J Biomol NMR (2013) 55:267–277
DOI 10.1007/s10858-013-9708-z

ARTICLE

Peakmatch: a simple and robust method for peak list matching

Lena Buchner · Elena Schmidt · Peter Güntert

Received: 18 October 2012 / Accepted: 9 January 2013 / Published online: 18 January 2013

Abstract Peak lists are commonly used in NMR as input data for various software tools such as automatic assignment and structure calculation programs. Inconsistencies of chemical shift referencing among different peak lists or between peak and chemical shift lists can cause severe problems during peak assignment. Here we present a simple and robust tool to achieve self-consistency of the chemical shift referencing among a set of peak lists. The *Peakmatch* algorithm matches a set of peak lists to a specified reference peak list, neither of which have to be assigned. The chemical shift referencing offset between two peak lists is determined by optimizing an assignment-free match score function using either a complete grid search or downhill simplex optimization. It is shown that peak lists from many different types of spectra can be matched reliably as long as they contain at least two corresponding dimensions. Using a simulated peak list, the *Peakmatch* algorithm can also be used to obtain the optimal agreement between a chemical shift list and experimental peak lists. Combining these features makes *Peakmatch* a useful tool that can be applied routinely before automatic assignment or structure calculation in order to obtain an optimized input data set.

Keywords Automated assignment · Peak list · Peak alignment · Spectrum referencing · CYANA

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9708-z) contains supplementary material, which is available to authorized users.

L. Buchner · E. Schmidt · P. Güntert (✉)
Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, and Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
e-mail: guentert@em.uni-frankfurt.de

Introduction

Protein structure determination by NMR spectroscopy has been accelerated by the development of programs that perform some or all of the necessary steps automatically (Baran et al. 2004; Guerry and Herrmann 2011; Güntert 2009; López-Méndez and Güntert 2006; Williamson and Craven 2009). The majority of these programs use the information from the NMR spectra in the form of peak lists rather than by accessing the spectra directly. For most applications a set of peak lists from different types of experiments is needed. It is important to have a consistently referenced data set for the resonance assignment, and automated NOE assignment and structure calculation require that the NOESY peak lists and the corresponding chemical shift list(s) are in optimal agreement.

Several programs exist for correcting the referencing of chemical shifts or optimizing the agreement between chemical shift assignments and general chemical shift statistics (Aeschbacher et al. 2012; Ginzinger et al. 2007; Wang et al. 2005; Wang and Wishart 2005). Methods are also available to adapt chemical shifts to NOESY spectra (Herrmann et al. 2002). However, to the best of our knowledge there is no program available that optimizes automatically the mutual referencing of several unassigned, multidimensional peak lists to achieve a consistently referenced data set prior to automated assignment or structure calculation.

Materials and methods

The new *Peakmatch* algorithm implemented in the CYANA software package (Güntert 2009; Güntert et al. 1997) calculates the optimal chemical shift referencing offsets

between two peak lists by maximizing a match score using either a grid search or downhill simplex method. One peak list is used as a reference and remains unchanged whereas each corresponding dimension in the second, target peak list is shifted by a constant offset. The offsets that yield the maximal match score represent the calculation result. An overview of the algorithm is given in Fig. 1.

Determination of corresponding dimensions

The user specifies the dimensions in the reference peak list. The algorithm can then determine the corresponding dimensions in the target peak list automatically based on the expected peak match. If more than one possibility is found, the one with the largest expected peak match is chosen.

To calculate the expected peak match score, the program generates expected peaks for the reference and target peak lists based on experiment type-specific connectivity patterns stored in the CYANA library and the covalent structure of the protein (Bartels et al. 1997; Schmidt and Güntert 2012; Schmucki et al. 2009). Through-space type experiments are approximated by the subset of short-range peaks, which is accurate enough for the present purpose. Details of the generation of expected peaks have been given elsewhere (Schmidt and Güntert 2012). The expected peak match score is calculated using Eq. 1,

$$E = \sum_{i=1}^{n_0} \sum_{j=1}^{m_0} \theta_{ij} \tag{1}$$

where n_0 and m_0 denote the number of expected peaks for the reference and target experiment, respectively, and θ_{ij} is one if the two expected peaks have the same assignment in all dimensions considered for match calculation, and zero otherwise. The expected peak match is independent from the experimental peak lists and thus not influenced by the lack of peaks or the presence of artifacts.

Match score

For a reference peak list with $i = 1,\ldots,n$ peaks at positions $\omega_{ik}^{(1)}$ and a target peak list with $j = 1,\ldots,m$ peaks at positions $\omega_{jk}^{(2)}$ in the $k = 1,\ldots,d$ corresponding dimensions, we define the match score S as a function of the offsets $\delta_1,\ldots,\delta_d$:

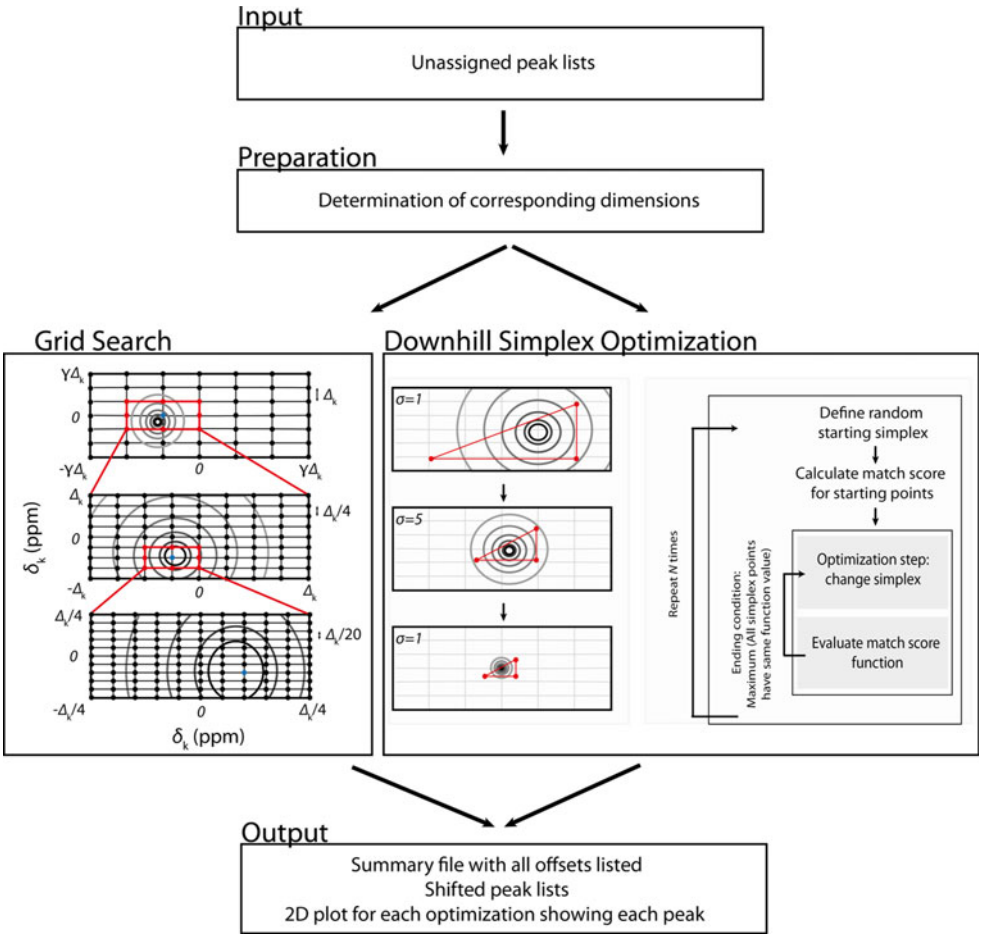

Fig. 1 Flowchart of the *Peakmatch* algorithm

$$S(\delta_1,\ldots,\delta_d) = \frac{1}{E}\sum_{i=1}^{n}\sum_{j=1}^{m}{}' \exp\left(-\sum_{k=1}^{d}\left(\frac{\omega_{ik}^{(1)} - \omega_{jk}^{(2)} + \delta_k}{\sigma\Delta_k}\right)^2\right) \quad (2)$$

The contribution of an individual peak pair to the match score is given by a Gaussian function of the normalized distance between the two peaks. The chemical shift tolerances Δ_k represent the accuracy of the peak positions. They should be set by the user such that the positions of any two peaks assigned to the same atom differ by less than the chemical shift tolerance in the given dimension. The default values are 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N dimensions. The dimensionless scaling factor σ determines the significance of a deviation. By default, $\sigma = 1$. Deviations of peak positions smaller than $\sigma\Delta_k$ yield score contributions close to one, whereas those from deviations much larger than $\sigma\Delta_k$ are negligible. The overall match S between two peak lists is calculated as a sum over all n peaks in the reference peak list. For each reference peak i, the $q = E/n_0$ largest contributions from peaks in the target peak list are included in the match score calculation, as indicated by the prime in Eq. 2. The parameter q represents the expected average number of peaks in the target peak list with the same assignment as a given reference peak in the corresponding dimensions. This results in larger q values when optimizing for instance ^{15}N-resolved [^{1}H,^{1}H]-NOESY against [^{15}N,^{1}H]-HSQC ($q \approx 13$) compared to HNCA against [^{15}N,^{1}H]-HSQC ($q = 2$), or two peak lists from the same experiment type ($q = 1$). Assignments for the input peak lists are not required.

The match score function of Eq. 2 approximately counts the number of peaks in the two peak lists whose position matches within the ranges $\sigma\Delta_k$, and does so with minimal influence from other, non-matching peaks. The match score S is normalized by the expected peak match E of Eq. 1. It thus has the value 1 for two ideally matched peak lists that contain exactly the expected peaks. In general, the match score shows one narrow optimum when using two or more corresponding dimensions (Fig. 2a) and gets broader as well as smoother with increasing σ (Fig. 2b, c).

Optimization procedures

Grid search

The grid search evaluates the match score function at every point of a grid and takes as result the offset values $\delta_1,\ldots,\delta_d$ that yield the maximum score value. With an appropriate grid size and spacing, this procedure guarantees the identification of the global maximum, which is in general the correct offset. In order to save computation time, the grid search procedure performs several steps at different grid sizes and spacings (Fig. 1, left side). The first grid covers the largest offset range, which should be chosen larger than the expected offset to ensure that the region of the global maximum is found. To this end, the user specifies a dimensionless parameter γ to define a rectangular grid of size $[-\gamma\Delta_k, \gamma\Delta_k]$ and spacing Δ_k in the corresponding dimensions $k = 1,\ldots,d$. Two subsequent grid searches are performed using smaller grids of sizes $[-\Delta_k, \Delta_k]$ and $[-\Delta_k/4, \Delta_k/4]$ with smaller spacings of $\Delta_k/4$ and $\Delta_k/20$, respectively, centered at the optimum found in the preceding search. This procedure allows finding the correct offset at high precision without having to search a large grid with very small spacing between the grid points. Nevertheless, depending on the size of the initial grid, calculation times can be significant.

Downhill simplex optimization algorithm

To further reduce the computation time, a downhill simplex minimization algorithm (Nelder and Mead 1965) can be used to find the optimal offsets $\delta_1,\ldots,\delta_d$ between two peak lists. This algorithm makes use of a simplex of $d + 1$ points in

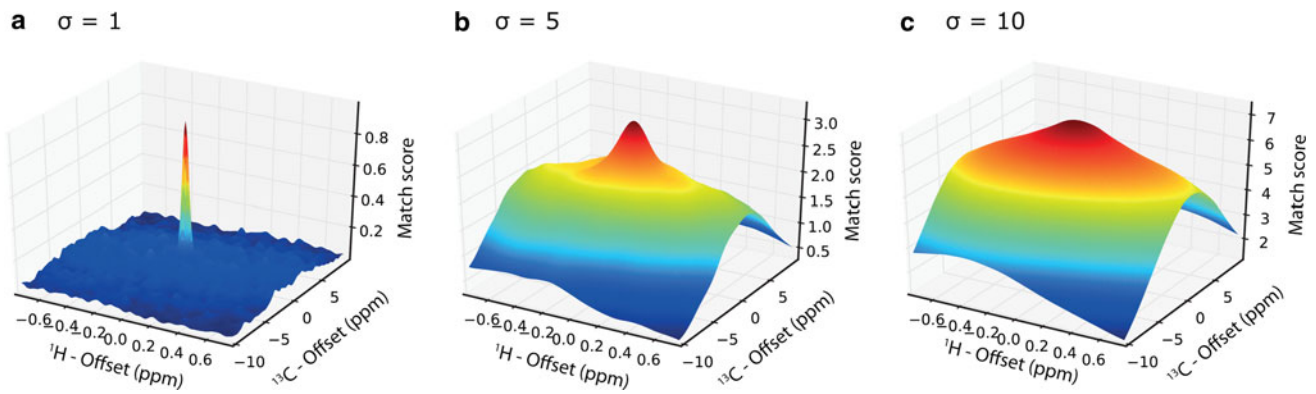

Fig. 2 Match score function for two corresponding dimensions and different σ values (see Eq. 2). The [^{15}N,^{1}H]-HSQC reference peak list and the CBCACONH target peak list are from the manually edited ENTH data set

d dimensions, e.g. a triangle in 2 dimensions, or a tetrahedron in 3 dimensions, that should be initialized such that it encloses the optimum. For two corresponding dimensions the algorithm uses triangular start simplexes with the vertices $(c\Delta_1, c\Delta_2)$, $(-c\Delta_1, -c\Delta_2)$, and $(-c\Delta_1, c\Delta_2)$, where Δ_k represents the chemical shift tolerance for dimension k, and c is a random number from a normal distribution with zero mean and user-defined standard deviation $\sigma^{(s)}$. Analogous choices are made for start simplexes in more than two corresponding dimensions. The program performs a specified number of optimization runs with different start simplexes of randomly varying size. The same random number is used for all vertices, i.e. the start simplexes vary only in size but not in shape or position. Beginning with the start simplex, the algorithm then performs a number of optimization steps that move vertices of the simplex to a new position. The optimization ends when either all vertices have the same function value within a specified tolerance or 10,000 optimization steps have been performed.

The downhill simplex optimization procedure requires a general slope towards the maximum of the function in order to reach the optimum. The match score function of Eq. 2 has in general a very narrow maximum when optimizing two or more corresponding dimensions and choosing the default scaling factor $\sigma = 1$. To increase the probability for reaching the global maximum, the optimization is divided into three steps. The first step is performed with $\sigma = 10$ and $\sigma^{(s)} = 40$. The smoothed match score results in a high percentage of runs that reach the global optimum and large start simplexes increase the range of potential offsets which are covered. However, the optimum may be slightly shifted at high σ values. Therefore, two further local optimization runs with smaller σ values are added to determine the offsets with high precision. The second optimization with $\sigma = 5$ and $\sigma^{(s)} = 10$ is started from the optimum found in the first optimization. The final optimization is performed with $\sigma = \sigma^{(s)} = 1$. The same number of runs with different random start simplexes is applied in the three optimization steps.

Algorithm input and output

The input to the *Peakmatch* algorithm consists of a reference peak list and one or more target peak lists in the format of the program XEASY (Bartels et al. 1995), the general CYANA library with the magnetization transfer pathway definitions for the corresponding NMR spectra (Schmidt and Güntert 2012; Schmucki et al. 2009), and the protein sequence. Parameters that can be set by the user include the chemical shift tolerances Δ_k (default 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N dimensions), the scaling factor σ for peak matching (default $\sigma = 1$), the choice of optimization strategy (default: downhill simplex), the initial grid size parameter γ for the grid search (default $\gamma = 30$), the standard deviation $\sigma^{(s)}$ for generating start simplexes (default $\sigma^{(s)} = 40$), and the dimensions of the reference peak list for which corresponding dimensions in the target peak list(s) should be searched. In general, the default values can be used.

The algorithm outputs a summary table with the calculated optimal offsets, the initial match score and the match score after optimization for each pair of peak lists (Fig. 3a), plots overlaying the peaks from the reference and target peak list in the corresponding dimensions (Fig. 3b), and the shifted target peak lists.

Test data sets

The algorithm was evaluated with experimental data sets of five different proteins, i.e. the 140-residue ENTH-VHS domain At3g16270(9–135) from *Arabidopsis thaliana* (ENTH) (López-Méndez et al. 2004), the 134-residue rhodanese homology domain At4g01050(175–295) from *Arabidopsis thaliana* (RHO) (Pantoja-Uceda et al. 2004, 2005), the 114-residue Src homology domain 2 from the human feline sarcoma oncogene Fes (SH2) (Scott et al. 2004, 2005), ubiquitin (Ikeya et al. 2009), and the DsbA. Stereo-array isotope labeling (SAIL) was used for ubiquitin and DsbA (Kainosho and Güntert 2009; Kainosho et al. 2006). Each data set includes typical backbone experiments for resonance assignment as well as through-space experiments, i.e. [^{15}N,^{1}H]-HSQC, [^{13}C,^{1}H]-HSQC, HNCO, HN(CA)CO, CBCANH, CBCA(CO)NH, HCCH-COSY, HCCH-TOCSY (in the case of RHO only for the aromatic region), (H)CCH-TOCSY (only for DsbA and ENTH), H(CCCO)NH, ^{15}N-resolved NOESY, and ^{13}C-resolved NOESY spectra. The peak lists of all five data sets were generated automatically using automatic peak-picking algorithms of the programs NMRView (Johnson 2004) and AZARA (http://www.ccpn.ac.uk/azara) without manual corrections (Ikeya et al. 2009; López-Méndez and Güntert 2006). In addition, peak lists for ENTH, RHO, and SH2 were also available from manual, or manually curated peak picking. Details about the peak lists are given in Tables S1–S8 in the Supplementary Material. These include the number of expected peaks, the number of measured peaks, the amount of artifact peaks (peaks in the measured peak list which cannot be explained by a chemical shift within the tolerance), the completeness (the amount of expected peaks that can be found in the peak list), as well as the match score to the given reference peak list.

Results and discussion

To evaluate the performance of the algorithm, we artificially introduced different offsets into the target peak lists

a

Reference: N15H1.peaks dimensions: H+N

Peaklist	Dimensions	Offset	Match (ini)	Score (ini)	Match (opt)	Score (opt)
HNCA	N+HN	-4.995,-0.499	44.2	0.16	269.8	0.98
CBCANH	N+HN	-5.009,-0.500	72.0	0.14	440.2	0.83
C_CO_NH	N+HN	-4.997,-0.499	57.1	0.13	347.3	0.77
N15NOESY	N+HN	-4.997,-0.499	126.1	0.06	1524.9	0.79

b

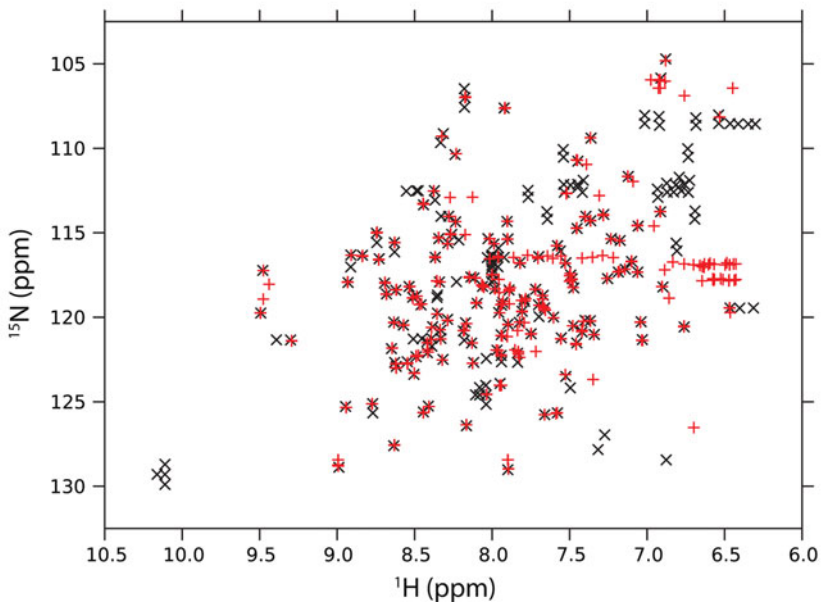

Fig. 3 Example output from the *Peakmatch* algorithm. **a** Summary file of the application of the *Peakmatch* algorithm to automatically generated peak lists for the protein ENTH with artificially introduced offsets of 5 ppm for heavy atoms and 0.5 ppm for protons. The reference peak list and the specified dimensions are reported in the first line of the output file. The match result is given for each target peak list in a separate line, which includes the corresponding dimensions, the offset for each dimension, and the absolute as well as the normalized match score prior to (ini) and after matching (opt). The absolute match score is the normalized match score *S* of Eq. 2, multiplied by the expected peak match *E* of Eq. 1, and represents the number of peaks that closely match a peak in the other peak list. **b** Example plot of the optimized HNCA target peak list, projected on the HN dimensions (*red*), and the corresponding [^{15}N,^{1}H]-HSQC reference peak list (*black*)

and back-calculated the offset using the *Peakmatch* algorithm under various conditions.

The *Peakmatch* score can be calculated for any number of corresponding dimensions in two peak lists. Almost every pair of peak lists contains one corresponding dimension. Two corresponding dimensions occur mostly for HSQC planes, and three corresponding dimensions only for few peak list pairs. The standard application of the *Peakmatch* algorithm is to optimize the offsets for two corresponding dimensions, and this will be the focus of the presentation. Offset optimization for one and three corresponding dimensions will be discussed in separate sections.

Determination of corresponding dimensions

Corresponding dimensions among two peak lists are determined automatically prior to peak list matching. To this end, expected peaks are generated for both experiment types and the expected peak match *E* of Eq. 1 is used to evaluate the different possible selections of corresponding dimensions. For many types of peak list pairs, such as typical triple-resonance backbone assignment experiments being matched to a [^{15}N,^{1}H]-HSQC spectrum, there exists only one solution with an expected peak match larger than zero. However, for example 3D NOESY spectra usually have more than one solution. By default, the solution with the largest expected peak match is chosen. Alternatively, it is also possible to optimize all solutions in independent runs. Table 1 shows expected peak match values for all NOESY spectra. In all cases the expected peak match has a significantly higher value when using the HSQC-plane for matching compared to the plane involving the other proton dimension, which means that the HSQC-plane will be the first choice for optimization.

Peak list matching for two corresponding dimensions

The performance of the *Peakmatch* algorithm was assessed using differently prepared peak lists. Manually generated peak lists are used as examples of high data quality and are thus expected to yield good results. Automatically picked peak lists, on the other hand, contain different levels of noise depending on the data set and the type of experiment. Finally, the robustness of the algorithm was evaluated systematically using a simulated SH2 data set by random deletion and addition of peaks. Downhill simplex optimization was used, unless noted otherwise.

Table 1 Expected peak match values for NOESY peak lists using the respective HSQC peak list as reference

Peak list	Dimensions	ENTH	RHO	SH2	Ubiquitin	DsbA
3D ^{15}N-resolved NOESY	N + HN[a]	1,941	1,617	1,466	871	1,495
	N + H[b]	270	211	195	150	181
3D ^{13}C-resolved NOESY	C + HC[a]	4,666	4,242	3,711	1,293	10,215
	C + H[b]	1,085	1,002	860	251	2,036

The expected peak match (Eq. 1) for ^{15}N-resolved NOESY and ^{13}C-resolved NOESY was calculated with respect to the respective HSQC peak list. Both combinations of corresponding dimensions were compared for each pair of peak lists

[a] H–N or H–C plane including the proton directly bound to the observed heavy atom

[b] H–N or H–C plane including the distant proton

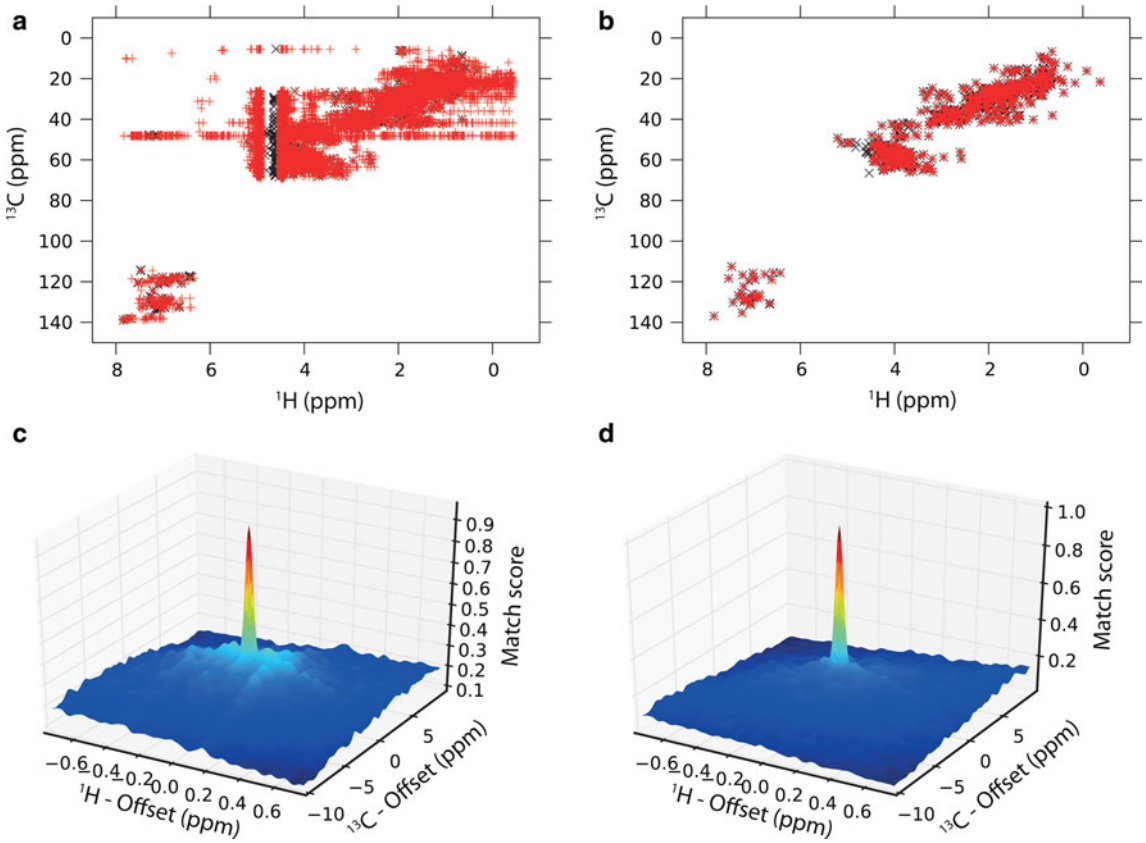

Fig. 4 Graphical representation of manually and automatically generated peak lists and the corresponding match score functions for two corresponding dimensions. Peak lists are taken from the protein ENTH. The target peak list is ^{13}C-resolved NOESY (*red* in **a** and **b**) and the reference peak list [^{13}C,^{1}H]-HSQC (*black* in **a** and **b**). **a** Peak list from automatic peak picking. **b** Peak list from manual peak picking. **c** Match score function for automatic peak picking. **d** Match score function for manual peak picking

Examples for automatically or manually prepared pairs of peak lists and the corresponding match score functions are shown in Fig. 4. The match score function for two corresponding dimensions shows a well-defined and narrow optimum at the optimal offset position even in the presence of many artifact peaks (Fig. 4c, d). This was true for every pair of peak lists in our test data sets. There might be exceptions in special cases. For instance, if there are systematic peak doublings due to very narrow lines in the spectrum of a small protein at high magnetic field, two maxima might occur in the match score function. For every doubled peak, the match score function will show two narrowly spaced maxima of approximately equal height. If one maximum is higher, a complete grid search will choose it, whereas the downhill simplex optimization might select either of the two maxima.

The *Peakmatch* algorithm was applied to all automatically or manually prepared data sets using either [^{13}C,^{1}H]-HSQC or [^{15}N,^{1}H]-HSQC as reference peak lists. This resulted in a total of 91 pairs of peak lists (Figs. S1–S8 in

the Supplementary Material). The quality of the automatically prepared peak lists depended strongly on the quality of the spectra. Especially some of the NOESY peak lists contained many noise peaks (see, for instance, Fig. 4a). Several offsets in the range between 0.1 ppm for heavy atoms and 0.01 ppm for protons and 10 ppm for heavy atoms and 1 ppm for protons were introduced into each target peak list and the *Peakmatch* algorithm was applied. Each optimization was performed using two different random starting simplexes and the offset with the highest match score was taken as the final result. An optimization result was considered correct if the difference between the introduced offsets and the calculation result was less than one-third of the chemical shift tolerance in all corresponding dimensions, i.e. if the offsets were correct within 0.01 ppm for ^{1}H and 0.13 ppm for ^{13}C and ^{15}N dimensions. Using this criterion, the *Peakmatch* algorithm found the correct offsets for all automatically or manually prepared pairs of peak lists and all four offsets tested for each peak list pair. In all cases tested the optimal offsets determined by downhill simplex optimization coincided with the correct solution. Therefore, also a complete grid search would certainly find the correct result as long as the initial grid size is larger than the offset. This shows that the maximum of the match score function of Eq. 2 describes correctly the optimal offsets also for peak lists that are far from perfect. The algorithm works reliably and with high precision over a large range of offsets for peak lists from a variety of spectra for backbone and side-chain assignment as well as ^{13}C- and ^{15}N-resolved NOESY experiments from five different proteins. The lower data quality from automatic peak picking did not have any significant effect on the offset determination by the *Peakmatch* algorithm.

The algorithm can match any combination of peak lists as long as they contain corresponding dimensions. Instead of using [^{15}N,^{1}H]-HSQC or [^{13}C,^{1}H]-HSQC reference peak lists, other suitable spectra can be chosen. For instance, we performed all offset determinations for the protein DsbA using the HSQC-planes of the (H)CCH-TOCSY and HNCO spectra as reference peak lists. In all cases the correct offsets were found.

The robustness of the algorithm was also investigated systematically with respect to missing peaks, additional artifact peaks, and small random shift changes for individual atoms among different peak lists, as might be caused for example by temperature or pH changes between the experiments. Starting from a simulated data set for the protein SH2 consisting of all expected peaks (see "Materials and methods"), randomly up to 90 % of the peaks in the reference and/or target peak lists were deleted. An offset was introduced in each target peak list and the *Peakmatch* algorithm was applied in the same way as with the experimental peak lists (Fig. 5). Deletions of up to 80 % of the peaks in either the reference or target peak list and deletion of up to 50 % of the peaks in both lists simultaneously had no effect on the offset determination. The percentage of incorrect offset determinations increased up to 4 % for deletion of 90 % of the peaks in the target peak list (Fig. 5, rhomboids), up to 20 % for 90 % deletion in the reference peak list (Fig. 5, circles), and up to 76 % for 90 % deletion in both lists simultaneously (Fig. 5, stars). The effect of noise was evaluated by adding artifact peaks at random positions up to ten times the amount of peaks in the original peak list. The addition of randomly placed artifact peaks had no effect on the offset determination up to 500 % the original number of peaks (4 times more artifact peaks than real ones), independent of whether the peaks were added to the reference peak list, the target peak list, or to both peak lists simultaneously. Adding more artifact peaks to the reference peak list, or to both peak lists simultaneously, yielded incorrect offsets in up to 16 or 47 % of the cases, respectively, for 1,100 % peaks, i.e. with 10 times more artifact peaks than real ones (Fig. 5). Addition of artifact peaks to the target peak list had no effect on the offset determination up to 1,100 % peaks.

The effect of small random shift changes for individual atoms among different peak lists was investigated starting from the same simulated data set for the protein SH2 that was also used for investigating the effect of missing peaks and artifact peaks. The chemical shift value of each atom was shifted by a normally distributed random number with

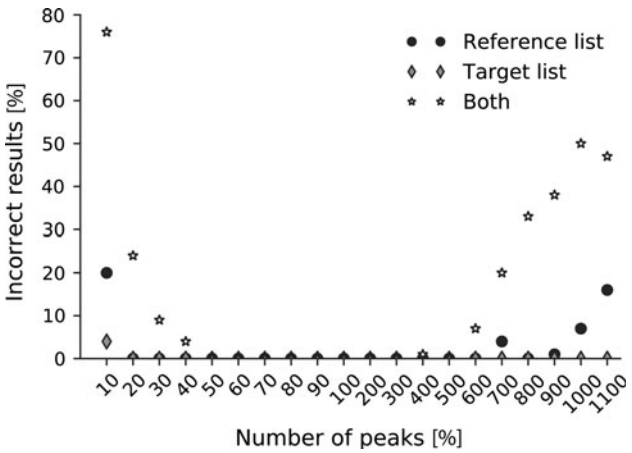

Fig. 5 Robustness of the *Peakmatch* algorithm with respect to missing peaks as well as additional artifact peaks. Starting from a simulated data set for the protein SH2 consisting of all expected peaks (see "Materials and methods"), randomly up to 90 % of the peaks were deleted, or artifact peaks were added at random positions up to ten times the amount of peaks in the original peak list in either the reference peak list (*circles*), the target peak list (*rhomboids*), or both peak lists simultaneously (*stars*). An offset was introduced into each target peak list, the *Peakmatch* algorithm was applied, and the percentage of incorrect offset determinations was measured. Every deletion or addition of artifact peaks was repeated five times using a different random number generator seed and results were averaged

a standard deviation of half the chemical shift tolerance of the respective atom type. Random shift changes were limited to a maximum of twice the respective chemical shift tolerance value. Each peak list was then generated using a different chemical shift file. A constant overall offset was introduced in each target peak list before applying the *Peakmatch* algorithm. It was found in all cases that random atom chemical shift changes up to twice the tolerance had no effect on the *Peakmatch* results. These investigations show that the algorithm is very robust with respect to data imperfections.

The number of independent downhill simplex optimization runs with different random start simplexes can be specified by the user. All optimizations mentioned were performed using two independent runs and the fact that no offset errors occurred indicates that in general two runs are sufficient. To calculate the probability that the correct offsets are found when performing n independent runs, we performed 100 runs for each optimization and took the fraction of successful runs as the probability P_1 to find the correct offsets in a single run. Assuming that individual runs are mutually independent, the probability to find the correct offsets in n runs is $P_n = 1 - (1 - P_1)^n$. Using manually prepared peak lists, the percentage of correct optimizations was on average 99 % and in all cases above 95 %. This corresponds to an average probability of 99.99 % and a minimum probability of 99.75 % that two independent runs will yield the correct result. When using peak lists from automatic peak picking, the percentage of correct optimizations was on average 97 %, and the minimal percentage was 88 %. This leads to an average probability of 99.91 % and a minimal probability of 98.56 % that two independent runs will yield the correct result.

The match score S of Eq. 2 is normalized by the expected match score E of Eq. 1. For a perfect match of two ideal peak lists one thus obtains $S = 1$. The optimal match score for experimental peak lists, however, depends strongly on the quality of the peak lists. Missing peaks decrease and additional peaks potentially increase the score, which makes it difficult to judge the result of an offset determination simply by the match score value. The normalized match score values of all individual calculations performed with manually prepared peak lists were 0.19–1.08 (average 0.70) for the correct results and 0.04–0.14 (average 0.11) for the optimizations yielding incorrect results. The corresponding score values for the automatically prepared peak lists were 0.27–1.62 (average 0.89) for the correct results and 0.08–0.85 (average 0.21) for the optimizations yielding incorrect results. On average the correct results have thus much higher score values than the incorrect ones. Nevertheless, correct and incorrect results cannot be separated clearly by their individual match score values. In particular, the results for automatically prepared peak lists include correct results with match scores as low as 0.27 as well as incorrect results with match scores up to 0.85. Since it is not straightforward to distinguish correct from incorrect results by the match score value, the overlay of the peaks (projected onto the corresponding dimensions, if necessary) in the reference peak list and the optimally shifted target peak list is visualized (Fig. 3b). Based on this diagram the user can evaluate the result and decide whether to use the optimized peak lists or not.

The runtime of the algorithm depends on the number of peaks in the reference and target peak lists, and on the number of evaluations of the match score function of Eq. 2. The number of function evaluations differs for the different optimization procedures. We compared the runtime for downhill simplex optimization, a grid search with a limited grid size of 2 ppm for heavy atoms and 0.15 ppm for protons, and a grid search with a larger grid of 10 ppm for heavy atoms and 0.75 ppm for protons using an Intel E5-2690 2.9 GHz processor. The shortest average runtime of 3.9 s occurred for the grid search with limited grid size (281 function evaluations). The average calculation time using the downhill simplex optimization procedure was 5.2 s (on average 500 function evaluations), and the largest average calculation time of 33.5 s was required for the larger grid search (2,731 function evaluations). Except in the case of small expected offsets, it is thus most efficient to use downhill simplex optimization.

Peak list matching for one corresponding dimension

There are target peak lists that have only one corresponding dimension in common with the reference peak list. This makes the correct matching more difficult than with two or more corresponding dimensions. We tested the performance of the *Peakmatch* algorithm using only one corresponding dimension. The match score function for one corresponding dimension does in general not show a single narrow maximum, but instead a larger number of local optima (Fig. 6c, d). Since the downhill simplex optimization might be trapped in local optima and the calculation time is not an issue for one-dimensional optimization, we limited the optimization procedure to a full grid search. We used again the aforementioned manually or automatically prepared peak lists and [^{13}C,^{1}H]-HSQC or [^{15}N,^{1}H]-HSQC as reference peak lists and performed a one-dimensional grid search to determine the optimal chemical shift offset, which was considered correct if it was within the chemical shift tolerance Δ_1 for the corresponding dimension, i.e. 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N.

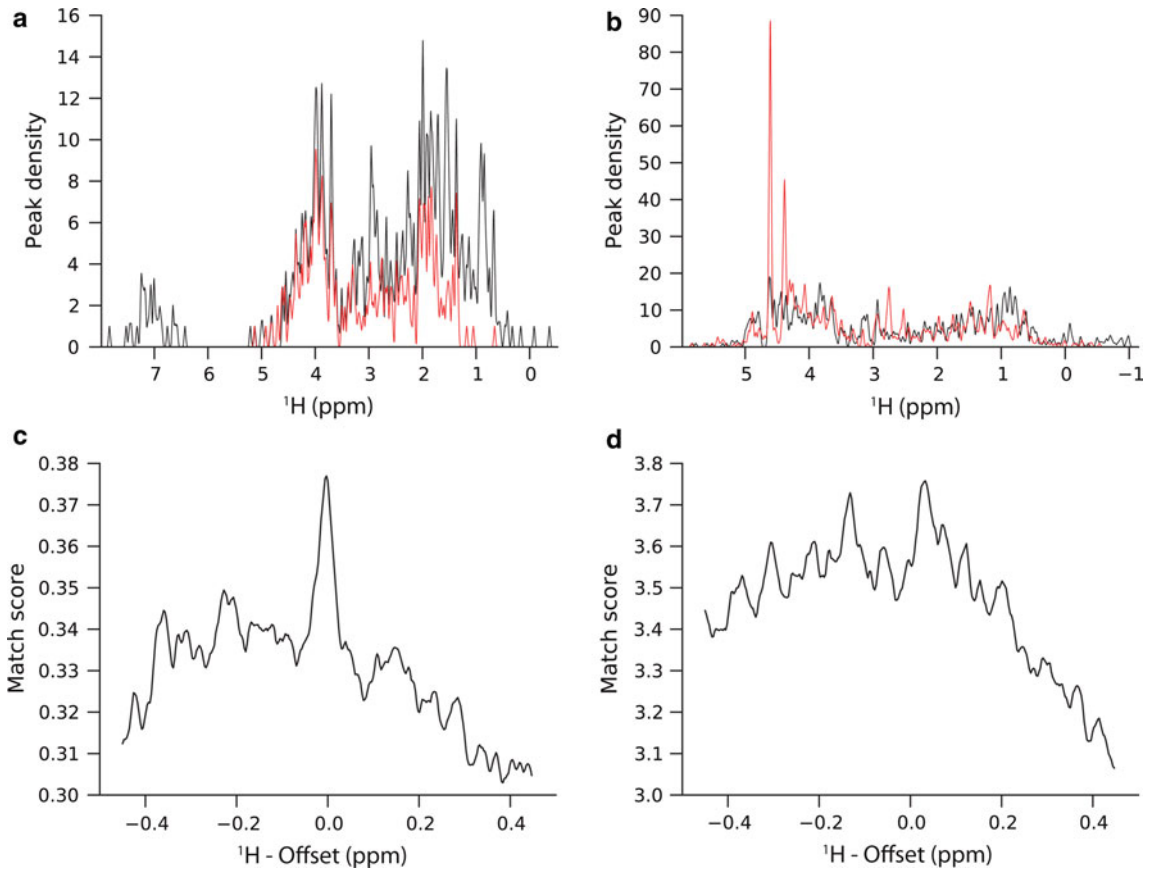

Fig. 6 Graphical representation of manually and automatically generated peak lists and corresponding match score functions for one corresponding dimension. **a** Peak density for manually prepared peak lists from HBHACONH (*black*) and [^{13}C,^{1}H]-HSQC (*red*) spectra of the protein ENTH, obtained by plotting a Gaussian lineshape of unit height and standard deviation 0.03 ppm at the ^{1}H position of each peak. **b** Peak density for automatically picked peak lists from HC(CO)NH (*black*) and [^{13}C,^{1}H]-HSQC (*red*) spectra of the protein RHO. **c** Match score function for the manually prepared peak lists from (**a**). **d** Match score function for the peak lists from automatic peak picking in (**b**)

When using manually prepared peak lists, the match score function showed in all cases a global optimum at the correct solution (Fig. 6c). However, there are many local optima with match score values of similar magnitude, which makes it difficult to distinguish the correct from incorrect solutions based on the match score value. The superposition of the one-dimensional projections of the peak lists (Fig. 6a) shows that the peak lists overlay nicely, which explains the fact that all match score functions have their global maximum at the correct offset. In the case of automatically picked peak lists, however, the global optimum does in many cases not represent the correct solution. One example is shown in Fig. 6d. The graphical representation of the one-dimensional projections of the peak lists (Fig. 6b) show that it is difficult to see how the two peak lists should be overlayed correctly, reflecting the fact that the match score function has multiple maxima of similar size.

These results indicate that the *Peakmatch* algorithm can also be used with only one corresponding dimension if good quality input data is available but results are much less reliable than with two or more corresponding dimensions and should be corroborated by visually checking the superposition of the one-dimensional projections of the peak lists.

Peak list matching for three corresponding dimensions

The Peakmatch algorithm can match any number of corresponding dimensions, even though in practice more than two corresponding dimensions occur rarely. One application is adapting the data from two three-dimensional spectra of the same type recorded under slightly different experimental conditions. As an example, we tested a pair of automatically picked peak lists for ENTH with three corresponding dimensions, CBCANH and CBCACONH, and performed a grid search as well as downhill simplex optimization. Both optimization strategies yielded the correct solution. The computation time for the grid search was 60 s due to a large number of function evaluations. In contrast,

Table 2 Results of automatic chemical shift assignment using peak lists with artificial random offsets

Protein	Correct chemical shift assignments for backbone/all atoms (%)		
	Small offsets[a]	Large offsets[b]	Optimized data set[c]
ENTH	94.4/88.4	81.6/76.4	95.5/90.8
RHO	93.6/88.0	88.4/74.4	96.2/90.1
SH2	97.7/89.8	96.9/85.2	98.4/90.8

Each of the automatically picked input peak list for the FLYA automated chemical shift assignment algorithm {Schmidt, 2012 #2030} was shifted independently by a uniformly distributed random number within the specified range

[a] Small offsets in a range of ±0.03 ppm for protons and ±0.4 ppm for heavy atoms. The *Peakmatch* algorithm was not used

[b] Large offsets in a range of ±0.045 ppm for protons and ±0.6 ppm for heavy atoms. The *Peakmatch* algorithm was not used

[c] The data set with large offsets was subjected to offset correction with the *Peakmatch* algorithm prior to automatic chemical shift assignment. All results are given as the percentage of correctly assigned atoms with respect to the reference assignment for either the backbone atoms (first number) or all atoms (second number)

the computation time for the downhill simplex optimization increased only to 11.4 s. Compared to the two-dimensional case, the number of function evaluations by the downhill simplex algorithm did not increase significantly and the increased runtime resulted mainly from the longer computation time for a single function evaluation.

Peak list matching against a chemical shift list

The *Peakmatch* algorithm can also be used to find the optimal offsets to match a peak list to a given chemical shifts list. For instance, NOESY peak lists can be matched to a chemical shift list obtained from through-bond spectra prior to automated NOE assignment and structure calculation. To this end, a reference peak list of the same spectrum type as the target peak list is simulated on the basis of the sequence and the chemical shift list, and then used as input to the *Peakmatch* algorithm treating all spectral dimensions as corresponding dimensions. Again, this approach has the advantage that it can be applied to unassigned peak lists.

Example *Peakmatch* application

Automatic chemical shift assignment is a possible application of the *Peakmatch* algorithm. We used the automatically picked data sets of the proteins ENTH, SH2, and RHO for automatic chemical shift assignment using the FLYA algorithm (Schmidt and Güntert 2012). To illustrate the consequences of chemical shift referencing inconsistencies among different peak lists of the same data set, we artificially introduced random constant offsets into each peak list prior to automatic chemical shift assignment. Offsets were introduced within either 1.0 or 1.5 times the assignment tolerance (0.03 ppm for protons and 0.4 ppm for heavy atoms), and assignment results were compared to those for the optimized input data set, which was obtained using the *Peakmatch* algorithm. A summary of results is presented in Table 2. In the presence of uncorrected offsets within the assignment tolerance the FLYA automated assignment algorithm yielded 88.0–89.8 % correct assignments for all atoms (first column in Table 2). Using peak lists with larger chemical shift referencing offsets of up to 1.5 times the tolerance, the amount of correct assignments decreased to 74.4–85.2 % (second column in Table 2). The latter results can be improved to 90.1–90.8 % when optimizing the offsets with the *Peakmatch* algorithm (third column in Table 2). This demonstrates the significant improvement of assignment results that can be achieved by applying the *Peakmatch* program prior to automatic chemical shift assignment with the FLYA algorithm.

Conclusions

In this paper we have presented a new algorithm that determines the optimal offset between two multidimensional peak lists that contain corresponding dimensions. The algorithm identifies corresponding dimensions automatically based on the expected peaks for the given experiments and then optimizes a match score function for the experimental peak lists. Extensive tests showed that the algorithm works very reliably also with input peak lists that are far from ideal, e.g. those generated by automatic peak picking programs, provided that there are at least two corresponding dimensions. Principal advantages of the algorithm are that (1) it can be applied to unassigned peak lists, (2) it is highly tolerant against the common imperfections of experimental peak lists, (3) the criterion for optimal matching is mathematically simple and largely captures what an experienced spectroscopist would do manually, and (4) its application is straightforward and quick.

The optimization can be performed using a complete grid search or a downhill simplex optimization procedure. In all test cases, both procedures performed equally well when using two corresponding dimensions. When using a single corresponding dimension a complete grid search is recommended as the downhill simplex algorithm has a higher chance of getting trapped in a local optimum and computation time is no issue in the one-dimensional case. For more than two corresponding dimensions both methods are equally reliable. However, the complete grid search can be time consuming depending on the grid size, whereas the computation time for the downhill simplex procedure rises

only slightly with increasing number of corresponding dimensions.

Acknowledgments We thank Dr. T. Ikeya, Dr. M. Takeda, and Prof. M. Kainosho for the DsbA peak lists. We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation, the Deutsche Forschungsgemeinschaft (DFG), and the Bio-NMR project of the European Commission.

References

Aeschbacher T, Schubert M, Allain FHT (2012) A procedure to validate and correct the ^{13}C chemical shift calibration of RNA datasets. J Biomol NMR 52:179–190

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

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

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

Ginzinger SW, Gerick F, Coles M, Heun V (2007) CheckShift: automatic correction of inconsistent chemical shift referencing. J Biomol NMR 39:223–227

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

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

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

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

Ikeya T, Takeda M, Yoshida H, Terauchi T, Jee J, Kainosho M, Güntert P (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. J Biomol NMR 44: 261–272

Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. Meth Mol Biol 278:313–352

Kainosho M, Güntert P (2009) SAIL—stereo-array isotope labeling. Q Rev Biophys 42:247–300

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440:52–57

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

López-Méndez B, Pantoja-Uceda D, Tomizawa T, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from *Arabidopsis thaliana*. J Biomol NMR 29:205–206

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313

Pantoja-Uceda D, López-Méndez B, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical rhodanese domain At4g01050 from *Arabidopsis thaliana*. J Biomol NMR 29:207–208

Pantoja-Uceda D, López-Méndez B, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Seki M, Shinozaki K, Yokoyama S, Güntert P (2005) Solution structure of the rhodanese homology domain At4g01050(175–295) from Arabidopsis thaliana. Protein Sci 14:224–230

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

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

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2004) NMR assignment of the SH2 domain from the human feline sarcoma oncogene FES. J Biomol NMR 30:463–464

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2005) Solution structure of the Src homology 2 domain from the human feline sarcoma oncogene Fes. J Biomol NMR 31:357–361

Wang YJ, Wishart DS (2005) A simple method to adjust inconsistently referenced ^{13}C and ^{15}N chemical shift assignments of proteins. J Biomol NMR 31:143–148

Wang LY, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13–22

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