successful automated assignment, for example by using global alignment as a pre-processor (Buchner et al. 2013).

Interestingly, we often superimpose multiple 2D-spectra to visually identify similarities or differences between them. Global matching has the advantage that the time it takes to solve the problem scales more favorably as the system gets larger in comparison to the conventional approach to backbone assignment, which is local and sequential. Humans are excellent at pattern recognition and often outperform computers, which accounts for the high amount of manual intervention necessary in analyzing densely populated NMR data sets. In crowded regions, local patterns can provide excellent initial guesses for data matches that can be subsequently validated with additional NMR data. We set out to replicate this natural human ability by borrowing from the field of image registration. The key concept that we are proposing is to display the two variables obtained from an NMR spectrum, in this case chemical shifts, to be matched as a two-dimensional plot. Therefore comparison of these variables obtained from different NMR spectra can be done using pattern recognition approaches. One of the most simple, effective, and fast methods to align and match two-dimensional patterns is fast Fourier transform (FFT) cross-correlation. In this study we will evaluate whether there is a benefit of implementing this global approach for resonance identification using protein backbone assignment as an example. We started by creating a two-dimensional plot of the two carbon frequencies ($C^{\alpha}$ and $C^{\beta}$) extracted from two 3D NMR data sets, which provided $i$ to $i - 1$ linkages when aligned and matched. In addition, the correlation between amide nitrogen and proton frequencies for all of the residues provided spin system information. These two global matching procedures ensured simultaneous assignment of all data. We evaluated this protocol on four experimental data sets and ten synthetic data sets for proteins up to 723 residues in size, which resulted in near-complete assignment in six cases in just a few seconds. In the other eight cases, increase in protein size, incomplete assignment and overlap in the two-dimensional plots were found to decrease the effectiveness of the algorithm. We also expanded the algorithm to carbonyl data sets for the 226-residue glutamine binding protein (GlnBP) to highlight the flexibility of the protocol.

The FFT cross-correlation algorithm (Srinivasa Reddy and Chatterji 1996) is excellent at handling uniform translation, shear, and expansion of the patterns but poor at handling non-uniform shifts. To test the robustness of the algorithm, we intentionally introduced a first order phase artifact along the carbon dimension in one of the NMR experiments that resulted in a non-uniform shifting of peak positions. By dividing the regions into distinct subsections, however, we could eliminate the effect of non-uniform shifts during the calculation. We also tested the noise tolerance of the algorithm and found that a grid search of input parameters drastically reduced the number of errors introduced by noise. Finally, our protocol is easy to implement and has enough flexibility to be easily incorporated into currently existing analysis schemes, offering overall improvement in efficiency and scalability with potential for broad applications.

## Materials and methods

### Protein purification

$^{15}N/^{13}C$-GB1, $^{15}N/^{13}C$-ubiquitin, the UEV domain of $^{15}N/^{13}C$-TSG101 (residues 1–145) and $^{15}N/^{13}C$-glutamine binding protein (without glutamine) were produced as previously described (Gronenborn et al. 1991; Lazar et al. 1997; Pornillos et al. 2002; Bermejo et al. 2009, respectively).

### NMR spectroscopy

All NMR experiments were collected at 300 K using a Bruker Avance 600 MHz spectrometer equipped either with a cryogenic or room temperature probe, with 8 scans, and a carbon spectral width of 8445.946 Hz unless otherwise indicated. The following three-dimensional triple resonance experiments were collected: GB1 CBCA(CO)NH (512 × 52 × 40) complex data points in the $^{1}H$, $^{13}C$, and $^{15}N$ dimensions, HNCACB (512 × 57 × 40), 12 scans; ubiquitin (room-temperature probe) CBCA(CO)NH (512 × 57 × 52), HNCACB (1024 × 57 × 32); TSG101 CBCA(CO)NH (512 × 52 × 40), HNCACB (512 × 57 × 40); GlnBP (cryoprobe) CBCA(CO)NH (512 × 50 × 44), HNCACB (512 × 57 × 42); GlnBP (room-temperature probe) HNCA (1024 × 64 × 25), 12 scans, with a carbon spectral width of 1562.5 Hz; HN(CO)CA (1024 × 40 × 30), 16 scans, and a carbon spectral width of 1758.715 Hz; HNCO (1024 × 47 × 32), 12 scans, with a carbon spectral width of 1785.71 Hz; HN(CA)CO (1024 × 40 × 30), 48 scans, and a carbon spectral width of 1758.715 Hz, all as previously described (CBCA(CO)NH, Grzesiek and Bax 1992a; HNCACB, Wittekind and Mueller 1993; HNCA, Grzesiek and Bax 1992b; HN(CO)CA, Bax and Ikura 1991; HNCO, Kay et al. 1994; HN(CA)CO, Clubb et al. 1992).

### TSG101 HNCACB non-linear shift addition

To test the robustness of the algorithm to non-linear shifts in the data, the HNCACB for TSG101 was altered to include an additional fixed 19.5 μs delay time during the carbon evolution period. This introduced a first order phase artifact in the spectrum and led to varying shifts in the peak position across the spectrum.

### Preparation of input data from experimental spectra

NMR spectra were processed using NMRPipe (Delaglio et al. 1995) and analyzed using either Pipp (Garrett et al. 1991) or CCPN Analysis 2.4.1 (Vranken et al. 2005). For experimental NMR data (GB1, ubiquitin, TSG101, and GlnBP), peaks were picked manually using CCPN Analysis. Any folded peaks were unaliased at this point. Peaks could also be picked automatically using the *Initialise root resonances* and *Pick and assign from roots* macros within CCPN Analysis, which produces generic spin systems by naming peaks in the HSQC and picking corresponding peak strips in the triple resonance spectra. This process has the advantage that the noise outside of the strips, as well as side chain resonances from glutamine and asparagine residues can be discarded before further data analysis takes place. Although manual checking of these peak lists is still necessary to remove noise within strips in both spectra and the weak $i - 1$ peaks in the HNCACB, HNCA, and HN(CA)CO spectra, the overall time taken to analyze each spectrum is reduced dramatically. Peak lists were then output from CCPN Analysis as text files containing one row for each spin system arranged into four columns ($H^N$, N, $C_1$, $C_2$ where $C_1 < C_2$ for all residues except glycine where $C^\alpha = C_1 = C_2$). Text files containing the one-letter amino acid sequence of the proteins were also prepared. Spectra were also manually assigned using CCPN Analysis in order to verify the automated assignments.

### Preparation of input data from BMRB chemical shift lists

Our protocol was tested on a range of proteins from the BMRB, including echidna domain 11 IGF2R, (BMRB code 17287, Williams et al. 2012), the N-terminal NEAr iron transporter (NEAT1) domain of the IsdB hemoglobin receptor (19056, Fonner et al. 2014), oxidized Fe-containing superoxide dismutase (SD, 4341, Vathyam et al. 1999), mupirocin didomain ACP (diACP, 17111, Haines et al. 2013), ATP-bound ATPase (5576, Hilge et al. 2003), GTPase-activing and Ral binding domains of RLIP76 (RalBP1, 17608, Rajasekar et al. 2012), CrkL (18321, Jankowski et al. 2012), VRK1 (16715, Shin et al. 2011) and malate synthase G (5471, Tugarinov et al. 2002). $H^N$, N, $C^\alpha$ and $C^\beta$ chemical shifts were extracted from BMRB chemical shift tables and were converted to corresponding HNCACB and CBCA(CO)NH tables with the same final format as for the experimental data.

### Computational algorithm

We used a modular approach and Python scripting language throughout to aid with future portability of the algorithms to other software (see Fig. 1 for a summary of the algorithms used and below for further details). The computations were performed on a MacMini with a 2.3 GHz Intel Core i7 and 8 GB of memory, running OS X 10.8.5 and consisted of the following seven steps.

1. Read: Read input data in the forms of HNCACB and CBCA(CO)NH peak lists and the backbone sequence.
2. Plot: Generate a pair of two-dimensional plots named (a) amide and (b) carbon. For the amide plot, correlate proton ($H^N$) and nitrogen (N) backbone amide peaks by plotting the chemical shifts for $H^N$ and N on the x- and y-axes, respectively. This should be done twice, one for each experiment (HNCACB and CBCA(CO)NH), resulting in two overlaid data sets. Both should roughly resemble the HSQC spectrum excluding side-chains. For the carbon plot, correlate $C^\alpha$ and $C^\beta$ peaks by plotting the carbon chemical shifts from each spectrum. In order to avoid too much manual intervention, the lower value chemical shift (in ppm) was plotted on the x-axis ($C_1$) and the higher value chemical shift on the y-axis ($C_2$). The second carbon shift for glycine is assigned to be the same as the first carbon shift ($C^\alpha$) to simplify the protocol, hence for glycine, $C_1 = C_2$. Before plotting, the chemical shift scale is converted from a continuous scale to a discrete scale in order to create the boxes needed for FFT cross-correlation. The user defines the ppm resolution for this discretization (e.g., the size of the boxes). These boxes are described by $H_{res.}$ and $N_{res.}$ (for the amide plot) and $C_{res.}$ (for the carbon plot). Therefore, the dimension of each plot is defined to be the range of the data in that dimension divided by the desired resolution in ppm. As such, a lower

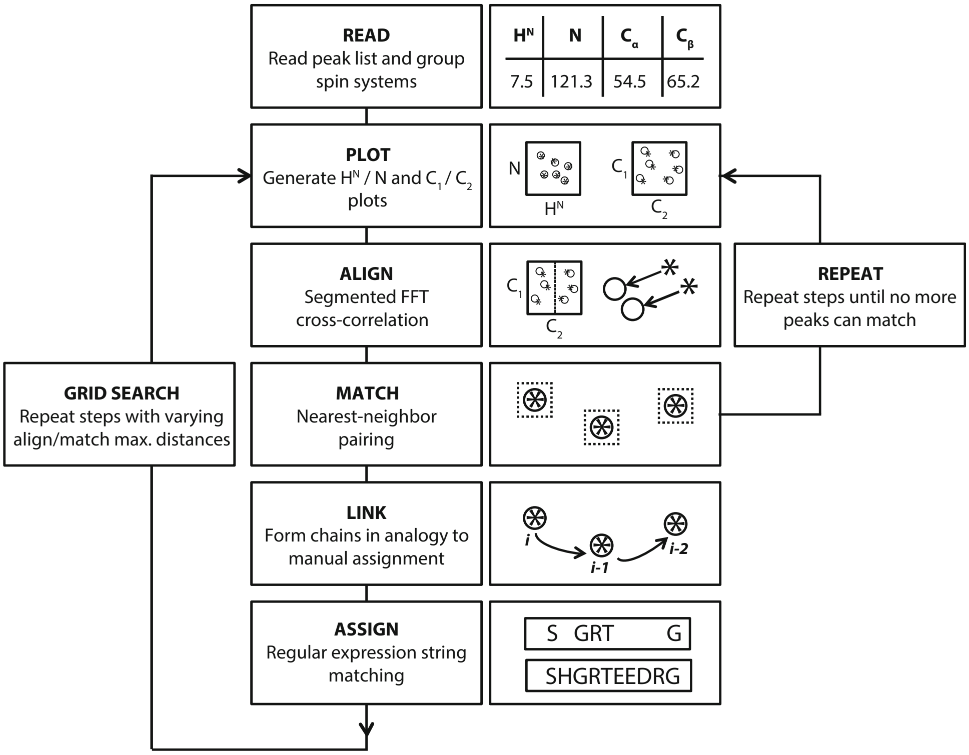
READ
Read peak list and group spin systems
H^N | N | Cα | Cβ
7.5 | 121.3 | 54.5 | 65.2
PLOT
Generate H^N / N and C1 / C2 plots
N
H^N
C1
C2
ALIGN
Segmented FFT cross-correlation
C1
C2
REPEAT
Repeat steps until no more peaks can match
GRID SEARCH
Repeat steps with varying align/match max. distances
MATCH
Nearest-neighbor pairing
LINK
Form chains in analogy to manual assignment
i
i-1
i-2
ASSIGN
Regular expression string matching
S GRT G
SHGRTEEDRG

**Fig. 1** Automated backbone assignment protocol. Our protocol (explained in detail in the text) consists of modular steps including 1. *Read*: Formation of spin systems from input peak lists. 2. *Plot*: Generation of a complementary pair of two-dimensional scatter plots using chemical shifts (ppm) for $H^N$/N and $C_1/C_2$ atoms on the x-/y-axes, respectively, from CBCA(CO)NH (*circles*) and HNCACB (*stars*) spectra. 3. *Align*: For the $H^N$/N plot—non-segmented FFT cross-correlation alignment; for the $C_1/C_2$ plot—segmented FFT cross-correlation alignment, to account for non-uniform shifts in data (see main text for default regions). 4. *Match*: Matching of peaks using nearest-neighbor pairing algorithm. 5. *Link*: Formation of linked chains of residues using pairings in analogy to manual assignment. 6. *Assign*: Use expected chemical shifts to assign glycine, alanine, serine and threonine (see main text for definitions). Sort chains by length and match to backbone string (sequence) using regular expression matching. 7. *Grid search*: Repeat steps 2–6, each time with a different set of input parameters, to find the optimal resolution needed to obtain the highest number of error-free matches. Resolution was related to maximum translation for fast Fourier transform cross-correlation (step 3) and maximum distance for nearest-neighbor pairing (step 4). Sequences were assessed for the percentage of amino acids that could be matched uniquely (unambiguously assigned to the backbone), non-uniquely (chains that could be matched in two or more places), or that couldn't be matched due to error. The best three results were chosen based on the best score in each category, respectively, with the caveat that all optimum results should score well in all three categories

value for $H_{res.}$, $N_{res.}$, or $C_{res.}$ (smaller boxes) will result in a longer computation time.

3. Align: Align the two plots using FFT cross-correlation (see Supplementary Materials and Supplementary Fig. 1). The user may optionally specify subregions of the data in order to segment the plot and align these subregions independently. Generally, the carbon plot was divided into four regions (glycines, serine/threonine, alanines, and others). The glycine region ($40 < C_1 < 50$ ppm and $40 < C_2 < 50$ ppm) was separated since these peaks were found on the diagonal of the plot and were best aligned by diagonal translational shifts. Since serine and threonine chemical shifts were reversed ($C^\alpha < C^\beta$) they needed to be separated from the rest of the data (53 ppm $< C_1 <$ max., 0 ppm $< C_2 <$ max., where max. is the highest chemical shift in that dimension). Alanines were also separated ($0 < C_1 < 25$ ppm, 0 ppm $< C_2 <$ max.) since their peaks on the plot were remote from the other types of residues. The amide plot was not segmented since non-linear shifts were not expected and the plot was generally of higher quality (better overlap). Maximum FFT cross-correlation translation distances were related to $H_{res.}$, $N_{res.}$, and $C_{res.}$ input parameters by the following equations.

$$Max.distance(H,N) = 2 \times \sqrt{H_{res.}^2 + N_{res.}^2} \quad (1)$$

$$Max.distance(C,C) = 7 \times C_{res.} \times \sqrt{2} \quad (2)$$

4. Match: Matching is carried out for the amide and carbon plots individually, using one-to-one nearest-neighbor peak matching where one experiment is

chosen as the base spectrum (e.g. HNCACB) for which its peaks are matched with their closest peaks in the second spectrum (CBCA(CO)NH). The order in which the peaks are matched does not matter since the matching is one-to-one. Any peak that has two neighbors at exactly the same distance will not be matched, but is not yet discarded. Any unmatched peaks will again be re-plotted, re-aligned, and matched until no more matches can be made or until a specified number of iterations is reached (the default value is eight iterations). As for the maximum FFT cross-correlation translation distance, maximum nearest-neighbor distance was also related to the $H_{res.}$, $N_{res.}$, and $C_{res.}$ (Equations 1 & 2) input parameters.

5. Link: Using the matches made in the previous step, sequentially link the residues in a manner analogous to manual assignment. Just like matching strips in HNCACB and CBCA(CO)NH spectra, pairs of $C^{\alpha}/C^{\beta}$ peaks can be used to link *i* and $i-1$ residues. $H^{N}/N$ pairs of peaks are used to match strips between HNCACB and CBCA(CO)NH spectra (e.g. both peak pairs will correspond to residue *i*). To begin, a peak is chosen at random in the amide plot that corresponds to residue *i* in the HNCACB. The matched peak from the CBCA(CO)NH spectrum corresponds to residue *i* as well, but will be linked to the $i-1$ residue via the previously formed arrays ($H^{N}$, N, $C_1$, $C_2$). By checking the match for the $i-1$ CBCA(CO)NH peak in the HNCACB spectrum, a second sequential link has been made, but this time between two different strips. Matches will continue to be read sequentially until no more matches are available. The algorithm then restarts reading matches, but this time in a forwards direction starting at residue *i* (to $i+1$, $i+2$, etc.) until no more matches are available. The algorithm then chooses a new *i* residue at random and reads as many matches as possible (both backwards and forwards). This process is repeated until the number of *i* residues that have been chosen matches the number of residues in the backbone string. This whole process produces chains of unassigned residues that are sequentially linked.
6. Assign: Regular expression matching is used to match sequentially linked chains to the backbone string. For example, a chain such as XXXGSA could match the sequence at KLMGSA or EFFGSA. The largest chain is matched first, followed by the next largest chain, etc., in decreasing order of size until no more chains can be matched. The assignment process uses alanine, serine, threonine, and glycine assignments derived from regions in the carbon plot. These regions are easily altered, which is recommended in the case of unusual shifts. We used the following as our default values: Alanine ($0 < C_1 < 25$ ppm, $0 < C_2 <$ max.), serine ($C_1 + C_2 > 115$ ppm, $C_1 + C_2 < 126.5$ ppm), threonine ($C_1 + C_2 > 126.5$ ppm, $C_1 <$ max., $C_2 <$ max.), and glycine ($40 < C_1 < 50$ ppm, $40 < C_2 < 50$ ppm, where $C_1 = C_2 = C^{\alpha}$). In addition, proline residues are blocked out in the regular expression matching since prolines cannot be assigned.
7. Grid search: If one run using the default input parameters was not sufficient to result in an error-free assignment, a grid search of the $H_{res.}$, $N_{res.}$, and $C_{res.}$ input parameters was used to find the best possible result. We found that, in general, changing $H_{res.}$ did not affect the results. Generally, a grid search consisted of 49 runs (1 value for $H_{res}$ at 0.0045 ppm, 7 values for $N_{res.}$ between 0.04 and 0.065 ppm, 7 values for $C_{res.}$ between 0.045 and 0.065 ppm, i.e. $1 \times 7 \times 7 = 49$ runs). Each run was usually complete in around a second, so a grid search would usually take around 1 min, depending on the size of the search space (number of parameter values to test) and the size of the protein. Table 1 summarizes the recommended values to use in a grid search for proteins of different sizes.

## Output

Assignments are output as an assigned HNCACB table, with the quality of the assignments assessed on a qualitative and quantitative basis. Linked chains will either match the backbone unambiguously (most likely correct assignments), ambiguously (where the chain can match in two or more places on the backbone), or erroneously (where an error has been made, either in linking or assigning, which results in no possible match to the backbone). These three categories (unambiguous, ambiguous, and erroneous) were scored as a percentage based on how many residues fit into each category. The best score from each category (the highest number of unambiguous and ambiguous assignments, and the lowest number of errors) would be used to choose the run with the optimal parameter input. Although this method of error assessment would not always put every residue in the correct category, it would usually find the run with the optimal parameter set and has the advantage that it requires minimal computation time in comparison to other methods of calculating errors.

## Extension to $C'$ chemical shifts

In order to show that the protocol is flexible and can be applied to any set of experiments where patterns are to be matched, we used our algorithm to assign GlnBP using HNCA, HN(CO)CA, HN(CA)CO, and HNCO experiments. As is the case for the HNCACB/CBCA(CO)NH

**Table 1** Results of backbone assignments using experimental or synthetic data

| Protein | BMRB code | Number of residues | | | | | Run time (s) | Opt. time (min:s) |
|---|---|---|---|---|---|---|---|---|
| | | Total | Assignable | Assigned (%) | Linked (%) | Errors | | |
| *Experimental (real) data sets* | | | | | | | | |
| GB1[a] | – | 56 | 55 | 100 | 100 | 0 | 0.89 | 0:02[e] |
| Ubiquitin[a] | – | 76 | 70 | 91.4 | 97.1 | 0 | 0.69 | 0:02[e] |
| TSG101[a] | – | 145 | 130 | 77.7 | 90.8 | 4 | 1.04 | 0:53[f] |
| GlnBP[a] | – | 226 | 204 | 46.1 | 86.8 | 15 | 4.26 | 4:38[f] |
| GlnBP[b] | – | 226 | 204 | 9.3 | 89.2 | 6 | 2.93 | 9:41[f] |
| *Synthetic data sets (BMRB)* | | | | | | | | |
| IGF2R[c] | 17287 | 140 | 133 | 99.2 | 100 | 0 | 1.18 | 0:17[g] |
| NEAT1[c] | 19056 | 163 | 131 | 96.9 | 100 | 0 | 1.19 | 0:18[g] |
| SD[c] | 4341 | 192 | 117 | 75.2 | 94.0 | 1 | 10.57 | 0:23[g] |
| diACP[c] | 17111 | 212 | 168 | 95.2 | 99.4 | 0 | 11.52 | 1:51[h] |
| ATPase[c] | 5576 | 213 | 155 | 54.8 | 87.1 | 1 | 8.50 | 1:10[h] |
| Tb24[c] | 18011 | 218 | 191 | 73.3 | 97.9 | 1 | 57.85 | 1:10[h] |
| RalBP1[c] | 17608 | 265 | 239 | 95.4 | 99.6 | 0 | 10.00 | 0:47[i] |
| CrkL[c] | 18321 | 303 | 222 | 35.6 | 73.9 | 1 | 1.63 | 1:12[h] |
| VRK1[c] | 16715 | 360 | 316 | 75.6 | 97.2 | 2 | 35.42 | 1:19[h] |
| MSG[c] | 5471 | 723 | 647 | 59.7 | 98.1 | 9 | 42.21 | 1:37[h] |
| *IGF2R noise addition* | | | | | | | | |
| 0× noise[d] | 17287 | 140 | 133 | 99.2 | 100 | 0 | 0.50 | 0:29[f] |
| 0.5× noise[d] | – | 140 | 133 | 99.2 | 100 | 0 | 0.58 | 0:30[f] |
| 1× noise[d] | – | 140 | 133 | 99.2 | 100 | 0 | 0.64 | 0:32[f] |
| 1.5× noise[d] | – | 140 | 133 | 79.7 | 97.0 | 3 | 0.93 | 0:35[f] |

The algorithm presented in the "Materials and methods" section was used to assign the backbone of a variety of proteins, which included segmented FFT cross-correlation to correct non-linear shifts in the data, followed by nearest-neighbor matching to link residues sequentially. For each protein, the BMRB code is given, if available, followed by the number of residues, and the number of assignable residues. Assignable residues exclude prolines, the N-terminus, and any residue that did not contain complete chemical shift data. The number of correctly assigned and correctly linked residues was calculated as a percentage of assignable residues. Residues were considered linked if they could form a chain of two or more residues. Any incorrect matches or assignments are labeled as errors. The run time is the time taken for one run of the assignment algorithm using the optimal parameter set. The opt. time (optimization time) was the amount of time necessary to perform the grid search of input parameters

[a] Assigned using experimental data sets (HNCACB and CBCA(CO)NH) and $H^N$/N, $C^{\alpha}$/$C^{\beta}$ plots

[b] Assigned using experimental data sets (HNCO, HN(CA)CO, HNCA and HN(CO)CA) and $H^N$/N, $C^{\alpha}$/$C'$ plots

[c] Assigned using corresponding peak tables (HNCACB and CBCA(CO)NH) produced from BMRB chemical shift data

[d] Noise added (see text)

[e] Grid search 1(N, 0.04 ppm) × 3(C, 0.04–0.07 ppm), recommended for <100 residues, real data

[f] Grid search 7(N, 0.045–0.065 ppm) × 7(C, 0.04–0.07 ppm), recommended for >100 residues, real data

[g] Grid search 2(N, 0.03–0.045 ppm) × 2(C, 0.01–0.03 ppm), recommended for <200 residues, BMRB

[h] Grid search 2(N, 0.03–0.045 ppm) × 2(C, 0.005–0.03 ppm), recommended for >200 residues, BMRB

[i] Grid search 2(H, 0.003–0.045 ppm) × 2(N, 0.03–0.045 ppm) × 2(C, 0.01–0.03 ppm), special case

spectral pair, these four experiments can be used to sequentially link backbone residues by exploiting known magnetization transfer pathways and are often used in the case of large proteins. For the carbon plot, in analogy to the outlined HNCACB/CBCA(CO)NH protocol, $C^{\alpha}$ and $C'$ were plotted on the x- and y-axes, respectively, using two pairs of two spectra – HNCA ($i$, $C^{\alpha}$, x-axis, pair 1) with HN(CA)CO ($i$, $C'$, y-axis, pair 1) and HN(CO)CA ($i-1$, $C^{\alpha}$, x-axis, pair 2) with HNCO ($i-1$, $C'$, y-axis, pair 2). The $H^N$/N plot was identical to the HNCACB/CBCA(CO)NH $H^N$/N plot. The algorithm was run in much the same way as for the HNCACB/CBCA(CO)NH pair of spectra (see Fig. 1), but with three exceptions. Firstly, the carbon plot was not segmented prior to FFT cross-correlation since there were no reversed (serine/threonine) or diagonal (glycine) peaks in the absence of $C^{\beta}$ chemical shifts. Secondly, only glycines ($0 < C^{\alpha} < 25$ ppm, 0 ppm $< C' <$ max.) could be used to assign linked chains to the backbone since no other peaks gave characteristic shifts in the carbon plot. Finally, for the grid search, a smaller default value was used for the input parameter $C_{res.}$ (0.01–0.04 ppm) since these particular spectra are known to be of a higher resolution than HNCACB/CBCA(CO)NH experiments.

### Addition of noise to synthetic peak lists

To test the robustness of the algorithm to noise, we added differing amounts of artificial noise to the synthetic IGF2R HNCACB and CBCA(CO)NH data tables. To give a general estimate of error, we measured the average standard deviation for correlated proton, nitrogen and carbon chemical shifts between experimental HNCACB and CBCA(CO)NH data sets for a protein of a similar size (TSG101, 145 residues). Using these values ($H^N = 0.0007$ ppm, $N = 0.03$ ppm, $C = 0.066$ ppm), we randomly added noise up to and including these values to our input data tables. Using the same input parameters for each run, we increased the noise using multiples of the average standard deviations. We categorized the different amounts of noise qualitatively, where zero noise represents perfect data and 0.5×, 1×, and 1.5× noise represent good quality data, normal data, and poor quality data, respectively. We ran the algorithm for each noise category using one parameter set (0.0045 ppm ($H^N$), 0.045 ppm (N), 0.045 ppm ($C_1$ and $C_2$)), or a grid search sampling 49 different parameter sets (0.0045 ppm, 1 parameter ($H_{res.}$); 0.04–0.065 ppm, 7 parameters ($N_{res.}$); 0.04–0.07 ppm, 7 parameters ($C_{res.}$)).

### Comparison with other assignment algorithms

Our protocol was compared to AutoAssign (Zimmerman et al. 1993) and Mars (Jung and Zweckstetter 2004) using the same input data as we used for TSG101. AutoAssign requires a FASTA sequence, as for our program, along with peak lists for HNCACB, CBCA(CO)NH, and HSQC spectra (our protocol does not require an HSQC). The peak lists were output directly from CCPN Analysis in Sparky format (Vranken et al. 2005) and uploaded to the AutoAssign WebServer (http://nmr.cabm.rutgers.edu/autoassign/cgi-bin/aaenmr.py). The output was compared to our manual assignment.

Mars requires a FASTA sequence, a PSIPRED secondary structure prediction, an input file (we used the default input parameters), and a chemical shift file containing N ($i$), $H^N$ ($i$), $C^{\beta}$ ($i-1$), $C^{\alpha}$ ($i-1$), $C^{\beta}$ ($i$) and $C^{\alpha}$ ($i$) chemical shifts in the form of labeled pseudoresidues (generic spin systems). The PSIPRED secondary structure prediction was performed using the PSIPRED server (Buchan et al. 2013; http://bioinf.cs.ucl.ac.uk/index.php?id=780). The chemical shift table was produced by combining our HNCACB and CBCA(CO)NH input data files. Mars 1.2 was run using Mars GUI 1.0 on a quad core 3.0 GHz Linux computer with 8 GB RAM running Ubuntu 14.10. The output was compared to our manual assignment.

## Results

Conventional backbone NMR resonance assignment of a protein involves matching resonance peak positions in at least one pair of complementary three-dimensional spectra, e.g. HNCACB/CBCA(CO)NH, HN(CA)CO/HNCO, or HN(CO)CA/HNCA. In this case we use HNCACB/CBCA(CO)NH as an example. The HNCACB spectrum contains six pieces of information for each amino acid – chemical shift information for the $H^N$, N, $C^{\alpha}$ and $C^{\beta}$ atoms for the current residue ($i$) and $C^{\alpha}$ and $C^{\beta}$ chemical shifts for the preceding residue ($i-1$). This information is contained in four carbon cross-peaks, all of which are typically displayed in a 'strip' at the $H^N$/N peak position (see Fig. 2a). In theory using the HNCACB spectrum alone, it is possible to assign the entire backbone of a protein by matching the $i$ and $i-1$ carbon peaks for each amino acid with other strips from the three dimensional spectrum. However, the $i-1$ carbon peak pair in any given strip are often weak, so the HNCACB spectrum is usually used in conjunction with the complementary CBCA(CO)NH, which only contains strong $i-1$ $C^{\alpha}$ and $C^{\beta}$ cross peaks.

In order to match $C^{\alpha}$ and $C^{\beta}$ peak pairs between strips, a search through the full set of strips from the complementary three-dimensional spectrum is carried out to find another pair of peaks with the same chemical shift in the carbon dimension. This is a tedious and slow process. Alternatively, we can plot the $C^{\alpha}$/$C^{\beta}$ peak pairs as a two-dimensional correlation plot (Fig. 2b), where peaks can be quickly matched, either by eye, or automatically using a nearest-neighbor search algorithm. $H^N$/N peak pairs can be matched in the same way, resulting in full assignment of a backbone of a protein with the exception of prolines and the N-terminal residue (Fig. 2c). Some amino acids, however, contain distinct carbon chemical shifts such as serine and threonine where the magnitude of their $C^{\alpha}$ chemical shift is smaller than their $C^{\beta}$ chemical shift. In addition, glycine has no $C^{\beta}$ atom. For creating correlation plots to match the peak positions from the two NMR data sets, it is irrelevant how the carbon shifts are labeled, as long as the two data sets are treated the same way. A convention was chosen where the smaller carbon chemical shift (designated as $C_1$) is plotted on the x-axis, while the larger one

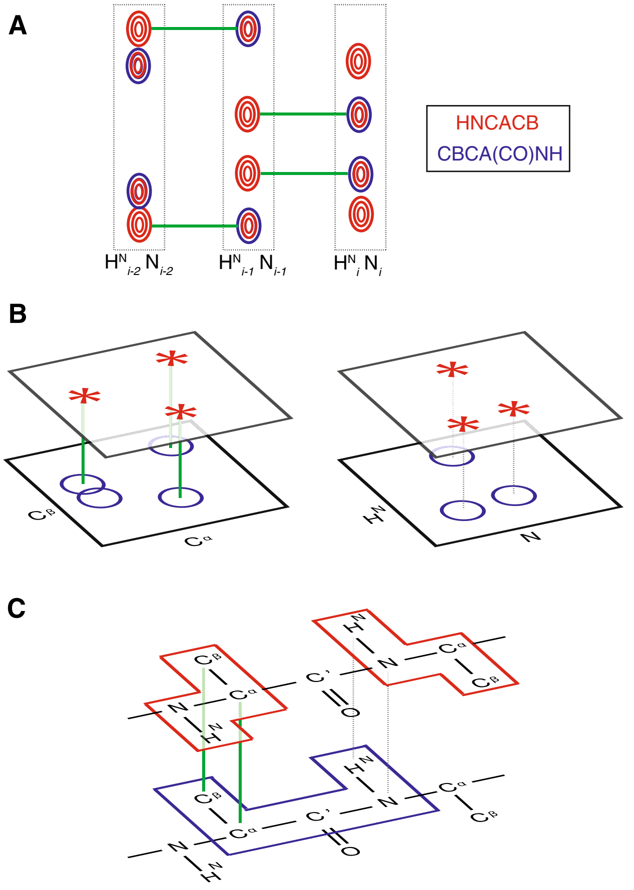

**Fig. 2** The use of correlations for NMR backbone resonance assignment. *Peaks* and *lines* in *red* and *blue* indicate data from the HNCACB and CBCA(CO)NH experiments, respectively. *Solid green lines* represent the same sequential linkages, in each case made through a $C^{\alpha}/C^{\beta}$ match. *Dotted black lines* represent intraresidue linkages (strips) made through an $H^N$/N match. **a** An illustration of traditional manual assignment based on strips. Each strip contains idealized peaks correlated with one residue's backbone amide ($H^N$ and N). These *peaks* include $C^{\alpha}$ and $C^{\beta}$ (residue *i*) from the HNCACB spectrum (*red*), and the $C^{\alpha}$ and $C^{\beta}$ (residue *i–1*) from the CBCA(CO)NH spectrum (*blue*). The weak HNCACB $i - 1$ peaks are also shown, but are not used in our protocol and are not shown in **b** and **c**. Sequential assignment involves linking pairs of *peaks* between strips, shown with *green lines*. **b** An illustration of overlay of two pairs of correlation plots showing idealized $C^{\alpha}/C^{\beta}$ and $H^N$/N chemical shifts from HNCACB (*red stars*) and CBCA(CO)NH (*blue circles*) spectra, signifying resonance assignment pairs of the backbone. The $C^{\alpha}/C^{\beta}$ pairs of chemical shifts from proline residues are not visible in the HNCACB spectra, shown by the *blue circle* without any pairing in the carbon correlation plot. The $H^N$/N correlation matches peaks found in both spectra coming from respective nuclei of residue *i*. Sequential assignment involves matching peaks between plots, shown with *green lines*. **c** The $C^{\alpha}/C^{\beta}$ correlation (*green lines*) matches peaks found in both the HNCACB (residue *i*) and CBCA(CO)NH (residue $i - 1$) spectra. The pathways for spin selection in HNCACB and CBCACONH are marked in *red* and *blue lines*, respectively

(designated as $C_2$ and $C_1 \leq C_2$) is on the y-axis. For glycine residues, $C_1$ and $C_2$ values were set to equal $C^{\alpha}$ chemical shifts, so they can then be matched as the diagonal peaks on the plot (Fig. 3a).

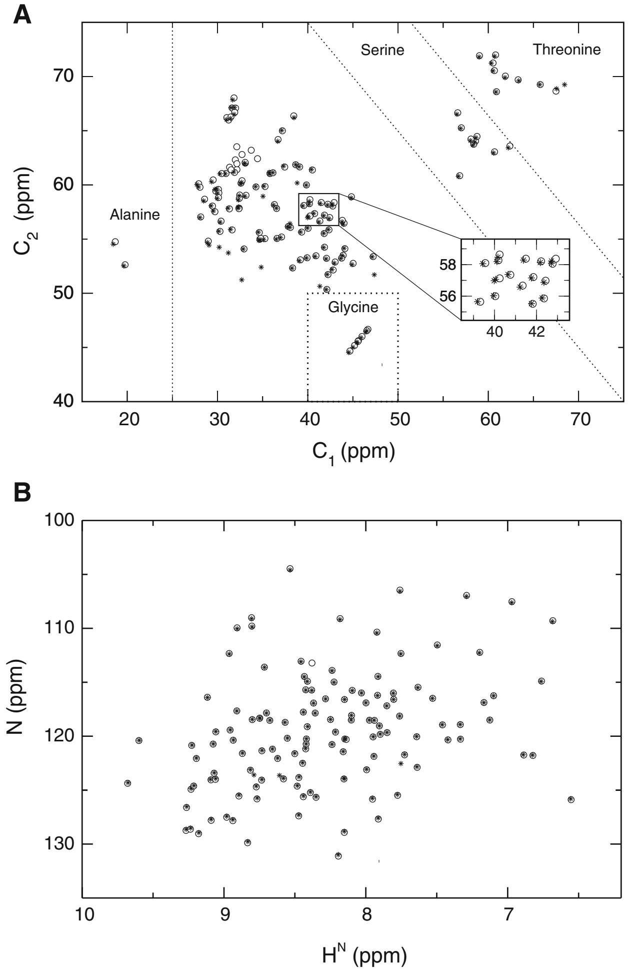

**Fig. 3** Carbon and proton-nitrogen correlation plots of TSG101 prior to segmented pattern matching Correlation of chemical shifts obtained from the HNCACB spectrum are plotted as *black stars*, while those from the CBCA(CO)NH spectrum are plotted as *white circles*. **a** Correlation of $C_1$ and $C_2$ chemical shifts for the *i* residue from the HNCACB data are overlaid on correlation of $C_1$ and $C_2$ chemical shifts for the $i - 1$ residue from the CBCA(CO)NH data. Non-uniform deviations in the position of the correlation points between the two experiments could be observed. Alanine, glycine, serine, and threonine peaks are separated from the remaining residues by their distinct chemical shift in the carbon correlation plot (regions shown in *dotted lines*). This property was used to apply segmented FFT translations and aided in assignment of chains to the backbone. The region of $C_1 = 38.5–42.5$ ppm and $C_2 = 54.6–58.4$ ppm is expanded in the inset to show how patterns in densely populated regions are easily recognized. **b** Correlation of $H^N$/N chemical shifts from the HNCACB and CBCA(CO)NH data sets are overlaid. Those correlation points that are not matched include missing peaks, proline residues and the N- and C-terminus

In practice, the carbon correlation plots from the two NMR data sets may not overlay exactly. Some regional shifts in the correlation plots caused by non-identical NMR experimental conditions may be observed. Figure 3a shows the $C_1/C_2$ correlation plot for the UEV domain of TSG101, a 17 kDa protein where the HNCACB experiment has been slightly modified (see "Materials and methods" section). Quick visual inspection, however, still readily reveals similarities in the clusters of points between the two data sets (see Fig. 3a). In fact, the human brain is astonishingly good at inferring how a cluster from one experiment should be shifted in order to align with a cluster from another experiment, at once providing several non-sequential linkages between the two NMR data sets. Our pattern matching approach, described in the "Materials and methods" section and Fig. 1, approximates this feat by individually aligning segments of these plots, allowing multiple individual linkages to be made simultaneously.

Using our segmented pattern search and nearest-neighbor peak matching, NMR resonances for the backbone of four proteins were assigned automatically using real experimental data, and compared to their assignment obtained manually (see Table 1). GB1 is a 56 amino acid protein that contains no proline residues. The $H^N$/N plot could be globally shifted directly by our pattern recognition algorithm, but with the $C_1$/$C_2$ correlation it was beneficial to divide the data into four regions to be shifted individually (e.g. serine/threonine, glycine, alanine and others). After nearest neighbor peak matching of the aligned data sets, a chain of 55 amino acids was found (Fig. 4). The whole process took 0.89 s. The assignment was carried out with no prior knowledge of expected chemical shifts since the only residue with missing information was the N-terminal methionine, as expected. This meant that we could confidently assign 100 % of the assignable residues in the protein, and they were found to be in perfect agreement with the manual assignment,.

However, we found that changing the user-defined input parameters had an effect on the outcome of the assignment ($H_{res.}$, $N_{res.}$, and $C_{res.}$, which essentially define the size of the boxes used to turn the continuous chemical shift scale to a discrete scale, a necessary process for the pattern alignment—see the "Materials and methods" section). We found that it was necessary to perform a grid search of the input parameters to maximize the number of correct assignments and minimize the number of errors. In practice, for GB1, we found that it was only necessary to alter $C_{res.}$ in the grid search, due to the higher accuracy of the proton and nitrogen dimensions of the original spectra, which meant that the assignment (100 % correct) was complete in only two seconds (see Table 1). Each run of the grid search was scored based on the number of unambiguously assigned residues, residues that could match the sequence in two or more places (ambiguous), and residues that couldn't possibly match the backbone (due to error in linking or assignment). The algorithm outputs the three runs that score best in each category with the added caveat that they also score well in all three categories, thus filtering out runs that contain errors.

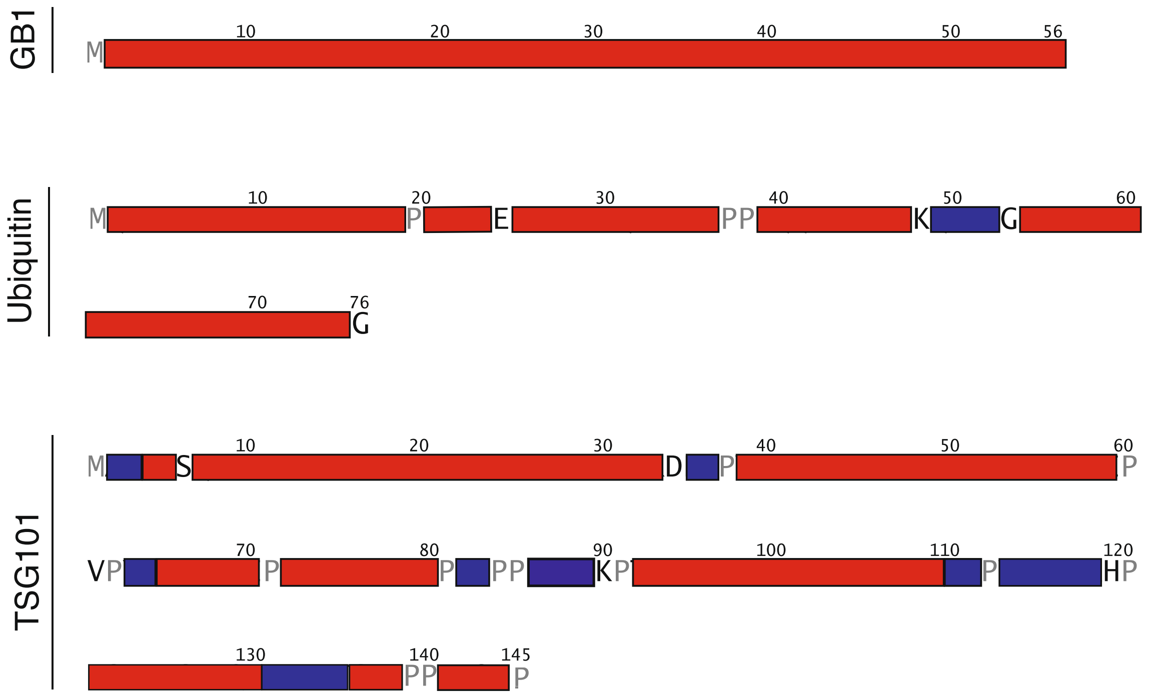

**Fig. 4** Results after application of FFT cross-correlation, nearest-neighbor matching, grid search, and assignment. Residues with correctly assigned backbone resonances are represented by *red boxes* overlaid on the sequence of GB1, ubiquitin, and TSG101. Residues whose resonances could be correctly linked but not assigned to the backbone are shown with *blue boxes*. Residues whose resonances could not be assigned are written explicitly in *black*. The N-terminal methionine and prolines are shown in *grey*

With the larger ubiquitin (76 amino acids), the $C_1/C_2$ correlation plot became slightly more crowded than for GB1. Even with the grid search of input parameter $C_{res.}$, it was impossible to assign linked chains to the backbone with no prior knowledge of expected chemical shifts for each residue type. GB1 could be assigned with no outside information since it formed only one chain of linked residues. For ubiquitin, however, chain breaks were either caused by missing spectral information due to prolines, resonance broadening (E24 and G53) or failed matches due to overlap of carbon shifts for glycine residues (G47 = 45.47 ppm and G75 = 45.36 ppm). By using the sectioning of the $C_1/C_2$ plot as a means of categorization of amino acids, serine, threonine, alanine, and glycine could be assigned. Using the assignments of these four residues along with the known linkages we had already formed, other residues could be assigned by inference. Using this protocol, all but one residue chain could be correctly assigned to the backbone of ubiquitin in comparison to manual assignment, resulting in coverage of 91.4 % of assignable residues (Fig. 4). As for GB1, we found that only the $C_{res.}$ value had an effect on the outcome of the grid search. The running time for the whole process including the grid search was two seconds, of which the optimized run took 0.69 s.

With the 145-residue TSG101, overlap and densely crowded regions became more challenging. We found it was necessary in this case to alter both the $C_{res.}$ and $N_{res.}$ input parameters, using a larger number of possible values (i.e. a 1 × 7 × 7 grid search). Using the optimal resolution, 22 amino acids (A2, V3, D34-K36, I86-K90, Y110, L111, Y113-W117, M131-F135) could not be assigned due to peak overlap of pairs of residues (V3/V89, K33/K108, P112/V130, all in the $C_1/C_2$ plots). In addition, twenty residues were either prolines or were next to proline and therefore could not be assigned (P37, P60-R64, P71, P81-P85, P91. P112, K118-P120, P139, P140, P145). One residue was missing assignments due to spectral overlap (S6). Despite these issues, 78 % of TSG101 residues could be correctly assigned in comparison to the manual assignment (see Fig. 4). A further 11 % of all residues could be linked correctly but not assigned. These unassigned chains typically resided in regions of the sequence between closely-spaced proline residues. Although these linked residues do not provide assignments directly, nevertheless, they provide linkage information that can be used in conjunction with the known chemical shift information to predict amino acid type manually, which would eventually result in complete assignment. Running time for the grid search was longer at 53 s, due to the increased size of the grid search in comparison to ubiquitin, including a time of 1.04 s for the run at the optimal resolution.

## Use of the assignment algorithm with synthetic data sets from the BMRB

To test the broader application of the backbone assignment algorithm to a wider variety of proteins, ten synthetic data sets were obtained from the BMRB (Fonner et al. 2014; Haines et al. 2013; Hilge et al. 2003; Jankowski et al. 2012; Rajasekar et al. 2012; Shin et al. 2011; Tugarinov et al. 2002; Vathyam et al. 1999; Williams et al. 2012; Xu et al. 2013). With proteins ranging in size from 140 to 723 residues, sampling a variety of secondary structures, including a paramagnetic-containing protein, a dimer, multidomain proteins, and liganded proteins, we felt that we surveyed a broad range of chemical shifts. Synthetic HNCACB and CBCA(CO)NH data sets were produced from BMRB chemical shift tables and run through the assignment algorithm (see Table 1). Ideally, in the absence of peak overlap, prolines, missing data, and assignments of residues outside of normal chemical shift ranges, the algorithm should attain complete assignment. For four of the proteins (IGF2R, NEAT1, diACP, and RalBP1), this was found to be the case with the exception of a few residues for each protein, with 99.2, 96.9, 95.2 and 95.4 % correct assignment, respectively. This is encouraging since the proteins range in size from 140 to 265 residues, with all types of secondary structure sampled. In particular, the near-complete assignment of the 212-residue diACP and 265-residue RalBP1 is surprising since all of the secondary structure for both proteins is α-helical, which should be detrimental due to increased overlap in the $C_1/C_2$ plot. What is most remarkable is that assignment for each of the four proteins was complete in a matter of seconds (for the optimized run), with a range from 1.18 s (IGF2R) to 10.00 s (RalBP1), with the time taken increasing concomitantly with protein size.

The algorithm does less well where many residues are missing complete chemical shift data, e.g. for the 192-residue dimeric and paramagnetic superoxide dismutase (39 % of residues missing some or all data), 213-residue ATP-bound ATPase (27 % missing) and the tridomain 303-residue CrkL (27 % missing). Missing data causes breaks in the linked chains formed by the algorithm, decreasing the average length of the chains. Since the assignment process uses only serine, threonine, alanine, and glycine chemical shifts to match chains of linked residues to the protein sequence, any breaks will be detrimental to unambiguous assignment. Despite this, the algorithm could correctly assign a useful quantity of residues for superoxide dismutase (75.2 %), ATPase (54.8 %), and CrkL (35.6 %), with only one error in each case. A much larger percentage of residues were correctly linked but could not be unambiguously assigned without further information (94.0, 87.1 and 73.9 % for each of the three proteins, respectively).

As always, larger proteins are known to be difficult to assign due to increased spectral overlap and line broadening. Using only HNCACB and CBCA(CO)NH data is not recommended since $C^{\alpha}$ and $C^{\beta}$ peak pairs readily overlap with an increased number of residues. Despite this, our algorithm was capable of assigning a considerable number of residues in a small amount of time for two large proteins, acting as the perfect starting point for a complete assignment using other methods. The 360-residue VRK1 backbone assignment took only 1 min and 19 s (75.6 % correctly assigned, 97.2 % correctly linked, 2 errors) in comparison to the assignment of the 723-residue malate synthase G in 1 min and 37 s (59.7 % correctly assigned, 98.1 % correctly linked, 9 errors). Although the time taken to assign these proteins was around five times longer than for the smaller proteins, due to the need for high resolution in the carbon plots especially (i.e. low values of $C_{res.}$), the assignments were fairly thorough and relatively error-free, which could significantly reduce the tedium associated with assigning such large proteins.

## Extension to C′ chemical shifts

In manual assignment, the use of additional spectra can reduce the number of overlap-associated errors while simultaneously increasing the number of correct assignments for larger proteins. Six experimental data sets for the 226-residue GlnBP were collected to highlight the flexibility of our assignment algorithm. Firstly, GlnBP was assigned using the standard HNCACB and CBCA(CO)NH data with the standard protocol of segmented FFT cross-correlation algorithm to correct for non-linear shifts in the data, nearest-neighbor pairing to make sequential links within $H^{N}$/N and $C^{\alpha}$/$C^{\beta}$ plots, and serine, threonine, glycine, and alanine assignments to match the linked chains to the backbone, which resulted in a 46.1 % assignment with 86.8 % linked residues and fifteen errors. In order to adapt the protocol for HNCA, HN(CO)CA, HNCO and HN(CA)CO spectra, rather than making the $C^{\alpha}$/$C^{\beta}$ plot, a $C^{\alpha}$/C′ plot was made by pairing HNCA (x-axis, $C^{\alpha}$, *i*, pair 1) with HN(CA)CO (y-axis, C′, *i*, pair 1) and HN(CO)CA (x-axis, $C^{\alpha}$, $i-1$, pair 2) with HNCO (y-axis, C′, $i-1$, pair 2), with each spectrum contributing one carbon-based chemical shift (see Fig. 5a). To make the sequential links between the chemical shifts, the HNCA/HN(CA)CO ($C^{\alpha}$, C′, residue *i*) pair was matched with the HNCO/HN(CO)CA ($C^{\alpha}$, C′, residue $i-1$) pair. One advantage with this method was that the $C^{\alpha}$/C′ plot was significantly better dispersed than the $C^{\alpha}$/$C^{\beta}$ plot, which along with the increased spectral resolution of these data sets, resulted in more matches being made (89.2 % residues linked versus 86.8 %). Despite the increased dispersion in the $C^{\alpha}$/C′ plot, only glycines could be distinguished based on unique carbon shifts from the plot itself (Fig. 5b). This resulted in a mere 9.3 % of residues being assigned unambiguously. However, by simply utilizing the $C^{\beta}$ shifts to help assign those residues linked in the $C^{\alpha}$/C′ experiments, three times the number of residues can be assigned (27.9 %) with eighteen of those assignments adding to those already assigned with the HNCACB and CBCA(CO)NH spectra.

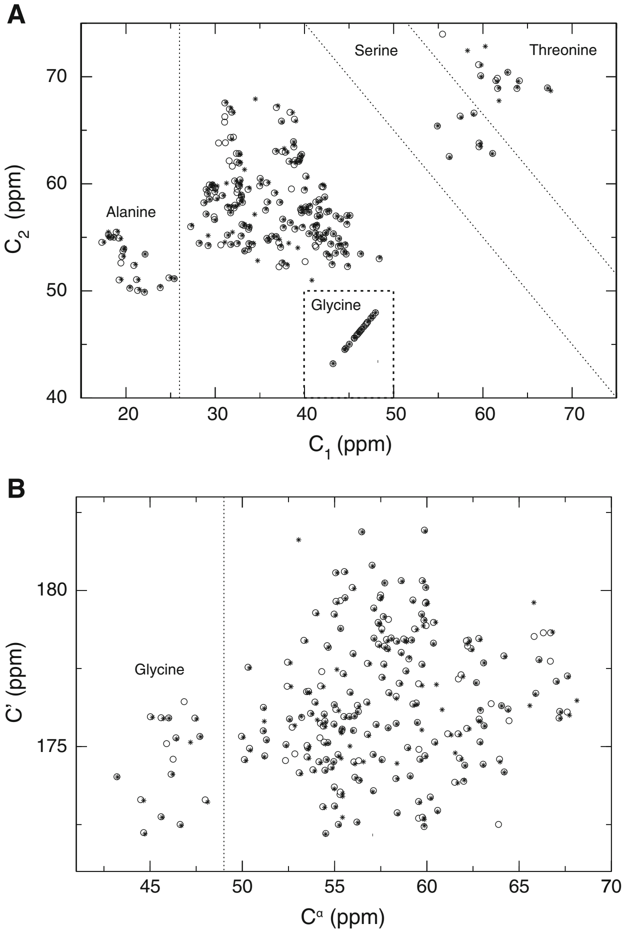

**Fig. 5** Carbon correlation plots of GlnBP prior to pattern matching. **a** Correlation of $C_1$ and $C_2$ chemical shifts for the *i* residue from the HNCACB data (*black stars*) are overlaid on correlation of $C_1$ and $C_2$ chemical shifts for the $i-1$ residue from the CBCA(CO)NH data (*white circles*). Non-uniform deviations in the position of the correlation points between the two experiments could be observed. Alanine, glycine, serine, and threonine peaks are separated from the remaining residues by their distinct chemical shift in the carbon correlation plot (regions shown in *dotted lines*). This property was used to apply segmented FFT translations and aided in assignment of chains to the backbone. **b** Correlation of $C^{\alpha}$ and C′ chemical shifts for the *i* residue from the HNCA ($C^{\alpha}$, *i*) and HN(CA)CO (C′, *i*) data (*black stars*) are overlaid on correlation of $C^{\alpha}$ and C′ chemical shifts from the $i-1$ residue from the HN(CO)CA ($C^{\alpha}$, $i-1$) and HNCO (C′, $i-1$) data (*white circles*). Glycine peaks are separated from the remaining residues by their distinct chemical shift in the carbon correlation plot. This property aided in assignment of chains to the backbone. No segmentation was used for the FFT cross-correlation of this plot

Both types of assignment protocols result in good sequence coverage of residues in linked chains (86.8 and 89.2 % for the $C^{\alpha}/C^{\beta}$ and $C^{\alpha}/C'$ methods, respectively), but when combined cover 98.5 % of residues. This means that for nearly all of the assignable residues we can obtain some kind of linkage information from our protocols even if they are not necessarily assigned. Most importantly, no errors are replicated between the two assignment protocols, which means that a higher confidence in assignment can be reached by combining the two methods.

### Robustness of the algorithm to the addition of noise

Protein size, sample quality, the presence of a cryogenic probe, magnetic field strength, and acquisition parameters can all contribute significantly to the final quality of NMR spectra. Since our algorithm involves the matching of peaks between sets of spectra, which can have varying degrees of accuracy and signal-to-noise ratio, we tested the robustness of our protocol to the addition of noise. We started with the synthetic HNCACB and CBCA(CO)NH IGF2R data sets (140 residues), which by their nature contained identical peak positions and are therefore representative of zero noise. As stated previously, the algorithm correctly assigned 99.2 % of assignable residues for IGF2R in just half a second. Using multiples of the noise calculated from a protein of similar size (TSG101, 145 residues), we artificially added random noise to the HNCACB and CBCA(CO)NH data sets for IGF2R where 0×, 0.5×, 1×, and 1.5× noise corresponds to perfect, good, normal, and poor quality data, respectively. As expected, with increased noise the number of correct assignments decreased and the number of errors increased (zero noise = 99.2 % correct, 0 errors; 0.5× noise = 25.6 %, 4 errors; 1× noise = 25.6 %, 4 errors; 1.5× noise = 6.8 %, 5 errors). Despite this large drop in assigned residues, even for poor quality data, the algorithm still correctly links the majority of residues (97.7 %) with an average chain length of 13 residues between chain breaks. The reason why the algorithm is not capable of assigning the residues is due to matching errors. However, by using a grid search of $N_{res.}$ and $C_{res.}$ input parameters and a scoring of the output, as discussed earlier for TSG101 and GlnBP, we can select for runs that result in a high percentage of correct assignments and with the least number of errors (see Table 1). With good quality and normal quality data (0.5× and 1× noise), 99.2 % of assignments could be made, which is on a par with perfect data (zero noise). Only with poor quality data (1.5× noise) were errors encountered—79.7 % of residues were correctly assigned and 3 errors were made. However, at this level of noise, the presence of errors would indicate that higher quality data would be desirable and could indicate the need for sample optimization before collection of further experiments. Interestingly, the increase in the amount of noise had minimal effect on the running time for each of the grid searches, which were complete in 30 s, 32 s, and 35 s for 0.5×, 1×, and 1.5× noise, respectively.

### Comparison with other backbone assignment algorithms

AutoAssign is one of the most widely used backbone assignment algorithms with nearly 300 citations in the PDB (as of 03/03/2015) (Williamson and Craven 2009; Zimmerman et al. 1993). Its popularity is most likely due to the fact that it can cope with peak lists from a variety of spectra, without any need for manual intervention, and runs in a matter of seconds. Using TSG101 HNCACB and CBCA(CO)NH peak lists and either AutoAssign or our protocol, 58.4 and 76.9 % of residues can be assigned, respectively. Although both programs make the same number of incorrect assignments (6.9 %), these errors are not replicated between the two programs. As such, the programs could be used in a complementary manner to reinforce correct assignments and reduce the overall number of errors of both programs, especially since AutoAssign has the capability of forming spin systems directly from peak lists.

Mars (Jung and Zweckstetter 2004) was also used to assess our protocol in comparison to other backbone assignment systems available. Using the same input data but with the addition of a PSIPRED secondary structure prediction (Buchan et al. 2013), Mars correctly assigned a large proportion of residues (97.7 %) with no errors, and in only 20 s (3 GHz, quad core Linux), plus 9 min for the PSIPRED prediction. Like our protocol, Mars requires the production of generic spin systems (pseudoresidues) and performs better from manually curated peak lists (Jung and Zweckstetter 2004).

Two of the larger BMRB proteins we used were also previously chosen in testing the effectiveness of Mars (Jung and Zweckstetter 2004). Using only $C^{\alpha}$ and $C^{\beta}$ data, Mars could assign 95.7 % of assignable superoxide dismutase residues and 76.5 % of malate synthase G residues with no errors (Jung and Zweckstetter 2004), which compares to 75.2 % (1 error) and 59.7 % (9 errors), respectively, for our protocol. It is apparent from these comparisons that the more information or data being incorporated into the assignment procedure, the better the outcome.

## Discussion and conclusions

We evaluated the possible benefits of our global approach in NMR resonance identification and showed that a segmented pattern matching method could rapidly assign the majority of NMR resonances of protein backbones. Manual backbone assignment of NMR resonances for the proteins used in this study would typically take anything between a few hours to a few weeks. Segmenting the pattern search areas into sections could solve the problem of when non-ideal experimental setup introduces non-uniform errors in the carbon chemical shifts. Optimizing the pattern search variables, such as the resolution chosen to discretize the chemical shift scale in each dimension, further improves the assignment outcome.

Even though we illustrated our method using HNCACB and CBCACONH NMR experiments, the approach can easily be used for matching any other NMR data sets, as we have shown for HNCA, HN(CO)CA, HNCO, and HN(CA)CO experiments. The general concept is to display any two variables in the NMR data sets that need to be compared and matched (in our example these are the two carbon chemical shifts or the proton and nitrogen chemical shifts) as a two-dimensional correlation plot where we can gain several advantages. First, we establish a global map of the problem and this is more conducive to efficient automation. We can choose to use a pattern recognition algorithm as a way to automate the process. This is a much faster process and can scale up to larger proteins much more favorably than the serial approach typically used in comparing strips of NMR spectra. Furthermore, missing or overlapped points in the data sets, typically encountered in larger proteins, are not detrimental to the pattern recognition algorithm. Second, any problems such as non-uniform chemical shift errors are easily identified in a global sense and a solution can be adopted in the analysis phase of the experiment in a straightforward manner. Finally, the approach can be optimized and the result be evaluated to provide a global goodness of fit.

It is important to point out that the FFT cross-correlation algorithm that we used in our pattern matching protocol is similar to the algorithm that Buchner et al. (2013) used in referencing chemical shifts in NMR experiments prior to analysis. Their algorithm, as well as ours, is a choice based on practical reasons: convenience and speed. In fact, the majority of time that it took to run our simple program was due to bookkeeping during the grid search. Once the system was optimized, it took merely a second on a modest computer to get the results. Our concept could easily be integrated into NMR analysis software where the two correlation plots can be displayed and the user can match the points visually, thus streamlining the backbone NMR resonance assignment. The limit of two variables at the moment is due to the pattern recognition algorithm. This could be overcome by creating multiple two dimensional correlation plots that can be compared simultaneously. Furthermore, prior knowledge can be added into the analysis, such as a predictive correlation network in the NMR resonances from expected distances between protons in regular secondary structure of a protein (Herrmann et al. 2002; Volk et al. 2008). This can potentially further improve the efficiency of the global matching process in NMR data analysis.

**Acknowledgments** This work was supported by the Intramural Research Program of the NIH, National Heart, Lung, and Blood Institute. Our software is available free of charge by contacting Nico Tjandra (tjandran@nhlbi.nih.gov).

## References

Bahrami A, Assadi A, Markley JL, Eghbalnia H (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5:e1000307

Bartels C, Günter 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

Bax A, Ikura M (1991) An efficient 3D NMR technique for correlating the proton and $^{15}$N backbone amide resonances with the α-carbon of the preceding residue in uniformly $^{15}$N/$^{13}$C enriched proteins. J Biomol NMR 1:99–104

Bermejo GA, Strub M-P, Ho C, Tjandra N (2009) Determination of the solution-bound conformation of an amino acid binding protein by NMR paramagnetic relaxation enhancement: use of a single flexible paramagnetic probe with improved estimation of its sampling space. J Am Chem Soc 131:9532–9537

Borkar A, Kumar D, Hosur RV (2011) AUTOBA: automation of backbone assignment from HN(C)N suite of experiments. J Biomol NMR 50:285–297

Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED protein analysis workbench. Nucleic Acids Res 41:W349–W357

Buchner L, Schmidt E, Güntert P (2013) Peakmatch: a simple and robust method for peak list matching. J Biomol NMR 55:267–277

Chen K, Delaglio F, Tjandra N (2010) A practical implementation of cross-spectrum in protein backbone resonance assignment. J Magn Reson 203:208–212

Clubb RT, Thanabal V, Wagner G (1992) A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue $^{1}$H$^{N}$, $^{15}$N, and $^{13}$C′ chemical shifts in $^{15}$N–$^{13}$C-labeled proteins. J Magn Reson 97:213–217

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

Dutta SK, Serrano P, Proudfoot A, Geralt M, Pedrini B, Herrmann T, Wüthrich K (2014) APSY-NMR for protein backbone assignment in high-throughput structural biology. J Biomol NMR 61:47–53

Fonner BA, Tripet BP, Lui M, Zhu H, Lei B, Copié V (2014) $^{1}$H, $^{13}$C, $^{15}$N backbone and side chain NMR resonance assignments of the N-terminal NEAr iron transporter domain 1 (NEAT 1) of the hemoglobin receptor IsdB of *Staphylococcus aureus*. Biomol. NMR Assign. 8:201–205

Garrett DS, Powers R, Gronenborn AM, Clore GM (1991) A common sense approach to peak picking two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. J Magn Reson 95:214–220

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–661

Grzesiek S, Bax A (1992a) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. J Am Chem Soc 114:6291–6293

Grzesiek S, Bax A (1992b) Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. J Magn Reson 96:432–440

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

Haines AS, Dong X, Song Z, Farmer R, Williams C, Hothersall J, Płoskoń E, Wattana-Amorn P, Stephens ER, Yamada E, Gurney R, Takebayashi Y, Masschelein J, Cox RJ, Lavigne R, Willis CL, Simpson TJ, Crosby J, Winn PJ, Thomas CM, Crump MP (2013) A conserved motif flags acyl carrier proteins for β-branching in polyketide synthesis. Nat Chem Biol 9:685–692

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

Hilge M, Siegal G, Vuister GW, Güntert Gloor SM, Abrahams JP (2003) ATP-induced conformation changes of the nucleotide-binding domain of Na, K-ATPase. Nat Struct Mol Biol 10:468–474

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102:10876–10881. doi:10.1073/pnas.0504818102

Jankowski W, Saleh T, Pai M-T, Sriram G, Birge RB, Kalodimos CG (2012) Domain organization differences explain Bcr-Abl's preference for CrkL over CrkII. Nat Chem Biol 8:590–596

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

Rajasekar KV, Campbell LJ, Nietlispach D, Owen D, Mott HR (2012) $^{1}$H, $^{13}$C and $^{15}$N resonance assignments of the GTPase-activating (GAP) and Ral binding domains (GBD) of RLIP76 (RalBP1). Biomol NMR Assign 6:119–122

Kay LE, Guang TX, Yamazaki T (1994) Enhanced-sensitivity triple-resonance spectroscopy with minimal $H_2O$ saturation. J Magn Reson A 109:129–133

Lazar G, Desjarlais JR, Handel TM (1997) De novo design of the hydrophobic core of ubiquitin. Protein Sci 6:1167–1178

Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

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

Lukin JA (1997) Automated probabilistic method for assigning backbone resonances of ($^{13}$C, $^{15}$N)-labeled proteins. J Biomol NMR 9:151–166

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

Pornillos O, Alam SL, Rich RL, Myszka DG, Davis DR, Sundquist WI (2002) Structure and functional interactions of the Tsg101 UEV domain. EMBO J 21:2397–2406

Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin T, Nilges M (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23:381–382

Shin J, Chakraborty G, Bharatham N, Kang C, Tochio N, Koshiba S, Kigawa T, Kim W, Kim K-T, Yoon HS (2011) NMR solution structure of human vaccinia-related kinase 1 (VRK1) reveals the C-terminal tail essential for its structural stability and autocatalytic activity. J Biol Chem 286:131–138

Srinivasa Reddy B, Chatterji BN (1996) An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Tans Image Process 5:1266–1271

Tugarinov V, Muhandiram R, Ayed A, Kay LE (2002) Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase G. J Am Chem Soc 124:10025–10035

Vathyam S, Byrd RA, Miller A-F (1999) Assignment of the backbone resonances of oxidized Fe-superoxide dismutase, a 42 kDa paramagnet-containing enzyme. J Biomol NMR 14:293–294

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

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

Williams C, Hoppe H, Rezgui D, Strickland M, Forbes BE, Grutzner F, Frago S, Ellis RZ, Wattana-Amorn P, Prince SN, Zaccheo OJ, Nolan CM, Mungall AJ, Jones EY, Crump MP, Hassan AB (2012) An exon splice enhancer primes IGF2:IGF2R binding site structure and function evolution. Science 338:1209–1213

Williamson MP, Craven CJ (2009) Automated protein structure calculation from NMR data. J Biomol NMR 43:131–143

Wittekind M, Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta-carbon resonances in proteins. J Magn Reson B 101:201–205

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

Xu X, Olson CL, Engman DM, Ames JB (2013) $^{1}$H, $^{15}$N, and $^{13}$C chemical shift assignments of the calflagin Tb24 flagellar calcium binding protein of *Trypanosoma brucei*. Biomol NMR Assign 7:9–12

Zimmerman DE, Kulikowski CA, Montelione GT (1993) A constraint reasoning system for automating sequence-specific resonance assignments from multidimensional protein NMR spectra. Proc Int Conf Intell Syst Mol Biol 1:447–455