J Biomol NMR (2012) 54:81–95
DOI 10.1007/s10858-012-9652-3

ARTICLE

TSAR: a program for automatic resonance assignment using 2D cross-sections of high dimensionality, high-resolution spectra

Anna Zawadzka-Kazimierczuk · Wiktor Koźmiński · Martin Billeter

Received: 5 April 2012 / Accepted: 29 June 2012 / Published online: 18 July 2012

Abstract While NMR studies of proteins typically aim at structure, dynamics or interactions, resonance assignments represent in almost all cases the initial step of the analysis. With increasing complexity of the NMR spectra, for example due to decreasing extent of ordered structure, this task often becomes both difficult and time-consuming, and the recording of high-dimensional data with high-resolution may be essential. Random sampling of the evolution time space, combined with sparse multidimensional Fourier transform (SMFT), allows for efficient recording of very high dimensional spectra (≥4 dimensions) while maintaining high resolution. However, the nature of this data demands for automation of the assignment process. Here we present the program TSAR (*T*ool for *S*MFT-based *A*ssignment of *R*esonances), which exploits all advantages of SMFT input. Moreover, its flexibility allows to process data from any type of experiments that provide sequential connectivities. The algorithm was tested on several protein samples, including a disordered 81-residue fragment of the δ subunit of RNA polymerase from *Bacillus subtilis* containing various repetitive sequences. For our test examples, TSAR achieves a high percentage of assigned residues without any erroneous assignments.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9652-3) contains supplementary material, which is available to authorized users.

A. Zawadzka-Kazimierczuk · W. Koźmiński
Faculty of Chemistry, University of Warsaw,
Pasteura 1, 02-093 Warsaw, Poland

M. Billeter (✉)
Biophysics Group, Department of Chemistry and
Molecular Biology, University of Gothenburg, Box 462,
405 30 Gothenburg, Sweden
e-mail: martin.billeter@chem.gu.se

Keywords Algorithm · Automated resonance assignment · High-dimensional fast NMR · Intrinsically disordered protein

Introduction

The first steps in protein characterizations by NMR consist of sample preparations, recording of spectra and assignment of resonances. Once chemical shifts are determined for (nearly) all nuclei considered, further experiments can provide a wealth of information on the protein, describing structure, dynamics, function and interactions. Besides the preparation of an NMR-friendly sample, the limiting barrier for a successful protein study by NMR is the resonance assignment; here the protein size, its folding state (multiple conformations, intrinsically disordered…) and, related to this, the complexity and quality of the spectra are critical. A central step is the detection of sequential connectivities between adjoining amino acid residues. Inspection of chemical shifts, such as those for Cβs, adds useful information by limiting the number of consistent types of residues.

3D triple-resonance spectra like HNCO (Kay et al. 1990), HN(CA)CO (Clubb et al. 1992), HNCA (Kay et al. 1990), CBCA(CO)NH (Grzesiek and Bax 1992a), CBCANH (Grzesiek and Bax 1992b) are widely used for this purpose. However, if the extent of spectral overlap of the NMR signals (peaks) is high, the spectra interpretation and assignment becomes complicated and time-consuming. Spectra of both high dimensionality and high resolution reduce overlap and ensure good precision in the determination of peak positions, both of which are critical for resonance assignment procedures. Such spectra can be effectively acquired with the use of non–Cartesian sampling in evolution time space by dramatically reducing the

number of acquired FIDs [recently reviewed in (Coggins et al. 2010; Freeman and Kupče 2012; Hiller and Wider 2012; Hyberts et al. 2012; Kazimierczuk et al. 2010a, 2012; Maciejewski et al. 2012; Orekhov and Jaravine 2011). One of the most general sampling schemes is random sampling, which allows for optimal resolution also in ≥4D experiments (Kazimierczuk et al. 2009). Employing this technique enabled us recently to develop a set of 4D (Zawadzka-Kazimierczuk et al. 2010), as well as 5D and 6D (Kazimierczuk et al. 2010b; Zawadzka-Kazimierczuk et al. 2012) experiments dedicated for easy resonance assignment. Data from these experiments can be processed using sparse multidimensional Fourier transform (SMFT) (Kazimierczuk et al. 2009), resulting in a set of 2D cross-sections. Each cross-section contains peaks originating from adjacent residues yielding sequential connectivities. An important feature is that a single common *basis* spectrum (usually 3D HNCO) is used to define cross-sections from several spectra. This allows collecting cross-sections from various spectra types that correspond to the same spin system, and thus assembling of resonances originating from various spectra into one spin system. Resonance assignment of data obtained by SMFT is realized by sorting cross-sections: peaks, which are identical in cross-sections for adjacent residues, are used to build chains of spin systems that correspond to fragments of the polypeptide. This task can be performed manually; however, it seems to be a perfect target for a computer algorithm.

Several algorithms have been designed for automated sequence-specific signal assignment, reviewed in (Baran et al. 2004; Altieri and Byrd 2004). The protocol of many programs consists of three steps: (1) collecting peaks into spin systems, (2) forming chains of these spin systems which correspond to consecutive residues, (3) mapping of these chains onto the (known) protein sequence. Step (1) is an important issue in automatic assignment, nonetheless, it is often omitted, which consequently requires user-defined spin systems as input, e.g. for the MARS program (Jung and Zweckstetter 2004). The main differences among the algorithms concern their approach to steps (2) and (3). Some algorithms exploit a global optimization approach, where a function describing assignment quality is defined and optimized [e.g. the program GARANT (Bartels et al. 1996, 1997)], other use a "best-first" strategy based on local optimization [e.g. the program AUTOASSIGN (Zimmerman et al. 1997)], a combination of both approaches [programs MARS (Jung and Zweckstetter 2004), MATCH (Volk et al. 2008), PASA (Xu et al. 2006)], or present all "acceptable" assignments so that the user can decide on their quality [program SAGA (Crippen et al. 2010)]. Input data for all above programs includes the amino acids sequence of protein and lists with spectral information. The variety of requirements for the lists is important for the motivation for the present work and thus requires deeper discussion. In some algorithms the peak lists from various experiments can be directly used, in others a special list has to be constructed from the peak lists by the user. Even in the former case the algorithms differ in the types of required peak lists. The program GARANT is very flexible: one can define any kind of experiment and use any peak list. In particular, the following feature is allowed, which will be referred to as *not-fully specified columns*: columns of the peak list may contain frequencies of various nuclei, e.g. CO_i or CO_{i-1} (like in a HN(CA)CO spectrum), $C\alpha_{i-1}$ or $C\beta_{i-1}$ (like in a CBCA(CO)NH spectrum) etc. A similar degree of flexibility is also implemented for MATCH. AUTOASSIGN can take lists with not-fully specified columns if each is complemented with a list containing a corresponding fully specified column, so that the program can distinguish between the two types of peaks in the former list. Moreover, AUTOASSIGN is able to find connectivities only via CO, Cα and Cβ chemical shifts. SAGA was designed to take exclusively 3D data from experiments pre-defined in the program. Similarly, the program AUTOBA (Borkar et al. 2011) is adjusted for input from the HN(C)N suite of experiments, either using the 3D versions or just 2D projections. The MARS and PASA programs only accept a single list containing frequencies of nuclei belonging to one spin system (the columns have to be fully specified) rather than experimental peak lists. Therefore preparation of such a list requires a priori thorough analysis of peak lists, which can prove difficult in cases with insufficient signal dispersion.

Despite the fact, that some of the above algorithms are flexible enough to analyze data from various (though usually not all) types of experiments presented in our recent works (Zawadzka-Kazimierczuk et al. 2010, 2012, Kazimierczuk et al. 2010b), none of them makes optimal use of the information present in a set of 2D cross-sections obtained with SMFT. This concerns both taking full advantage of all available input spectra, and in particular providing results in a form that allows feedback for user intervention in ambiguous situations. Regarding the input, while most of the programs can use spectra which involve two adjacent residues, only few can process data from magnetization pathways that involve three consecutive residues, e.g. (H)NCO(NCA)CONH contains peaks of type N_i–CO_{i-1}–CO_i–N_{i+1}–H^N_{i+1} which join i − 1, i and i + 1 residues. Furthermore, while the type of some peaks can be uniquely characterized, this is not always the case. Consider the (frequent) example of more than one peak on each cross-section, e.g. on the CA-CO planes in 4D HNCACO spectra there are $C\alpha_i$–CO_i and $C\alpha_{i-1}$–CO_{i-1} peaks, which cannot be distinguished at this stage. Some programs have problems in handling these ambiguities. Similarly, when assigning amino acid types to a particular spin system, a program should not only consider chemical shift statistics but also the signs of peak amplitudes.

For instance, in a 5D HN(CA)CONH the signs of all peak amplitudes in the corresponding cross-section change if the preceding residue is glycine (due to the lack of Cα–Cβ coupling). The signs depend on the pulse sequence employed and on the phase correction applied, and are thus defined in the input file. Finally, the output of an assignment program for SMFT data should, besides listing chemical shifts, also identify the corresponding 2D cross-sections in the case of unambiguous assignments; this would enable collecting the spectral data necessary for manual inspection of ambiguous parts of the assignment.

In this paper we present a program called TSAR (*T*ool for *S*MFT-based *A*ssignment of *R*esonances) for automatic resonance assignment in proteins, which is designed to use 2D cross-sections of one or several high-dimensional spectra obtained with SMFT, exploiting all features of this type of input. The types of spectra may be chosen out of the set proposed by us (Zawadzka-Kazimierczuk et al. 2010, 2012, Kazimierczuk et al. 2010b), but the program accepts data of any type of experiments.

Methods

Principle of operation

The experimental data required by the program is one spectrum of lower dimensionality (*basis spectrum*) and one or more spectra of high dimensionality ($\geq$4D). The basis spectrum should contain exactly one peak per amino acid residue and the peaks should ideally be well-separated (this issue will be discussed in detail in the "Discussion" section). In all experimental examples presented here, a 3D HNCO spectrum was used as basis spectrum. The high dimensionality spectra should contain some (or all) dimensions of the basis spectrum and two additional dimensions in order to enable SMFT processing (Kazimierczuk et al. 2009). Note that in various spectra different dimensions of the basis spectrum can be used for calculating SMFT cross-sections. It is essential to calculate cross-sections of various spectra relying on the same peak list from the basis spectrum (i.e. containing the peaks in the same order). This ensures that information of corresponding cross-sections from various spectra can be properly collected in a structure referred to as cross-section spin system or CSSS. Each CSSS contains chemical shifts corresponding to the HN, N, CO, Cα, Cβ, Hα and Hβ nuclei of up to three consecutive residues. All spectra together should contain sufficient information to form sequential connectivities.

The principle of operation of the program is shown in Fig. 1 and details of the algorithm are explained in the sections below. In short, after reading the input data, one CSSS per peak in the basis spectrum is created. At first only the chemical shifts known from the basis spectrum are entered into the appropriate sites in each CSSS. Next, peaks from cross-sections (originating from all other

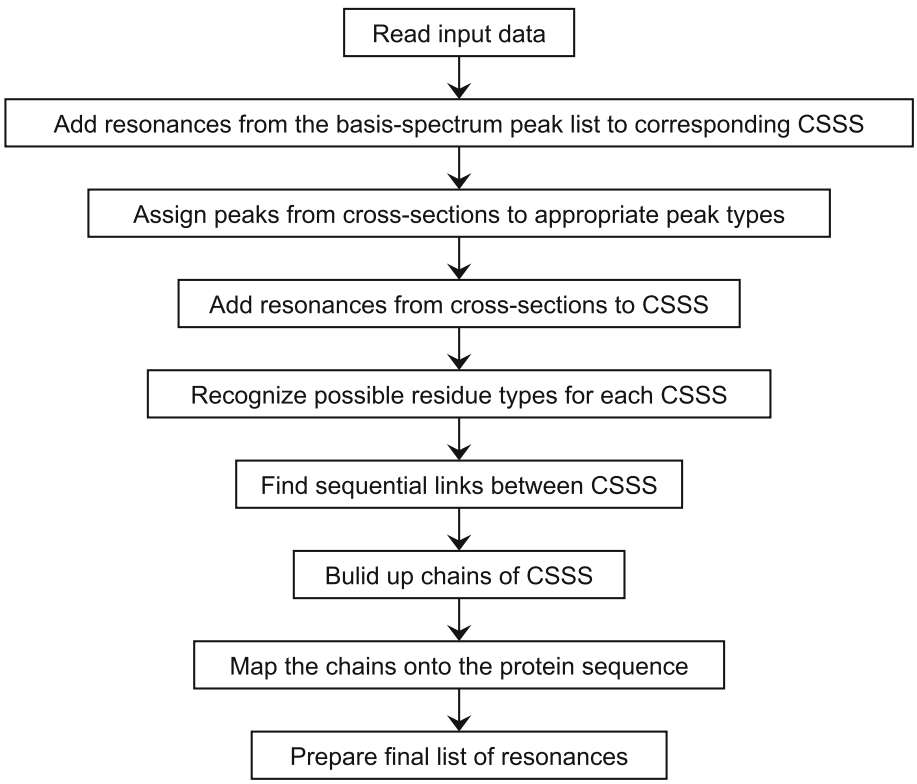

Fig. 1 Principle of operation of the TSAR program. The individual steps of this flow-chart are explained in "Methods"

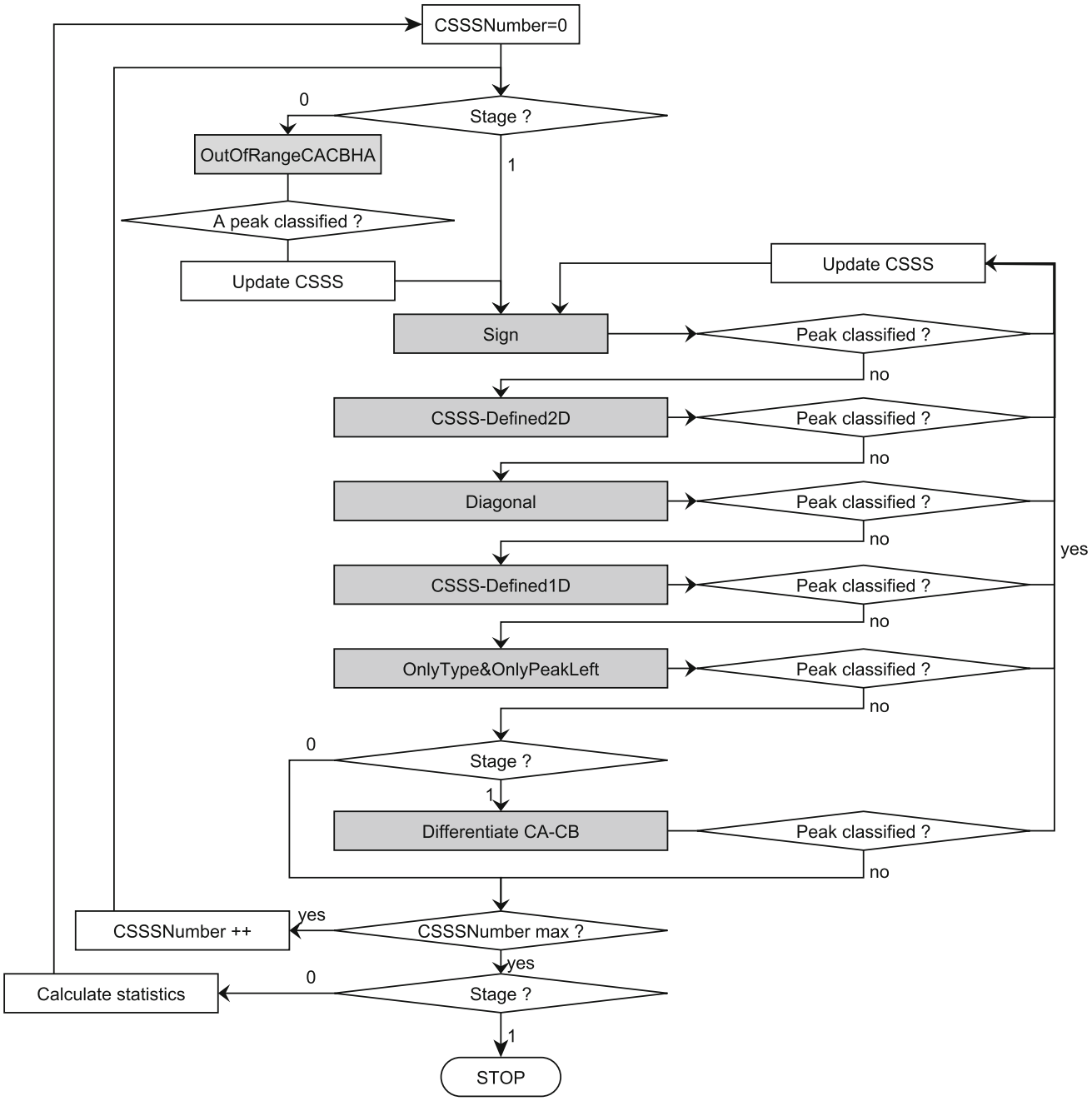

Fig. 2 Flow-chart summarizing the assignment of peaks to peak types. See "Methods" for explanations

spectra) are assigned to peak types (i.e. specific nuclei combinations), which are defined for each spectrum in an input file. During this process the CSSS structures are successively filled with newly obtained chemical shift values. When completed, the types of amino acid residues compatible with a given CSSS are determined. The next step is finding sequential connectivities between pairs of CSSSs. Chains of CSSSs are then formed by sorting the cross-sections, such that the order of the residues, from which they originate, corresponds to the protein sequence. During this mapping step information about possible residue types is exploited.

Input

The input includes the protein sequence, descriptions of the experiments used and peak lists (Tables S1 and S2 in Supplementary Material), all in the form of text files. The description of the basis spectrum consists of the names and relative positions (in terms of residues) of nuclei for all

dimensions (e.g. CO_{i-1}, N_i and H_i^N). The definitions of higher-dimensional experiments are more complex and contain information about all types of peaks expected in the cross-sections: types of nuclei and relative positions, amplitude sign and whether the peak can switch sign under specific conditions (e.g., when the preceding residue is a glycine). Moreover, for each dimension of each spectrum the chemical shift tolerances are needed; these are initially best set to the inverse of the maximum evolution time for the corresponding dimension (in ppm). A more thorough discussion of these tolerances is given in the next section. It should be noted that relative positions of nuclei in individual spectra can include, besides $i - 1$ and i, also $i + 1$.

Assignment of peaks to peak types

On each cross-section there may be several peaks of various types, e.g. in a 4D HNCACACB spectrum four peaks are expected on any (non-glycine) CA-CAB cross-section: $C\alpha_i$–$C\alpha_i$ (positive) and $C\alpha_{i-1}$–$C\alpha_{i-1}$ (positive), $C\alpha_i$–$C\beta_i$ (negative) and $C\alpha_{i-1}$–$C\beta_{i-1}$ (negative). The program uses all available information in order to attribute a peak type to each peak. The corresponding algorithm, shown in Fig. 2, also needs to handle situations where cross-sections do not show the expected number of peaks due to artifacts (false peaks) or missing peaks.

There are several methods enabling recognition of a peak type. They are arranged such that the most reliable ones are applied first. Importantly, every new peak assignment may add information about chemical shifts to a CSSS, which can help in the assignment of other peaks. Therefore the methods are used in a loop (separately for each CSSS): following any new peak assignment, the procedure is restarted from the beginning. Some methods to assign a peak type rely on comparisons of chemical shifts, which in turn are based on tolerances. The whole procedure is performed in two stages: At first the usually rather large initial tolerances, which are read from the input file and typically equal to the inverse of corresponding maximum evolution times, are used. At this stage only unambiguous assignments are made. Once as many assignments as possible are obtained with these initial tolerances, statistical values of differences in corresponding chemical shifts for each pair of spectra become available: mean and standard deviation are calculated, and the initial tolerances are replaced with four times the appropriate standard deviation (provided the new value is smaller than the old one).

The following describes the methods for peak type recognition (see Fig. 2):

- *OutOfRangeCACBHA*—for spectra in which Cα and Cβ (Hα and Hβ) resonances appear in the same dimension: if an aliphatic carbon chemical shift is outside of the Cα range ($C_{aliph} < 39.5$ ppm) for a peak, all peak types which originate from a Cα coherence are excluded. A similar procedure is applied to aliphatic carbon resonances outside of the Cβ range (54.5 ppm $< C_{aliph} < 58.5$ ppm) and to aliphatic proton resonances outside of the Hα range ($H_{aliph} < 1.85$ ppm). The ranges in all three cases were established using the mean and four standard deviations taken from the Biological Magnetic Resonance Bank (BMRB, (Ulrich et al. 2008)) statistics for the "restricted" set with only diamagnetic proteins.
- *Sign*—for each peak the peak types with opposite amplitude sign are excluded. If the peak can change its sign (e.g. in the presence of glycine) this function is not used until another method clarifies whether the signs are standard or changed.
- *CSSS-Defined2D*—if a particular type of peak has both chemical shifts already present in the CSSS chemical shifts structure, then this type is either excluded or confirmed for each peak on this cross-section. If there is more than one peak which is close to the defined position and the statistics is available (second stage of the procedure), the closest one is evaluated according to the so-called distribution condition (for details see the bottom of this section). If there is no peak at the defined position, it is artificially added (but with unknown sign).
- *Diagonal*—if a particular type of peak is diagonal on a cross-section, it is excluded for peaks with two different chemical shift values. If there is only one diagonal type not assigned to any peak and one diagonal peak left, the corresponding assignment is confirmed.
- *CSSS-Defined1D*—if a particular type of peak has one dimension already present in the CSSS structure, this type is excluded for peaks with different value of this chemical shift. If there is exactly one peak close to the defined position, it is assigned to the considered peak type. If there is more than one peak close to the defined position, there are at least as many peaks on the cross-section as expected and the statistics is available (second stage of the procedure), then the closest one is evaluated according to the distribution condition (for details see the bottom of this section).
- *OnlyType&OnlyPeakLeft*—if in a particular spectrum there is only one type of peak left (not assigned to any peak) and on a given cross-section there is only one peak left (not assigned to any peak type), then this peak is assigned to the remaining peak type.
- *DifferentiateCA-CB*—if on a cross-section there is the expected number of peaks, only two of them are not assigned to any type and the types left originate from different nuclei type, namely Cα and Cβ, this method

tries to distinguish which peak corresponds to Cα and which to Cβ. If both chemical shifts are lower than 54.5 ppm (upper limit for Cβ for residues other than serine or threonine) and the difference in these chemical shifts is bigger than a safety margin (set to 1 ppm), the peak with the lower chemical shift is assigned to the peak type originating from Cβ and the peak with the higher chemical shift to the type originating from Cα. If both chemical shifts are higher than 58.5 ppm (lower limit for Cβ for residues of serine or threonine) and the difference in these chemical shifts is bigger than the safety margin, the peak with the higher chemical shift is assigned to the peak type originating from Cβ and the peak with the lower chemical shift to the type originating from Cα.

During the second stage of the procedure, when statistics from early assignments is available, two methods (*CSSS-Defined2D* and *CSSS-Defined1D*) rely on a distribution condition. If several peaks are close to the position known from CSSS, the closest is accepted to be of the considered type provided the following condition is fulfilled. Two peaks are considered: the geometrically closest from the known position (peak *i*) and the second closest (peak *j*). Peak *i* is accepted if the following inequality is fulfilled (for *CSSS-Defined2D* it is checked in both dimensions):

$$\begin{aligned}&(cdf(|pos_i - pos - \mu|) - 0.5)\\ &\quad \cdot 1.5 < \left(cdf\left(\left|pos_j - pos - \mu\right|\right) - 0.5\right)\end{aligned} \tag{1}$$

where *pos* is the position known from CSSS, pos_i is the position of the peak *i*, pos_j is the position of the peak *j* (all positions in the considered dimension), cdf is the cumulative distribution function for the corresponding normal distribution with mean μ and standard deviation σ, describing the probability that the function's argument takes on a value less or equal to that argument. Thus for positive x:

$$cdf(x) - 0.5 = \int_0^x \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(-\frac{(x' - \mu)^2}{2\sigma^2}\right)dx' \tag{2}$$

If the condition is not fulfilled for the *CSSS-Defined2D* method, the closest peak is still accepted, but its intensity sign is changed to 'unknown'.

Recognition of amino acid types

Identification of compatible amino acid types for each CSSS strongly limits the number of possible choices for mapping of the CSSS onto the protein sequence. However, at the initial stage the conditions for excluding certain amino acid types are formulated in a weak way not to accidentally exclude the correct one; further elimination, when necessary, will be performed during the mapping stage. For the initial recognition of amino acid types the TSAR program employs two types of methods: structural and statistical.

The structural methods are based on the knowledge of the chemical structure of amino acids. Some of them examine whether a certain type of nuclei is present in the given CSSS (individually for all relative positions). Thus, if a H^N chemical shift is recognized, the residue cannot be proline; if a Cβ or Hβ chemical shift is present, glycine is excluded from the set of possible amino acid types; if there are two different Hα chemical shifts (corresponding to two nuclei of a pair of prochiral methylene protons), all residue types except for glycine are excluded, and—similarly—if there are two different Hβ chemical shifts, all types which contain only one Hβ proton (or a methyl group) are excluded: alanine, isoleucine, threonine and valine. Other types of structural methods can be exploited depending on the experimental technique used. If Cα–Cβ scalar coupling evolves during an experiment for the time of approximately $^1/_{J_{C\alpha-C\beta}}$, the sign of all peak amplitudes are inverted, except for those, which originate from glycine. This results in a relative change of sign of some glycine-related peaks. For example, in the HN(CA)CONH spectrum, on HN–N cross section corresponding to CO_{i-1}–N_i–H_i^N HNCO peak, two peaks with opposite signs are expected for H_i^N–N_i and H_{i-1}^N–N_{i-1}. However, if residue i − 1 is glycine, both signs are reversed.

The statistical methods are based on BMRB statistics and involve Cα, Cβ, Hα and Hβ chemical shifts. Statistics on β-nuclei are performed for each type of amino acid: a residue type is excluded, if the observed chemical shift is over four standard deviations away from the BMRB mean value. Additionally, in the case of Cβ the two forms of cysteine (oxidized and reduced) were treated separately. Since only joint statistics for both forms are directly available from the BMRB, the parameters for the two distributions were calculated based on the joint BMRB histogram by fitting the sum of two Gaussian functions, yielding the following values:

- oxidized form: mean = 40.89 ppm, standard deviation = 3.97 ppm
- reduced form: mean = 29.09 ppm, standard deviation = 2.51 ppm

As Cα and Hα chemical shifts exhibit less diversity compared to Cβ and Hβ, only glycine residues are recognized at this stage using chemical shifts of the α-nuclei: if Cα or Hα is above the mean plus four standard deviations for glycine, this amino acid type is excluded.

Should the procedure of amino acid type recognition exclude all types of residues (due to some input error), then all types are allowed again, and the program proceeds after issuing a warning with the request for a manual check.

Sequential connectivities between CSSSs, chains of CSSSs

Prior to establishing sequential links, pairs of mismatching CSSSs are identified; these are pairs which have at least two peaks of corresponding types which do not match (within at least one type of spectrum). To find mismatches, also peaks not assigned to any type are taken into account, provided that all of them can be potentially assigned to the considered type of peak and among them there is at least one, which can be potentially assigned only to that type. In such a case all potential pairs of peaks are compared and if none of them matches, the mismatch is established. At this stage the initial tolerances (inverse maximum evolution times) are used. CSSSs pairs which have a mismatch cannot be linked later on, even if some peak positions are consistent.

Note that for each CSSS two types of both mismatches and links exist: forward ones (to a successor) and backward ones (to a predecessor). The two types of links/mismatches can be established using different peak types. For example, for a given CA-CO cross-section of an HNCACO spectrum, the $C\alpha_{i-1}$–CO_{i-1} peak is used to find backward links and $C\alpha_i$–CO_i to find forward links. Technically, the program builds only forward links, the backward links are formed artificially based on already known forward links.

For the establishing of links between CSSSs, statistics of distance differences between corresponding peaks within each type of spectrum are needed. This relies on pairs of cross-sections with unambiguous links (only one cross-section contains a peak whose position is within the initial tolerances from a given peak of the given cross-section). Having formed all such reliable pairs, the standard deviation of distance differences is calculated for both dimensions of each spectrum. These values can then be used to cope with some ambiguous cases, as described below.

The next step is to establish the links. Each peak of the CSSS (below called main peak) is compared to the peaks from the same spectrum but different CSSS (called test peaks). If only one test peak can be found within the initial tolerances from the main peak, the link between the two CSSSs is established immediately. However, if several test peaks fall into this range, a stricter criterion employing the above calculated standard deviations of distance differences is needed. The procedure is then similar to that employed during assignment of peaks to peak types. Firstly, if the distance between the main peak and the closest test peak is in any dimension smaller than three standard deviations, all test peaks whose distance in this dimension from the main peak exceeds four standard deviations are excluded. Note that in this procedure we do not compare peaks from various spectra, but only from various cross-sections of the same spectrum. This allows us to avoid the problem of spectra calibration. The problem may appear in the very infrequent situation when a peak is artificially added (in the *CSSS-Defined2D* function); then its coordinates originate from other spectra and the match may be not perfect. If after applying the above criterion there are still more than one potential links from a single CSSS, then for all test peaks i, except for the closest one, the following condition is checked in both dimensions:

$$\begin{aligned}&(cdf(|pos_{closest} - pos_{main}|) - 0.5) \cdot 3 \\ &\quad < (cdf(|pos_i - pos_{main}|) - 0.5)\end{aligned} \tag{3}$$

where pos_{main} is the position of the main peak, $pos_{closest}$ is the position of closest test peak, pos_i is the position of the peak i (all positions in the considered dimension), cdf is the cumulative distribution function for the corresponding normal distribution with mean equal 0. If the above condition is fulfilled the test peak i is excluded.

The above algorithm allows establishing several links from one CSSS. Therefore, building up the chains of CSSSs is not straightforward. The scheme of this procedure is presented in Fig. 3. If CSSS i forms only one forward link to CSSS j and CSSS j forms only one backward link to CSSS i then the $i - j$ pair becomes a fragment of a chain (Fig. 3a). The situation is more complex, when the CSSS under consideration forms more links. If CSSS i forms only one forward link to CSSS j, but CSSS j forms several backward links, then the $i - j$ fragment is formed provided that that all other CSSS linked to j form alternative forward links (Fig. 3b). If CSSS i forms several forward links, then the $i - j$ fragment is formed only if from those CSSSs that have backward links to i only CSSS j has a single backward link (Fig. 3c). When all possible fragments of chains are formed, they are joined together e.g. the fragment ending with CSSS k is joined with the fragment starting with CSSS k.

This chain-forming procedure is not 100 % reliable, i.e. chains with false links occur. Such cases are usually detected in the next stage of mapping the chains onto the protein sequence, where they also can be corrected.

Mapping the chains of CSSSs onto the protein sequence

The mapping is performed according to the chains' lengths, starting from the longest chain (see also Fig. S1 in Supplementary Material). Longer chains have a better chance for a unique sequence of possible chemical shifts, which reduces the number of possible mapping sites and facilitates the starting of the mapping procedure. For each chain all possible sites in the sequence are found. Amino acid types compatible with each CSSS have to match those present in the sequence and there can be no conflicts with the chains already positioned. The latter condition means not only that

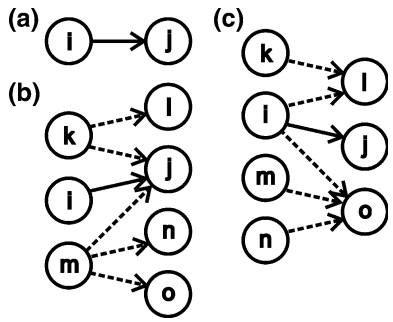

Fig. 3 Different situations occurring during the formation of CSSSs chains. *Circles with letters* represent individual CSSSs, and *arrows* are links for chain formation. *Solid arrows* link the CSSSs that constitute a fragment of a formed chain, *dashed arrows* represent other links (see text). In all cases, a $i - j$ chain fragment is formed. **a** CSSS i forms exactly one forward link and CSSS j forms exactly one backward link. **b** CSSS j forms several backward links, but out of its potential predecessors just one CSSS (i) forms exactly one forward link. **c** CSSS i forms several forward links, but out of its potential successors just one CSSS (j) forms exactly one backward link

chains cannot overlap, but also that if they are positioned one directly after the other, no mismatch between the last CSSS of the first chain and the first CSSS of the second chain is allowed.

If several sites in the sequence fulfill the above criteria, the best one can be chosen using BMRB statistics. For all the sites the average deviation from the BMRB mean (in units of BMRB standard deviations) for all residues of the chain is calculated. Initially, it is calculated only for Cβ chemical shifts. If the smallest result is at least three times smaller than the second smallest (or five times in the case of chains consisting of only one CSSS), the site is accepted. In the opposite case the deviations for Hβ are added to Cβ. If this still does not identify a unique solution, Cα deviations are calculated, but these values are not used for chains of length one. If also this last criterion fails, the chain remains unmapped at this step; it may be mapped when positioning of another chain limits the number of potential sites for the first one.

If there is already only one site, where the chain can be positioned, it is checked whether there will not be any conflict with the positioning of chains of similar length (at least 70 % of the length of the currently considered chain). If there is such a conflict, the chain remains unmapped until placing of another chain allows deciding which of the two conflicting chains should be mapped in this region. However, if a chain A has only one conflict—with a chain B—and the chain B has already at least one other conflict, then chain A is positioned. If a chain with one potential site is not in any conflict, it is positioned and cannot be moved any more.

If a chain cannot be positioned at any site, an error in the chain is assumed. It is split into shorter chains in an attempt to find a sub-chain that can be positioned. This sub-chain should be as long as possible, but not shorter than the next unmapped chain. Positioning of the sub-chain follows the method described above. If the procedure succeeds, the remaining chains from the termini of the original chain (one or two chains) are put into the array of unassigned chains according to their length. If the attempt to position the sub-chain fails, the full chain is built up again, assuming that the split was also incorrect; attempts of positioning it, possibly including a new splitting, may then be repeated at a later stage.

After positioning of any chain, the whole procedure is repeated starting from the longest unmapped chain.

Output

Four text files are produced as output of the program: 'peaks.txt', 'links.txt', 'chains.txt' and 'resonance_list.txt', containing the assignment result and all the information allowing evaluation of the program performance. The 'peaks.txt' file contains a list of peaks of unknown peak type, and a complete list of all peaks with assigned peak types (Table S3 in Supplementary Material). The 'links.txt' file contains a list of forward links for each CSSS and a list of all mismatches between CSSSs (Table S4 in Supplementary Material). In the 'chains.txt' file the protein sequence with assigned CSSSs is printed; in addition all assigned and unassigned chains, with possible assignments, are listed (Table S5 in Supplementary Material). The 'resonance_list.txt' file (not shown) contains all chemical shifts obtained by the program and is the main output of TSAR. In the presentation of the results, references are made to the CSSSs numbers (corresponding to the initial order of the peaks in the basis spectrum). This makes it easy to find spectral data of doubtful fragments of assignment or unassigned CSSSs, allowing checking and complementing the assignment manually.

Implementation

The TSAR program was written in Python, version 2.5. For all examples presented in the "Results" sections, execution time did not exceed 10 s on a single CPU. For academic users the program is available free of charge from the authors upon request.

Experimental

To test the algorithm, data from four proteins were used: the 5–79 fragment of bovine Ca^{2+}-loaded calbindin from the d9k P47 M mutant (below called calbindin), the protein interacting with NIMA-kinase from *Cenarcheaum symbiosum* (below called CsPin), the cupredoxin azurin from *Pseudomonas aeruginosa*, and the δ subunit of RNA polymerase from *Bacillus subtilis* (below called delta). All the

samples were uniformly ^{13}C, ^{15}N-labeled. The concentrations and pH were as follows: 1.0 mM and pH = 6.0 for calbindin, 1.5 mM and pH = 7.5 for CsPin, 1.0 mM and pH = 5 for azurin and 0.7 mM and pH = 6.6 for delta.

For calbindin (Kazimierczuk et al. 2010b), CsPin (Zawadzka-Kazimierczuk et al. 2010) and delta (Motáčková et al. 2010; Zawadzka-Kazimierczuk et al. 2012) the data have been acquired previously, on a Varian NMR System 700 spectrometer, equipped with a Performa XYZ PFG unit, using the standard 5 mm ^{1}H, ^{13}C, ^{15}N triple resonance probe. The azurin sample was measured on a Varian NMR INOVA 900 spectrometer, equipped with a cold probe. The calbindin, delta and azurin samples were measured at 298 K and the CsPin sample was measured at 289 K. All acquisition parameters are summarized in Table 1.

Results

The TSAR program was tested on several data sets recorded for both structured and unstructured proteins and therefore featuring various level of chemical shift degeneracy. The sizes of the proteins varied from 8.5 to 20 kDa. In each case the basis spectrum was a 3D HNCO spectrum, and various 4D and 5D techniques were chosen for the higher-dimensional experiments as listed in Table 2 for each protein, together with experimental times. Completeness of each data set is shown in Table 3 as the percentage of cross-sections with all expected peaks present, with one peak missing, with two peaks missing etc. Table 4 summarizes the performance of the program for each protein and various sets of input spectra by providing percentages of assigned CSSSs and resonances. The latter is with respect to the theoretically possible number of the given set of techniques (i.e. only nuclei types with chemical shifts observed in any of the experiments used are counted). Also reported in Table 4 is the extent of assignment for the chains of CSSSs formed by the program; these were divided into three groups according to their length (*long*: ≥8 CSSSs, *medium*: 3–7 CSSSs, *short*: ≤2 CSSSs). Table 4 also shows the numbers of assigned CSSSs from long, medium and short chains. Note that a larger number of long chains may contain fewer CSSSs than a smaller number of (very) long chains, like for the experiment sets 'A B C D F' and 'A B D F' for azurin. The chain lengths provide valuable information about the assignment quality: the longer a resulting chain, the more reliable is the result (but note that Pro will always end a chain; thus, due to Pro36 and Pro40 in azurin, there should always be a chain of maximal length three). In Table 5, the fractions of assigned resonances with respect to various types of nuclei, i.e. H^N, N, CO, Cα, Cβ, Hα and Hβ, are presented. Before presenting results for the individual proteins, it is important to state that for all proteins all assignments are correct.

Calbindin (Forsen et al. 1990; Skelton et al. 1995), the smallest of the proteins tested, consists of 75 residues with little problem with respect to signal dispersion. Three

Table 1 Experimental parameters for the indirect-dimensions for all spectra

Sample	Exp[a]	Ni[b]	Dimension 1			Dimension 2			Dimension 3			Dimension 4		
			Nucl	t^c_{max}	sw^d	Nucl	t^c_{max}	sw^d	Nucl	t^c_{max}	sw^d	Nucl	t^c_{max}	sw^d
Calbindin	A	300	CO	30	2.8	N	50	2.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Calbindin	E	725	HAB	6.5	4.0	CAB	7.1	14.0	CO	28	3.0	N	28	2.5
Calbindin	F	675	HN	5.5	6.0	N	27.5	2.5	CO	8.9	3.0	N	27.5	2.5
CsPin	A	2,000	CO	40	3.8	N	50	2.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CsPin	C	1,800	CO	20	3	CA	10	6.2	N	30	2.5	n.a.	n.a.	n.a.
CsPin	D	1,800	CA	7.1	14.0	CAB	10	14.0	N	30	2.5	n.a.	n.a.	n.a.
Azurin	A	300	CO	20	4.5	N	30	3.6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Azurin	B	810	CO	8	4.5	CA	10	8.0	N	28	3.6	n.a.	n.a.	n.a.
Azurin	C	1,000	CO	20	4.5	CA	8	8.0	N	28	3.6	n.a.	n.a.	n.a.
Azurin	D	1,100	CA	7	20.6	CAB	10	20.6	N	28	3.6	n.a.	n.a.	n.a.
Azurin	F	1,000	HN	7	8.0	N	28	3.6	CO	20	4.5	N	28	3.6
Delta	A	484	CO	100	3.7	N	120	2.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Delta	E	880	HAB	10	4.0	CAB	7.1	14.0	CO	30	3.0	N	50	2.8
Delta	F	740	HN	20	6.0	N	50	2.8	CO	30	3.0	N	50	2.8
Delta	G	1,570	N	50	2.8	CO	45	3.0	CO	45	3.0	N	75	2.8

[a] See Table 2 for experiment symbols

[b] Number of increments in indirectly detected dimensions (together)

[c] Maximum evolution time in indirectly-detected dimensions, in milliseconds

[d] Spectral widths in indirectly-detected dimensions, in kilohertz

Table 2 Types of experiments run for the tested samples and corresponding experimental times (in hours)

Symbol of experiment	Name of experiment	Dimensionality	Calbindin	CsPin	Azurin	Delta
A (basis spectrum)	HNCO	3D	1.9	12.9	1.9	3.1
B	HNCOCA	4D	–	–	10.4	–
C	HNCACO	4D	–	23.2	12.9	–
D	HNCACACB	4D	–	23.2	14.2	–
E	HabCabCONH	5D	18.7	–	–	22.7
F	HN(CA)CONH	5D	17.4	–	26.7	19.1
G	(H)NCO(NCA)CONH	5D	–	–	–	25.8

Table 3 Data completeness given as percentages of the cross-sections

Experiment	Peaks[a]	Calbindin	CsPin	Azurin	C delta
HNCOCA	1	n.a.	n.a.	100/0	n.a.
HNCACO	2	n.a.	90.8/8.1/1.1	85.4/13.0/1.6	n.a.
HNCACACB	4	n.a.	91.9/5.8/2.3/0/0	65.0/16.3/17.9/0.8/0	n.a.
HabCabCONH	2	85.9/14.1/0	n.a.	n.a.	96.3/3.7/0
HN(CA)CONH	2	98.6/1.4/0	n.a.	65.1/26.0/8.9	90/8.7/1.3
(H)NCO(NCA)CONH	2	n.a.	n.a.	n.a.	87.5/8.7/3.8

[a] Number of peaks expected on each cross-section of the given spectrum

The first number refers to complete cross-sections, the following to cross-sections with one missing peak, with two missing peaks etc. The types of peaks are not differentiated

spectra were acquired for this sample: 3D HNCO (basis spectrum), 5D HN(CA)CONH and 5D HabCabCONH. Two data sets were constructed: utilizing only the HN(CA)CONH (as only this spectrum provides sequential links) and utilizing both 5D spectra. In both cases 100 % of CSSSs were correctly assigned and the lengths of the formed chains (38, 16, 16, 1) were maximal, spanning complete fragments between the prolines. Unambiguous positioning of the chains was achieved by recognition of CSSSs corresponding to the five glycine residues. When using also the HabCabCONH spectrum, the additional information on Cα, Cβ, Hα and Hβ chemical shifts enabled recognition of more amino-acid residue types. In the case of using a single 5D spectrum, 100 % of resonances were obtained, while using both 5D spectra reduced this number to 87.7 %. This apparent inconsistency (better result from less data) can be explained by noting that this percentage is calculated relatively to all theoretically available resonances: only H^N, N and CO in the former case, Cα, Cβ, Hα and Hβ added in the latter. It is noteworthy that when using the HabCabCONH spectrum, the only methods applicable to the assignment of peaks to particular peak types were *OutOfRangeCACBHA* and *DifferentiateCA-CB* (see "Methods"). The fact that no other experiment providing any of the Cα, Cβ, Hα or Hβ chemical shift was available to help in recognition of peaks types explains the relatively low percentage of assigned resonances.

CsPin (Jaremko et al. 2011) is a 97-residue protein, but six initial residues of the N-terminus were not labeled, thus the assignable fragment included 91 residues. Three spectra were acquired for this sample: 3D HNCO (basis spectrum), 4D HNCACO and 4D HNCACACB. All of them were used by the program to achieve the assignment. The fraction of assigned resonances was 97.7 % and the fraction of assigned CSSSs was 94.4 %. The lengths of the chains were rather long (six chains of at least 8 CSSSs), but did in all but one cases not correspond to the possible maximum. Three chains of length 1 remained unassigned. The inability of the program to complete the assignment was caused by the fact that on these planes none of the peaks could be automatically assigned to a peak type (due to CO degeneracy in two consecutive residues or to missing peaks). Consequently, the links could not be formed. Nevertheless manual fitting of the chains onto the protein sequence was still possible (based on visual analysis of the cross-sections instead of peak picking, and on testing of all alternatives).

Azurin (Parr et al. 1976) with its 128 amino-acid residues (14 kDa) is the largest of the fully folded proteins tested here. The data sets feature low completeness (see Table 3). Several reasons are probably contributing to this: higher relaxation rates cause lower sensitivity, this is further reduced by the N–Cα coupling, and finally three of the azurin spectra were recorded with the shortest experiment times of Table 2 (except for the basis spectra). While sensitivity is really a spectroscopic rather than a TSAR problem, the low completeness represents an interesting challenge for TSAR. Altogether five spectra were acquired:

Table 4 TSAR assignment results

Protein	Set of experiments[a]	Total experimental time (h)	Assignments (%)		Number of assigned/unassigned chains			Number of assigned CSSSs in Long/medium/short chains
			CSSSs	Resonances	Long (≥8)	Medium (3–7)	Short (≤2)	
Calbindin	A E F	38	100	87.7	3/0	0/0	1/0	70/0/4
	A F	19.3	100	100	3/0	0/0	1/0	70/0/4
CsPin	A C D	59.3	94.4	97.7	6/0	1/0	2/3	76/5/3
Azurin	A B C D F	65.2	100	98.9	6/0	0/0	3/0	118/0/5
	A C D F	54.8	96.7	96.0	7/0	3/0	1/4	108/10/1
	A B D F	52.3	100	98.7	8/0	4/0	6/0	95/21/7
	A B C F	51	100	100	6/0	0/0	3/0	119/0/4
	A B C D	39.5	100	98.9	5/0	4/0	4/0	95/21/7
Delta	A E F G	70.7	100	96.8	3/0	6/0	7/0	40/29/11

[a] See Table 2 for experiment symbols

Table 5 TSAR assignment results with specified types of nuclei

Protein	Data set[a]	Assigned resonances (%)							
		HN	N	CO	CA	CB	HA	HB	total
Calbindin	A E F	100	100	100	71.8	84.8	71.8	84.8	87.7
	A F	100	100	100	n.a.	n.a.	n.a.	n.a.	100
CsPin	A C D	94.4	94.4	100	100	100	n.a.	n.a.	97.7
Azurin	A B C D F	100	100	100	100	94.0	n.a.	n.a.	98.9
	A C D F	96.7	96.7	97.7	97.7	90.6	n.a.	n.a.	96.0
	A B D F	100	100	100	99.2	94.0	n.a.	n.a.	98.7
	A B C F	100	100	100	100	n.a.	n.a.	n.a.	100
	A B C D	100	100	100	100	94.0	n.a.	n.a.	98.9
Delta	A E F G	100	100	100	96.3	92.5	96.3	92.5	96.8

[a] See Table 2 for experiment symbols

3D HNCO (basis spectrum), 4D HNCOCA, 4D HNCACO, 4D HNCACACB and 5D HN(CA)CONH, out of which five data sets were constructed: a first one utilizing all high-dimensional spectra and four others utilizing different combinations of three out of four available high-dimensional spectra. The percentage of assigned CSSSs was 96.7 % in the case of combining the HNCACO, HNCACACB and HN(CA)CONH spectra, while in the remaining four cases it was 100 %. In spite of the low completeness of the given in Table 3, the percentage of assigned resonances was 96 % or more. The level of difficulty of these data sets is reflected in the lengths of the formed chains. In one case seven assigned chains were *short*. Although in this case all the assignments were correct, with this number of short chains manual checking is strongly recommended. Interestingly, one of the data sets utilizing three 4D spectra (HNCOCA, HNCACO and HNCACACB; last line for azurin in Tables 4, 5) provides identical results to the full set, where the additional 5D HN(CA)CONH alone requires 26.7 h compared to 39.5 h for all three 4D spectra together. The only difference in the result lies in the lengths of formed chains—in the latter case they are slightly longer, which makes the assignment more reliable.

Delta (Lopez de Saro et al. 1995) is a 20 kDa protein (172 amino-acid residues), but the automatic assignment was performed only for its unstructured C-terminal part with 81 residues. This case was especially challenging due to various repetitive sequences in the unstructured part (glutamates: one stretch of four residues, one stretch of three residues, four stretches of two residues; aspartates: one stretch of four residues, four stretches of two residues; lysines: one stretch of four residues, one stretch of two residues) and small diversity of the types of residues involved (25 aspartates, 23 glutamates, 8 lysines and 8 leucines in the 81 residue long unstructured chain). Moreover, the unstructured part of delta contains a very limited number of residues that are easily recognizable by their shifts: no glycine, no serine, one threonine, four

alanines. As a result, the chemical shift ranges are outstandingly narrow, even comparing to other intrinsically disordered proteins (Motáčková et al. 2010), e.g. for the amide hydrogens it is 8.03–8.48 ppm and for the amide nitrogens: 120.2–126.3 ppm (excluding two outliers). Therefore the set of experiments employed to automatically assign the resonances included, besides 3D HNCO (basis spectrum), three 5D experiments: HN(CA)CONH, (H)NCO(NCA)CONH and HabCabCONH. The first and second spectra provide sequential links, while the last one yields Cβ, Hβ, Cα, Hα resonances used for positioning of the resulting chains. This set allowed to obtain a complete assignment: 100 % of assigned CSSSs and 96.8 % of assigned resonances. The chains were relatively short (three *long*, six *medium*, seven *short*); nevertheless all of them were correctly assigned (again, with this number of short chains manual checking is recommended). Such a good result may be surprising considering the presence of many short chains, the low residue type diversity and the small number of easily recognizable residues. It can be explained by the use of HabCabCONH data. This technique provides Hα, Hβ, Cα and Cβ resonances, allowing to exploit all the methods of chain positioning. This was neither the case for CsPin nor for azurin, where only Cα and Cβ were available.

Discussion

Using the TSAR program, resonance assignment of proteins can be based on various SMFT data sets. The choice of the high-dimensional experiments should depend on the protein size and chemical shift dispersion. Firstly, the dimensionality of experiments should not be too small (see below for a discussion of overlapping cross-sections). The number of experiments providing sequential links should depend on the degree of chemical shift dispersion in the cross-sections' dimensions and on data completeness (connected with the sample concentration and the experiment sensitivity). For example, in the case of calbindin a single spectrum providing connectivities was sufficient to form chains of maximal lengths, which could then easily be positioned in the protein sequence. However, usually more spectra providing links are required. The problem is especially difficult in the case of unfolded proteins, where the chemical shifts of aliphatic carbons and protons exhibit particularly low dispersion. In this situation, spectra providing sequential links via these types of nuclei (e.g. 4D HNCACACB, 4D HabCabNH) are not sufficient and links via better-separated dimensions (H^N, N, CO) should be provided. For instance, for the delta protein two 5D spectra providing the connectivities were necessary: HN(CA)CONH, with links via H^N and N, and (H)NCO(NCA)CONH, with links via N and CO. Besides link-providing spectra also spectra enabling positioning of CSSSs chains are required. If the chains are long and there are several glycines in the protein sequence, simple recognition of these glycines may allow for unambiguous mapping. While this was the case for calbindin, where the entire assignment could be achieved based only on the 5D HN(CA)CONH, this is usually not sufficient and the mapping step requires Cβ and possibly also Hβ chemical shifts. In the case of azurin and CsPin, the 4D HNCACACB spectrum was used to obtain Cβ chemical shifts (at the same time it also provided sequential connectivities). The delta protein required a 5D spectrum (HabCabCONH) for this purpose.

A condition for using TSAR is that data resulting from all experiments can be SMFT-processed using a single basis spectrum. This means that each of the high-dimensional experiments must contain some dimensions common with the basis spectrum (a ND experiment requires at least N-2 common dimensions). The basis spectrum should contain one peak per residue. Any residue not represented in the basis spectrum (missing peak due to insufficient sensitivity of the basis spectrum, but also normal absence of peaks, e.g. for prolines if H^N is involved) prevents the calculation of the corresponding cross-sections of the high-dimensional spectra. This hinders the formation of sequential connectivities at the corresponding site of the protein sequence and consequently breaks the chain of CSSSs. Thus, the sensitivity of the basis spectrum is of high importance, and the completeness of the basis spectrum peak list should be checked such that at least the expected number of peaks is found. If false peaks appear in the basis-spectrum, usually no peak appears in the corresponding cross-sections of the high-dimensional spectra. Thus, it has no negative impact on the final result, and it is recommended to pick rather too many basis-spectrum peaks than too few. We used Sparky (Goddard and Kneller 2002) for automated peak picking of the basis-spectra (but any other peak-picking may be used here). Besides good sensitivity, the basis spectrum should also be chosen and recorded with the goal of high resolution. This ensures that the cross-sections can be properly selected, avoiding any decrease of peak intensities due to inaccurate shifts of the dimensions corresponding to the basis spectrum, which define the position of the cross-sections.

Cross-sections are calculated using the SMFT procedure by setting frequencies of some of the dimensions to the values obtained from the basis spectrum peak list (fixed dimensions). If the basis spectrum frequencies in the fixed dimensions are well separated, only peaks from one spin system will appear on each cross-section. However, it can happen that some sets of the fixed frequencies are sufficiently close to produce cross-sections on which peaks originating from several spin systems appear, i.e. some peaks show up on more than one cross-section. The current TSAR program

cannot cope with this situation. The input peak lists should contain for each cross-section only the peaks originating from the corresponding spin system. However, the task of preparing such lists is usually rather straightforward. A peak usually "belongs" to that cross-section on which it has the highest absolute intensity (note, that this is not equivalent to choosing the most intensive peaks for each cross-section). Normally, if the dimensionality of the experiments was well adjusted to the degree of resonance degeneracy of the sample, there should be only a few pairs of such overlapping cross-sections and the problem can be solved manually or by using a script detecting overlaps. More problematic are situations with triplets of overlapping cross-sections. This typically means that the dimensionality of the conducted experiments was not sufficient. Therefore, the dimensionality should be adjusted to the given sample, meaning that the peaks in the spectrum consisting of the fixed dimensions in the higher-dimensional spectrum should (almost) not exhibit any overlaps. For instance, if cross-sections of a 4D spectrum are to be calculated based on amide proton and nitrogen frequencies, the peaks in ^{15}N-HSQC should be well separated. So far we found that 5D experiments were sufficient even for very difficult cases (e.g. the delta protein with several repetitive sequences in a disordered part).

False peaks which appear in high-dimensional spectra can be detected at the stage of assignment of peak types: sometimes (depending on the set of experiments used and on the spectrum in which the false peak appears) such peaks do not fulfill the conditions of any peak type and thus are disregarded. However, some methods of peak-type recognition (i.e. *OnlyPeak&OnlyTypeLeft* and *DifferentiateCA-CB*) are not able to identify false peaks. If a false peak is already assigned to a certain peak type, wrong matches and/or mismatches can be established, which can lead to the lack of assignment of a certain CSSSs chain or to an incorrect assignment. To prevent such situations, the peak-type recognition methods, which do not allow identifying false peaks, are applied only if no other methods could recognize the peaks types for a certain CSSS. In general, it is better to lose some weak peaks in higher-dimensional spectra than to include false ones.

Problems with the present algorithm may arise due to unusual Cα, Cβ, Hα or Hβ chemical shifts (i.e. shifts outside of four standard deviations from the statistical average). This can cause an error at two stages of the program operation: assignment of peak to peak type (*OutOfRangeCACBHA* method) and statistics-based amino acid recognition. In the former case, the peaks of the given CSSS may remain unassigned to any peak type, which can prevent link formation in the following stage, or they may be assigned incorrectly, which can cause formation of an incorrect link. Amino-acid recognition problems can be caused not only by untypical chemical shifts, but also by incorrect peak-type recognition. Most problematic are confusions of Hα–Cα peaks with Hβ–Cβ peaks; this may cause the chemical shift statistics to propose a wrong amino acid type. Furthermore, recognition of a pair of prochiral methylene protons in the case of wrongly-recognized peak types will be incorrectly interpreted. That is why the *DifferentiateCA-CB* method is not a very safe criterion and only used when no other, more reliable methods for Cα and Cβ peaks are available. This method was used in two of the presented data sets (calbindin 'A E F' data set and delta); while no incorrect peak-type assignments were made by this method, a number of peaks remained unassigned. Generally, in the case of incorrect amino-acid recognition, the most favorable situation is that all amino acid types are excluded; then, all amino acids types are included back again and the mapping may be done properly. Such a situation occurred in one of the test data sets (for calbindin with one of the Hβ chemical shifts of Lys25 being smaller than 0.5 ppm). A more severe problem occurs, when the correct type is excluded, while some other types are left. If the considered CSSS is part of a longer chain, this chain will probably not be mapped, unless it is split. If it belongs to a short chain (originally or as a result of splitting) it may be mapped incorrectly. This is one of the reasons for which the assignment of short chains is less reliable and postponed to the end.

The "best-first" strategy for the mapping of chains onto the protein sequence utilized by the TSAR program has one major disadvantage. The incorrect assignment of one chain will affect the rest of the assignments, which is especially harmful if the chain is long. Some protection against this is given by the detection of conflicts with other chains of at least 70 % of the length of the considered chain. Therefore, errors in an assignment often lower the percentages of assigned resonances and CSSSs, without necessarily yielding erroneous assignments.

The TSAR program was primarily designed for the analysis of SMFT-processed data, which means that at least 4D spectra have to be run. This represents a spectroscopic limitation, but means that TSAR will be mostly used on unfolded proteins or on folded proteins up to 20 kDa, since for larger folded proteins such high-dimensional spectra may be not achievable due to fast signal relaxation. The TSAR program can be used also for the analysis of data obtained using methods other than SMFT provided that they are organized in a way accepted by the algorithm. However, one spectrum (with arbitrary dimensionality), containing one peak per residue, has to serve as a basis spectrum. The definitions of other spectra must contain exactly two additional frequency axes, and the order of the data equivalent to the cross sections in these spectra has to correspond to the order of peaks in the basis-spectrum peak list.

Attempts to compare TSAR with other computational assignment tools involve parameter choices and typically

Table 6 GARANT assignment results with specified types of nuclei

Protein	Data set[a]	Resonances correctly/incorrectly assigned by GARANT (%)							
		HN	N	CO	CA	CB	HA	HB	Total
Calbindin	A E F	98.6/1.4	98.6/0	98.6/0	82.9/0	98.5/0	74.3/0	90.8/9.2	91.7/1.5
	A F	The assignment could not be achieved using this data set							
CsPin	A C D	98.9/0	98.9/0	100/0	100/0	100/0	n.a.	n.a.	99.5/0
Azurin	A B C D F	97.6/0	97.6/0	96.9/0	96.9/0	94.9/0	n.a.	n.a.	96.8/0
	A C D F	96.7/0	96.7/0	96.1/0	96.9/0	94.9/0	n.a.	n.a.	96.3/0
	A B D F	95.9/0	95.9/0	96.1/0	96.9/0.8	94.9/0	n.a.	n.a.	96.0/0.2
	A B C F	92.7/0	91.9/0	93.8/0	92.2/0	n.a.	n.a.	n.a.	92.6/0
	A B C D	97.6/0	97.6/0	96.9/0	96.9/0	94.9/0.9	n.a.	n.a.	96.8/0.2
Delta	A E F G	92.5/0	90/0	97.5/0	95.0/2.5	90.0/2.5	98.8/1.2	83.8/5.0	92.5/1.6

[a] See Table 2 for experiment symbols

lead to incompatibilities of input data that may be handled in various ways. More generally, relevant comparisons should be based on complete approaches (data acquisition and analysis), for example in the context of efforts such as CASD-NMR (Rosato et al. 2009). For a qualitative comparison, we run GARANT (Bartels et al. 1996, 1997), which accepts almost exactly the same input as TSAR. The GARANT results for individual nuclei types are collected in Table 6. In general, the results of the two algorithms for most of the data sets are comparable. We attribute the slightly lower extent of assignments in GARANT to the more general type of approach of the latter, making not full use of the SMFT data, but also to the possibly non-optimal choice of parameters made by us. The fact that GARANT analyzes many assignment possibilities for each nucleus and then chooses the best one, results in several suggestions from independent runs with different starting seeds (we run GARANT twelve times). The choice of how many runs need to coincide for accepting an assignment may explain the sometimes higher assignment rates of GARANT (e.g. for CsPin) but at the same time the occurrence of some incorrect assignments (e.g. for delta).

Conclusions

The TSAR program is a new algorithm for automatic resonance assignment designed to exploit all advantages of 2D cross-sections of spectra of high dimensionality ($\geq$4D) obtained by SMFT processing. Focusing on this type of input makes the approach robust in terms of assignment completeness and reliability, and inexpensive with respect to computational time. Furthermore, the user interface is adapted for work with a set of cross-sections, which makes the program easy to use. An important feature of TSAR is the automatic determination of chemical shift tolerances. This makes the approach unique and insensitive to user experience. Due to its flexibility in defining experiments, the program is prepared for emerging techniques, including e.g. ^{13}C-detected experiments (Bermel et al. 2006, 2009; Nováček et al. 2011). Contrary to many other algorithms, the TSAR program is able to cope with experiments which involve three consecutive residues. An important feature of the program is its ability to exploit information about glycines which may arise from changing peak signs in the absence of Cβ carbons.

The robustness of the TSAR program was confirmed by a series of experimental tests. It was shown, that even while working with very difficult cases, such as an unfolded protein fragment with various repetitive amino acid sequences, the result can be complete and reliable. Currently, the bottleneck of the proposed assignment strategy is input data preparation. Good quality peak lists have to be prepared for the basis-spectrum and all higher-dimensional spectra.

Acknowledgments This work has been supported by the Bio-NMR project under the 7th Framework Programme of the EC grant agreement 261863 for conducting the research. A.Z.-K. thanks the Foundation for Polish Science for supporting her with the MPD Programme that was co-financed by the EU European Regional Development Fund. The experiments were performed in the Structural Research Laboratory at the Faculty of Chemistry, University of Warsaw, Poland (calbindin, CsPin and delta), and the Swedish NMR Centre, University of Gothenburg, Sweden (azurin). We thank Jiří Nováček of the Masaryk University for providing data from ^{13}C-detected experiments for the delta protein, which were used for testing of an early version of the TSAR program. We are grateful to Göran Karlsson of the Swedish NMR Centre for the loan of labeled azurin.

References

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

Baran MC, Huang YJ, Moseley HN, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3556

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

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

Bermel W, Bertini I, Felli IC, Piccioli M, Pierattelli R (2006) 13C-detected protonless NMR spectroscopy of proteins in solution. Prog Nucl Magn Reson Spectrosc 48:25–45

Bermel W, Bertini I, Csizmok V, Felli IC, Pierattelli R, Tompa P (2009) H-start for exclusively heteronuclear NMR spectroscopy: the case of intrinsically disordered proteins. J Magn Reson 198: 275–281

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

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

Coggins BE, Venters RA, Zhou P (2010) Radial sampling for fast NMR: concepts and practices over three decades. Prog Nucl Magn Reson Spectrosc 57:381–419

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

Forsen S, Drakenberg T, Johansson C, Linse S, Thulin E, Kordel J (1990) Protein engineering and structure/function relations in bovine calbindin D9 k. Adv Exp Med Biol 269:37–42

Freeman R, Kupče E (2012) Concepts in projection-reconstruction. Top Curr Chem 316:1–20

Goddard TD, Kneller DG (2002) SPARKY 3. University of California, San Francisco

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) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson 99:201–207

Hiller S, Wider G (2012) Automated projection spectroscopy and its applications. Top Curr Chem 316:21–47

Hyberts SG, Arthanari H, Wagner G (2012) Applications of non-uniform sampling and processing. Top Curr Chem 316:125–148

Jaremko L, Jaremko M, Elfaki I, Mueller JW, Ejchart A, Bayer P, Zhukov I (2011) Structure and dynamics of the first archaeal parvulin reveal a new functionally important loop in parvulin-type prolyl isomerases. J Biol Chem 286:6554–6565

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

Kay LE, Ikura M, Tschudin R, Bax A (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Kazimierczuk K, Zawadzka A, Koźmiński W (2009) Narrow peaks and high dimensionalities: exploiting the advantages of random sampling. J Magn Reson 197:219–228

Kazimierczuk K, Stanek J, Zawadzka-Kazimierczuk A, Koźmiński W (2010a) Random sampling in multidimensional NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 57:420–434

Kazimierczuk K, Zawadzka-Kazimierczuk A, Koźmiński W (2010b) Non-uniform frequency domain for optimal exploitation of non-uniform sampling. J Magn Reson 205:286–292

Kazimierczuk K, Misiak M, Stanek J, Zawadzka-Kazimierczuk A, Kozminski W (2012) Generalized fourier transform for non-uniform sampled data. Top Curr Chem 316:79–124

Lopez de Saro FJ, Woody AY, Helmann JD (1995) Structural analysis of the *Bacillus subtilis* delta factor: a protein polyanion which displaces RNA from RNA polymerase. J Mol Biol 252:189–202

Maciejewski MW, Mobli M, Schuyler AD, Stern AS, Hoch JC (2012) Data sampling in multidimensional NMR: fundamentals and strategies. Top Curr Chem 316:49–77

Motáčková V, Nováček J, Zawadzka-Kazimierczuk A, Kazimierczuk K, Žídek L, Šanderová H, Krásný L, Koźmiński W, Sklenář V (2010) Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. J Biomol NMR 48:169–177

Nováček J, Zawadzka-Kazimierczuk A, Papoušková V, Žídek L, Šanderová H, Krásný L, Koźmiński W, Sklenář V (2011) 5D 13C-detected experiments for backbone assignment of unstructured proteins with a very low signal dispersion. J Biomol NMR 50:1–11

Orekhov VY, Jaravine VA (2011) Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. Prog Nucl Magn Reson Spectrosc 59:271–292

Parr SR, Barber D, Greenwood C (1976) A purification procedure for the soluble cytochrome oxidase and some other respiratory proteins from *Pseudomonas aeruginosa*. Biochem J 157:423–430

Rosato A, Bagaria A, Baker D, Bardiaux B, Cavalli A, Doreleijers JF, Giachetti A, Guerry P, Guntert P, Herrmann T, Huang YJ, Jonker HR, Mao B, Malliavin TE, Montelione GT, Nilges M, Raman S, van der Schot G, Vranken WF, Vuister GW, Bonvin AM (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626

Skelton NJ, Kordel J, Chazin WJ (1995) Determination of the solution structure of Apo calbindin D9 k by NMR spectroscopy. J Mol Biol 249:441–462

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

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

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

Zawadzka-Kazimierczuk A, Kazimierczuk K, Koźmiński W (2010) A set of 4D NMR experiments of enhanced resolution for easy resonance assignment in proteins. J Magn Reson 202:109–116

Zawadzka-Kazimierczuk A, Koźmiński W, Šanderová H, Krásný L (2012) High dimensional and high resolution pulse sequences for backbone resonance assignment of intrinsically disordered proteins. J Biomol NMR 52:329–337

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