Journal of Biomolecular NMR, **17:** 125–136, 2000.
KLUWER/ESCOM

A tracked approach for automated NMR assignments in proteins (TATAPRO)

H.S. Atreya, S.C. Sahu, K.V.R. Chary* & Girjesh Govil
Department of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400005, India

Received 25 January 2000; Accepted 18 April 2000

Key words: automated NMR assignments, *Borrelia burgdoferi* OspA, drosophila numb phosphotyrosine-binding domain, *Eh*-CaBP, *Escherichia coli* maltose binding protein, fibroblast collagenase, sequence specific resonance assignments, triple resonance experiments

Abstract

A novel automated approach for the sequence specific NMR assignments of $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'/^{1}H^{\alpha}$ and ^{15}N spins in proteins, using triple resonance experimental data, is presented. The algorithm, TATAPRO (**T**racked **A**u**T**omated **A**ssignments in **Pro**teins) utilizes the protein primary sequence and peak lists from a set of triple resonance spectra which correlate $^{1}H^{N}$ and ^{15}N chemical shifts with those of $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$. The information derived from such correlations is used to create a '*master_list*' consisting of all possible sets of $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$, $^{13}C'_{i}/^{1}H^{\alpha}_{i}$, $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$ and $^{13}C'_{i-1}/\,^{1}H^{\alpha}_{i-1}$ chemical shifts. On the basis of an extensive statistical analysis of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift data of proteins derived from the BioMagResBank (BMRB), it is shown that the 20 amino acid residues can be grouped into eight distinct categories, each of which is assigned a unique two-digit code. Such a code is used to tag individual sets of chemical shifts in the master_list and also to translate the protein primary sequence into an array called pps_array. The program then uses the master_list to search for neighbouring partners of a given amino acid residue along the polypeptide chain and sequentially assigns a maximum possible stretch of residues on either side. While doing so, each assigned residue is tracked in an array called assig_array, with the two-digit code assigned earlier. The assig_array is then mapped onto the pps_array for sequence specific resonance assignment. The program has been tested using experimental data on a calcium binding protein from *Entamoeba histolytica* (*Eh*-CaBP, 15 kDa) having substantial internal sequence homology and using published data on four other proteins in the molecular weight range of 18–42 kDa. In all the cases, nearly complete sequence specific resonance assignments (> 95%) are obtained. Furthermore, the reliability of the program has been tested by deleting sets of chemical shifts randomly from the master_list created for the test proteins.

Introduction

Sequence specific resonance assignments (hereafter abbreviated as ssr_assignments) in proteins are an important and essential step towards complete three dimensional (3D) structural characterization (Wüthrich et al., 1986). In recent years, several double and triple resonance experiments have been proposed to carry out ssr_assignments in isotope labeled proteins (Bax and Grzesiek, 1993). However, for large proteins, manual assignment becomes a tedious and time consuming task. This has led to an increasing demand of the development of algorithms for automation of ssr_assignments, following which a number of strategies have been proposed (see review by Moseley and Montelione, 1999). These include approaches which utilize information from various triple resonance experiments and methods such as simulated annealing

*To whom correspondence should be addressed. E-mail: chary@tifr.res.in

(Buchler et al., 1997; Lukin et al., 1997), bayesian statistics and artificial intelligence (Zimmerman et al., 1997; Montelione et al., 1999), characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of individual amino acid residues (Grzesiek and Bax, 1993; Friedrichs et al., 1994), threshold accepting algorithm (Leutner et al., 1998), connectivity tracing algorithms (Olson and Markley, 1994) and neural networks (Choy et al., 1993; Hare and Prestegard, 1994). For side chain assignments, methods have been proposed which utilize side chain topologies of spin systems (Li and Sanctuary, 1997), side chain ^{13}C chemical shift patterns of amino acid residues (Zimmerman et al., 1994) and semi-automated approaches (Meadows et al., 1994). Other strategies utilize information from homologous proteins and chemical shift prediction for complete ssr_assignments (Bartels et al., 1996; Gronwald et al., 1999).

In this paper, we propose a novel algorithm for automated ssr_assignments of $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'/^{1}H^{\alpha}$ and ^{15}N spins in proteins, called TATAPRO, using the protein primary sequence and a set of triple resonance experiments which correlate the $^{1}H^{N}$ and ^{15}N chemical shifts with those of $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'/^{1}H^{\alpha}$. This approach is demonstrated using data from CBCANH (Wittekind and Mueller, 1993), CBCA(CO)NH (Grzesiek and Bax, 1992), HN(CA)CO (Clubb et al., 1992a) and HNCO (Kay et al., 1990) spectra.

Two important parameters which determine the success of any automated approach for ssr_assignments in proteins are good input of peaks, both in terms of resolution and sensitivity, and a reliable classification of individual spin systems (Moseley and Montelione, 1999). With the development of a TROSY-based approach for the implementation of various triple resonance pulse sequences (Salzmann et al., 1998, 1999) and also with the modification of pulse sequences for deuterated proteins (Gardner and Kay, 1998), it is now possible to acquire triple resonance spectra with high resolution and sensitivity for proteins with molecular weights up to 50 kDa (Loria et al., 1999). These techniques, combined with an efficient peak picking algorithm, help in overcoming ambiguities in resonance assignments to some extent. However, a satisfactory classification of individual spin systems is still an important and difficult task. Few of the algorithms in the past have utilized the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of individual amino acid residues (Grzesiek and Bax, 1993; Friedrichs et al., 1994; Lukin et al., 1997) for such a classification. However, due to extensive overlap of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts for most of the amino acid residues, identifying each residue by its characteristic chemical shifts can result in ambiguous assignments. The problem is further aggravated by unusual chemical shifts which can lead to erroneous assignments. Hence, in order to obtain insight into the distribution of NMR chemical shifts for amino acid residues, we have carried out an extensive statistical analysis using $^{13}C^{\alpha}$ (∼25 000) and $^{13}C^{\beta}$ (∼21 000) chemical shift information of all the proteins deposited in the BMRB (Seavey et al., 1991). This analysis aided in grouping the 20 amino acid residues into eight distinct categories, based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts. These categories are then distinguished from each other by assigning them a unique two-digit code. This grouping of amino acid residues dramatically reduces the problem of overlapping and unusual chemical shifts and results in a deterministic approach to the problem of ssr_assignments.

The algorithm has been tested for assignments, using experimental data, on a calcium binding protein from *Entamoeba histolytica* (*Eh*-CaBP, M_r ∼15 kDa), which possesses a substantial internal sequence homology, and on four other proteins with published assignments. The complete ssr_assignments have been accomplished in three stages, using a separate program at each stage. These programs have been written in ANSI C code and can be compiled on any Unix-based workstation or Windows-based system equipped with a C compiler. The execution time of the program is of the order of a few seconds on an R10000-based solid impact workstation (SGI). The program can be obtained on request at the following e-mail address: chary@tifr.res.in

Methodology

$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift statistics

The algorithm for ssr_assignments proposed here primarily makes use of the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of each individual amino acid residue except Pro residues. For this purpose, an extensive statistical analysis has been carried out, utilizing the chemical shift data available in the BMRB.

The histograms in Figures 1a and 1b depict the percentage of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, respectively, spanning the range 12–78 ppm (28–78 ppm for $^{13}C^{\alpha}$) for individual amino acid residues. As evident from Figure 1a, Gly($^{13}C^{\alpha}$) always resonates upfield of 50 ppm in a region well separated from the $^{13}C^{\alpha}$ chemi-

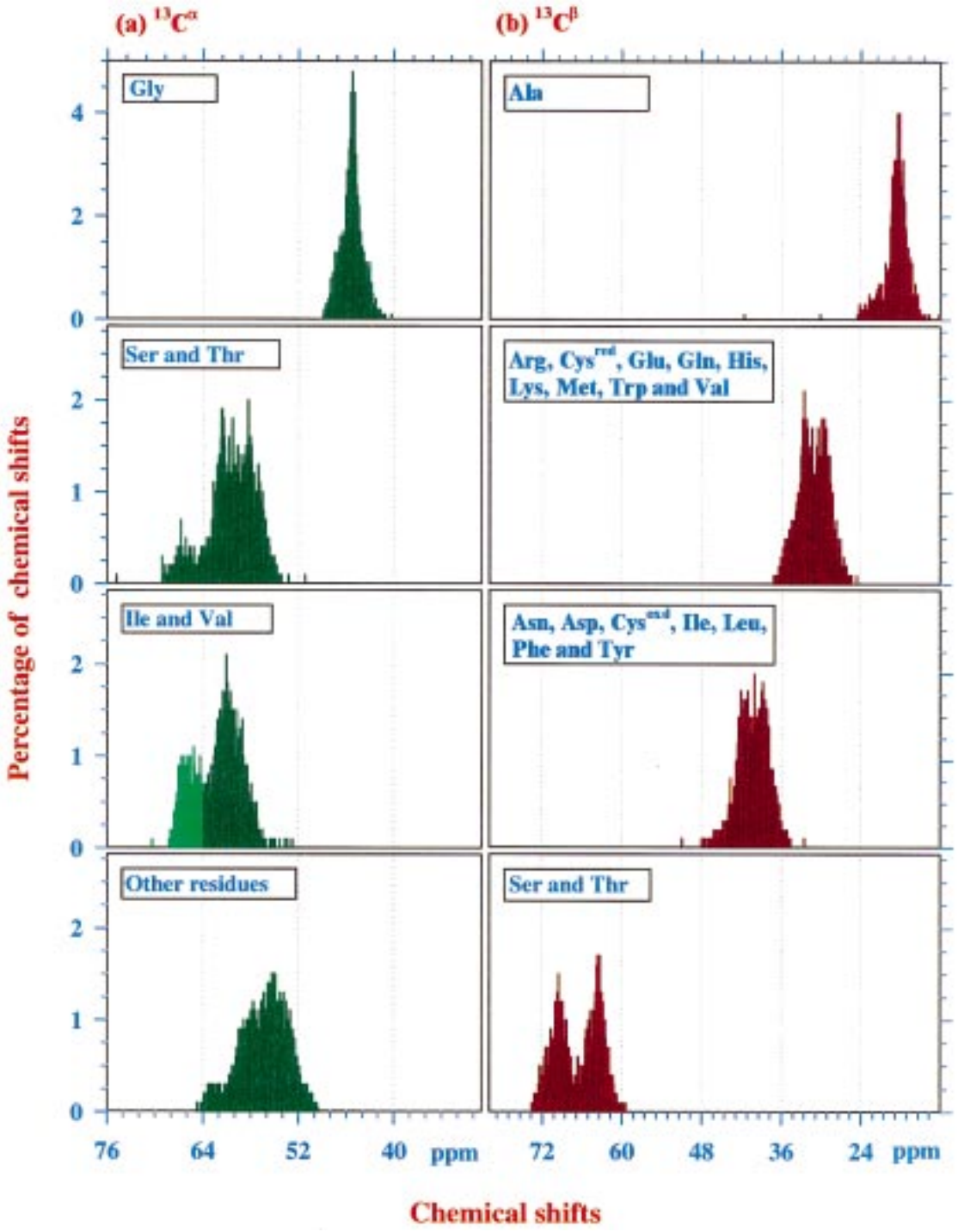

Figure 1. Distribution of (a) $^{13}C^{\alpha}$ and (b) $^{13}C^{\beta}$ chemical shifts for various amino acid residues in proteins selected from the BMRB. The histograms depict the percentage of amino acids having a particular chemical shift within a range of 0.1 ppm. In the case of Ile and Val residues, $^{13}C^{\alpha}$ chemical shifts greater than 64 ppm are shown in a different colour for clarity.

cal shifts of all other residues. On the other hand, no $^{13}C^{\beta}$resonates between 50–58 ppm (Figure 1b). It is also seen that the amino acid residues can be classified into five distinct categories based entirely on the characteristic $^{13}C^{\beta}$ chemical shifts (Figure 1b): (i) Gly having no $^{13}C^{\beta}$; (ii) less than 24 ppm – Ala; (iii) 24–36 ppm – Arg, Cys^{red}, Gln, Glu, His, Lys, Met, Val, Trp; (iv) 36–50 ppm – Asp, Asn, Cys^{oxd}, Ile, Leu, Phe and Tyr; and (v) more than 58 ppm – Ser and Thr.

In our algorithm, each of these five categories is distinguished from the rest by a single digit code. For example, all Gly residues are given a code **1**, Ala residues **2** and so on (Table 1). Besides Ser and Thr residues, Val and Ile residues (about 26% of them) also have their $^{13}C^{\alpha}$ chemical shifts downfield of 64 ppm (Figure 1a). No other amino acid residue has resonances in this region. This facilitates a further classification for Val and Ile residues by appending a second digit to the single digit codes assigned earlier. This second digit is chosen as **1** for all the residues with $^{13}C^{\alpha}$ chemical shifts downfield of 64 ppm and $^{13}C^{\beta}$ chemical shift upfield of 58 ppm and is chosen as **0** otherwise. Thus, for example, a Val residue with its $^{13}C^{\alpha}$ chemical shift downfield of 64 ppm acquires a code **4 1**, while a Val residue with its $^{13}C^{\alpha}$ chemical shift upfield of 64 ppm acquires a code **4 0** (Table 1).

The last two columns in Table 1 indicate the percentage of amino acid residues which violate the two-digit code assigned to them. This might happen if a given residue exhibits unusual $^{13}C^{\alpha}$ or/and $^{13}C^{\beta}$chemical shift(s), as a result of which it acquires a code different from the one generally expected. For example, it is evident from Table 1 that Ser, Thr and Ala residues deviate the least from their expected range of $^{13}C^{\beta}$chemical shifts and rarely do other amino acid residues fall in their range. Further, since these three residues, along with Gly, are given individual codes, it is easy to identify these spin systems uniquely. Hence, these four residues serve as primary markers in ssr_assignments, as has been observed earlier (Metzler et al., 1993).

Experimental inputs for TATAPRO

Several automated assignment strategies that have been proposed in the past require correlating peak lists from a large number (six or more) of triple resonance experiments (Friedrich et al., 1994; Olson and Markley, 1994; Zimmerman et al., 1995; Lukin et al., 1997). These strategies suffer from the fact that there are likely to be some chemical shift variations for the same spin in different spectra because of changes in experimental conditions such as pH and decoupling heating during the experiments. Also, 4D experiments required as input for some of these algorithms suffer from low digital resolution and sensitivity. These factors can contribute to incomplete or/and erroneous assignments. Further, for proteins with low stability, it is imperative that all data be acquired in a short duration of time. Thus, it is desirable to restrict the number of experiments required as inputs to a minimum and all experiments should be performed under identical conditions of pH and temperature, preferably with the same sample. In the present study, four triple resonance experiments, namely CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO, are found to be sufficient for complete ssr_assignment of all the $^{1}H^{N}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and ^{15}N spins. Other 3D triple resonance experiments which can serve as inputs for TATAPRO are HN(CA)HA (Clubb et al., 1992b) and HN(COCA)HA (Clubb and Wagner, 1992) in place of HN(CA)CO and HNCO, respectively. Peak lists obtained from these spectra consisting of chemical shift co-ordinates of the peaks, $(\omega_1, \omega_2, \omega_3) = (^{13}C/^{1}H^{\alpha}, ^{15}N, ^{1}H^{N})$, along with their intensities and phases, are taken as inputs for TATAPRO.

Description of the algorithm

We have considered a deterministic approach here which takes into account the characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of all the 20 individual amino acid residues. The approach can be divided into three important steps, namely, peak list preparation, assignment of two-digit codes (Table 1) to the individual amino acid residues in the primary sequence and the rows in the *master_list* and finally, carrying out ssr_assignments. These steps are described below:

(a) Peak list preparation. Peak lists derived from CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO (or alternatively, HN(CA)HA and HN(COCA)HA spectra) are used to group the chemical shifts as follows. An automatically picked CBCA(CO)NH peak list has information about $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H^{N}_i$. From such a list, the chemical shifts of $^{13}C^{\alpha}_{i-1}$ and $^{13}C^{\beta}_{i-1}$ are identified for each specific pair of $^{15}N_i$ and $^{1}H^{N}_i$ chemical shifts within the user defined tolerance limits and grouped into a single set. Owing to the fact that all peaks in CBCA(CO)NH are seen with positive intensity, distinction between those arising from $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ is based on the following criteria:

Table 1. Two-digit codes assigned to different amino acid residues based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift ranges. The last two columns indicate the percentage of residues which violate these codes

Sr. no.	$^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (δ in ppm) characteristics	Amino acids	Two-digit code	Percentage of $^{13}C^{\beta}$ chemical shift violations	Percentage of other residues taking the code
1	Absence of $^{13}C^{\beta}$	Gly	**1 0**	0.0	0.00
2	$15 < \delta(^{13}C^{\beta}) < 24$	Ala	**2 0**	0.8	0.09
3	$\delta(^{13}C^{\beta}) > 58$	Ser and Thr	**3 0**	0.5	0.04
4	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) < 64$	Lys, Arg, Gln, Glu, His, Trp, Cys^{red}, Val and Met	**4 0**	3.1	2.4
5	$24 < \delta(^{13}C^{\beta}) < 36$ & $\delta(^{13}C^{\alpha}) \geq 64$	Val	**4 1**	1.3	0.6
6	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) < 64$	Asp, Asn, Phe, Tyr, Cys^{Oxd}, Ile and Leu	**5 0**	3.0	2.4
7	$36 < \delta(^{13}C^{\beta}) < 50$ & $\delta(^{13}C^{\alpha}) \geq 64$	Ile	**5 1**	6.8	0.3
8	–	Pro	**6 0**	–	–

(i) If the ^{13}C chemical shift of one of the peaks at ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) is below 50 ppm and the other is more than 50 ppm, then the former is treated as due to $^{13}C^{\beta}$ and the latter to $^{13}C^{\alpha}$.
(ii) If the ^{13}C chemical shifts of both the peaks ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) are more than 50 ppm, then they belong to Ser/Thr residues. Since either peak may be due to the $^{13}C^{\alpha}$ or $^{13}C^{\beta}$ spin, two possible combinations of chemical shifts are considered.
(iii) If only one peak ($^{13}C_{i-1}$, $^{15}N_i$, $^{1}H_i^N$) is seen with its ^{13}C chemical shift below 50 ppm, then it is categorically treated as due to Gly($^{13}C^{\alpha}$) and the corresponding $^{13}C^{\beta}$ chemical shift is set to zero.
(iv) In the event of degeneracy in $^{15}N_i$ and $^{1}H_i^N$ chemical shifts for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks, all possible combinations of chemical shifts are chosen.

On the other hand, an automatically picked CBCANH peak list has information about $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts for a given pair of $^{15}N_i$ and $^{1}H_i^N$ chemical shifts. For each set of $^{15}N_i$, $^{1}H_i^N$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts grouped from the CBCA(CO)NH spectral peak list, the search is now carried out in the CBCANH peak list to identify $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ chemical shifts, within the user defined tolerance limits. In principle, for a given pair of $^{15}N_i$ and $^{1}H_i^N$ chemical shifts, one observes four (^{13}C, $^{15}N_i$, $^{1}H_i^N$) peaks (for non-Gly residues): one pair belonging to ($^{13}C_i^{\alpha}/^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks and the other to ($^{13}C_{i-1}^{\alpha}/^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks. Hence, it may seem straightforward to identify ($^{13}C_i^{\alpha}$, $^{15}N_i$, $^{1}H_i^N$) and ($^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks which are positive and negative in intensity respectively, given that the sequential peaks ($^{13}C_{i-1}^{\alpha}/^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) are already identified. In practice, due to overlap in cross peaks ($^{13}C^{\alpha}/^{13}C^{\beta}$,$^{15}N_i$, $^{1}H_i^N$) of self and sequential residues, one may not find four distinct peaks. Our algorithm then makes use of the following criteria for identifying $^{13}C_i^{\alpha}$ and $^{13}C_i^{\beta}$ peaks:
(i) Since only Gly($^{13}C^{\alpha}$) spins resonate below 50 ppm, all peaks (^{13}C, ^{15}N, $^{1}H^N$) with positive intensity and ^{13}C chemical shift below 50 ppm are ignored (Figure 1a).
(ii) Since no $^{13}C^{\beta}$ spin resonates between 50–58 ppm, all peaks (^{13}C, ^{15}N, $^{1}H^N$) with negative intensity and ^{13}C chemical shift within this range are ignored (Figure 1b).
(iii) The most intense positive peak, excluding the one belonging to the sequential residue, is identified as the ($^{13}C_i^{\alpha}$, $^{15}N_i$, $^{1}H_i^N$) peak.
(iv) The most intense negative peak, excluding the one belonging to the sequential residue, is identified as the ($^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peak.
(v) If the absolute intensity of either or both of the ($^{13}C_i^{\alpha}/^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks happens to be less than twice the intensity of the corresponding sequential peak, or if no peak other than the sequential peaks ($^{13}C_{i-1}^{\alpha}/^{13}C_{i-1}^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) is seen, then the sequential peaks are themselves treated as those for the self ($^{13}C_i^{\alpha}/^{13}C_i^{\beta}$, $^{15}N_i$, $^{1}H_i^N$) peaks. This can happen if either or both of the ($^{13}C^{\alpha}/^{13}C^{\beta}$, ^{15}N, $^{1}H^N$) peaks of self and sequential residues are degenerate.

Once $^{1}H_{i}^{N}$, $^{15}N_{i}$, $^{13}C_{i}^{\alpha}$, $^{13}C_{i}^{\beta}$, $^{13}C_{i-1}^{\alpha}$ and $^{13}C_{i-1}^{\beta}$ chemical shifts are grouped into individual sets, $^{13}C_{i}^{\prime}$ and $^{13}C_{i-1}^{\prime}$ chemical shifts are obtained using automatically picked HN(CA)CO and HNCO peak lists, respectively. Thus, such grouping of chemical shifts results in a peak list containing individual sets of $^{1}H_{i}^{N}$, $^{15}N_{i}$, $^{13}C_{i}^{\alpha}$, $^{13}C_{i}^{\beta}$, $^{13}C_{i}^{\prime}$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}^{\prime}$ chemical shifts. This list, referred to as the '*master_list*', forms the input for the next step in our algorithm.

Each individual set of chemical shifts in the master_list will hereafter be referred to as a *row*. In principle, the number of rows should correspond to the number of amino acid residues in the protein minus the number of Pro residues. In practice, owing to the near degeneracy in $^{1}H^{N}$ and ^{15}N chemical shifts, all possible pairs of chemical shifts within the user defined tolerance limits are accounted for in the master_list. Hence the number of rows usually exceeds the number of amino acid residues. In the case of *Eh*-CaBP with 134 residues, the master_list contained 216 rows.

(b) Assignment of two-digit codes. As discussed earlier, we classify the amino acid residues into eight different categories based on their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, rather than characterizing them individually as has been done in the past. This method of classification helps in a deterministic approach for resonance assignment. The two-digit code assigned to individual amino acid residues (Table 1) is used to tag the individual rows in the master_list depending on the observed $^{13}C_{i}^{\alpha}$ and $^{13}C_{i}^{\beta}$ chemical shift values. In the next step, the master_list is rearranged such that rows belonging to Gly residues are grouped together at the beginning of the list, followed by Ala etc., in the same order as in Table 1. When a polypeptide stretch of amino acid residues is assigned, the two-digit codes associated with the individual rows in that stretch are put into an array, referred to as *assig_array.* Simultaneously, all the amino acid residues in the protein primary sequence are assigned the two-digit code given in Table 1 (Ile and Val residues are given codes **4 1** and **5 1**, respectively). All Cys residues in the primary sequence are assigned a code **5 0**, corresponding to the oxidized state. The reduced Cys residues in the protein then, can be considered as having unusual chemical shifts, which are still assigned unambiguously. However, if most of the Cys residues present in the protein under investigation are in the reduced form, the user can interactively assign these residues a code **4 0**. Thus, on assigning these codes to all the individual amino acid residues in the protein primary sequence, it gets translated into an array of two-digit codes referred to as *pps_array*.

(c) Sequence specific resonance assignment. The algorithm uses the master_list for ssr_assignments. As described earlier, each row in the master_list consists of $^{1}H_{i}^{N}$, $^{15}N_{i}$, $^{13}C_{i}^{\alpha}$, $^{13}C_{i}^{\beta}$, $^{13}C_{i}^{\prime}$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}^{\prime}$ chemical shift values. To begin with, the algorithm reads in the $^{13}C_{i}^{\alpha}$, $^{13}C_{i}^{\beta}$ and $^{13}C_{i}^{\prime}$ chemical shift values from the first row in the master_list and searches for a row where, within the user-defined tolerance limits, these three chemical shifts are seen as $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}^{\prime}$ chemical shifts. If the search is successful, the two-digit code associated with the new row is stored in an assig_array. This procedure corresponds to forward assignment in the primary sequence, which is continued until a break is encountered. The break can be due to a Pro residue, (a) missing peak(s) or the fact that the C-terminal end of the polypeptide chain has been reached. Once a stretch of amino acid residues has been assigned in the forward direction, the algorithm continues with the assignment in the backward direction starting again from the first row in the master_list. For backward assignment, the program reads in the $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$, and $^{13}C_{i-1}^{\prime}$ chemical shifts for a given row in the master_list and searches for the row where these chemical shifts are seen as $^{13}C_{i}^{\alpha}$, $^{13}C_{i}^{\beta}$ and $^{13}C_{i}^{\prime}$ chemical shifts, within the user-defined tolerance limits. If the search is successful, the two-digit code associated with the new row is stored in the same assig_array, as was done in the case of forward assignment. The assignment is continued until a break is encountered. Thus, after assigning the residues in both forward and backward directions, the program maps the assig_array onto the *pps_array*. A one-to-one correspondence with the pps_array results in the sequence specific resonance assignment of that polypeptide stretch. Following this, all the assigned rows are deleted from the master_list before the next round of assignment commences, for which the first row in the updated master_list is chosen as the next starting point. In principle, the above procedure suffices to assign all the amino acid residues in the protein except the Pro residues. In practice, however, several problems can arise when assigning and mapping a stretch of amino acid residues onto the pps_array. We consider each of these in detail:

(i) During the assignment procedure, more than one possible pair of $^{1}H_{i}^{N}$ *and* $^{15}N_{i}$ *chemical shifts satisfy the assignment condition.* The program continues with the assignment along each possible pathway until a break is encountered. Each of the assigned polypeptide stretches, represented in the form of a specific assig_array, is then mapped onto the pps_array. If only one of the assig_arrays gets mapped uniquely onto the pps_array, the rest of the assig_arrays are ignored and resonance assignment is continued. If more than one assig_array correspond to different stretches in the pps_array, no assignment is carried out and all the rows are retained in the master_list. The algorithm then continues the assignment with the next top row in the master_list as the starting point.

(ii) An assigned stretch of amino acid residues occurs more than once in the primary sequence. This happens mostly if the assigned polypeptide stretch of amino acid residues is of a short length (2–4 residues). In the case of proteins with substantial internal sequence homology, larger stretches (5–6 residues) are also found to be redundant. An insight in the statistics of short stretch *polypeptide redundancies* (2–8 residues in length) has been obtained by scanning eight proteins of different lengths ranging from 134 to 370 amino acid residues (Supplementary material, Table 1). In all these proteins, the number of polypeptide redundancies (2–7 residues in length) increases when the 20 amino acid residues are grouped into eight distinct categories compared to when they are considered independently. However, the amino acid stretches comprising eight or more residues are found to be unique. Such stretches can thus be assigned unambiguously. If mapping of assig_array onto pps_array results in multiple matches, the polypeptide stretch is not considered to be assigned and the assignment is continued with the next upper-most row in the master_list. Once a large fraction of amino acid residues are assigned, the number of polypeptide redundancies reduces considerably, leading to unambiguous assignment of stretches spanning even two to three residues.

(iii) One or more residues in the assigned stretch have unusual $^{13}C^{\alpha}$ *or* $^{13}C^{\beta}$ *chemical shifts and therefore do not belong to their expected category.* For example, a Val residue with $^{13}C^{\beta} = 24$ ppm may acquire a code corresponding to an Ala residue. In such a situation, referred to as a '*mismatch*', the mapping of assig_array onto the pps_array will result in either an incorrect mapping or no mapping. For polypeptide stretches spanning eight residues or more, incorrect mapping is unlikely, as these stretches will be unique in the primary sequence (Supplementary material, Table 1). In view of this, the program assigns polypeptide stretches of eight or more residues without a limit on the number of mismatches. For polypeptide stretches of 4–7 residues in length, the program allows only two mismatches, while for stretches spanning 2–3 residues, only one mismatch is allowed. At all stages of assignments, mismatches are reported to the user along with their ^{13}C chemical shift(s).

A statistical survey of $^{13}C^{\beta}$ chemical shifts in the 100 proteins chosen from the BMRB (accession numbers of proteins are listed in the Supplementary material, Table 2) reveals that in a given protein, on average the maximum number of mismatches is 2.3. Figure 2a shows the number of mismatches observed in each of the 100 proteins. Further, three proteins with the largest number of mismatches were analyzed to check the positions of these mismatches along the primary sequence. As shown in Figure 2b, the mismatches in a protein are generally distributed throughout the primary sequence and it is unlikely for a given polypeptide stretch of less than 10 residues to have more than two mismatches. This implies that such a stretch with two mismatches can still be mapped uniquely onto the primary sequence.

(iv) Assignment of a lone residue flanked by two polypeptide segments. During the process of ssr_assignments described above, one may end up with several unassigned lone residues other than prolines, that are flanked by assigned polypeptide stretches. This can happen either because of degenerate chemical shifts or due to the absence of a ($^{13}C_{i}$, $^{15}N_{i}$, $^{1}H_{i}^{N}$) peak in the respective triple resonance spectra. In such an event, the information about the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of the residue preceding the unassigned one is used to assign the ^{15}N and ^{1}H chemical shifts of the latter by utilizing CBCA(CO)NH and HNCO peak lists. Thus, by following this procedure, ^{15}N and ^{1}H chemical shifts of all the lone residues except prolines are assigned unambiguously. On the other hand, the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts of all unassigned lone residues and those of Pro residues are obtained from the row corresponding to their succeeding residue in the master_list.

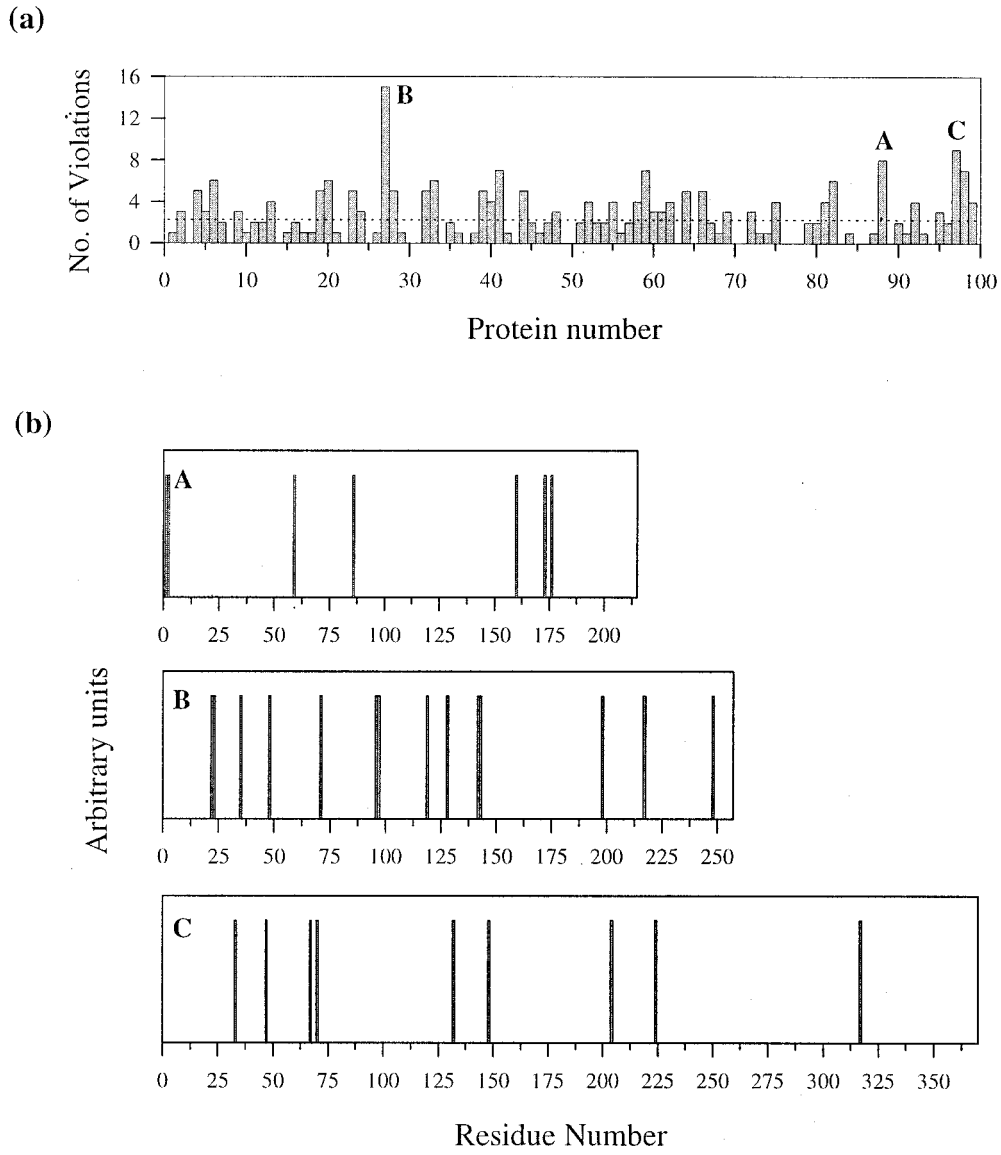

Figure 2. (a) Number of mismatches (amino acid residues which violate our classification of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ and chemical shifts, see text) observed in each of the 100 proteins chosen for statistical analysis. The dotted line indicates the average number of mismatches. (b) Mismatch locations in the primary sequences of three proteins, marked as A, B and C in (a). The respective BMRB accession numbers are: A – 4318, B – 4076 and C – 4354.

Results and discussion

Assignments in Eh-CaBP using experimental data

The algorithm has been tested for NMR assignments in *Eh*-CaBP (134 amino acid residues, 15 kDa) using the experimental data. Ssr_assignments for this protein have been reported elsewhere (Sahu et al., 1999), and were utilized to check the results using TATAPRO. *Eh*-CaBP has the characteristic EF-hands of calcium binding proteins possessing substantial internal sequence homology within the four calcium binding loops. This is evident from its primary sequence shown below, where highly homologous loop segments are highlighted:

MAEALFKEI**DVNGDGAVSYEE**VKAFVSKKRAIKNEQLLQ
LIFKSI**DADGNGEIDQNE**FAKFYGSIQGQDLSDDKIGLK
VLYKLM**DVDGDGKLTKEE**VTSFFKKHGIEKVAEQVMKA
DANGDGYITLEEFLEFSL

Table 2. Details of test proteins and percentage of assignments obtained in each case

Sr. no.	Proteins	BMRB accn. no.	Mol. wt (in kDa)	No. of amino acid residues	No. of mismatches[a]	No. of Pro residues	Percentage of assignments obtained on random deletion of peaks		
							0%	15%	30%
1	Calcium binding protein from *Entamoeba histolytica*	4271	15	134	3	0	96	85	77
2	Drosophila numb phosphotyrosine binding domain	4263	17.8	160	11	6	100	89	68
2	Fibroblast collagenase	4064	18.7	169	0	11	100	87	70
3	*Borrelia burgdorferi* OspA	4076	28	257	14	1	100	92	74
4	*Escherichia coli* maltose binding protein	4354	42	370	9	21	100	90	65

[a]Number of mismatches include reduced cysteines, as all cysteines are given a code **5 0** corresponding to the oxidized form (see text).

Such internal sequence homology complicates the resonance assignment. First, it results in multiple matches when an assigned stretch of amino acid residues in the loop region is mapped onto the primary sequence. Secondly, chemical shifts of ($^{13}C_i^{\alpha}/^{13}C_i^{\beta}/^{13}C_i'$) spins belonging to similar residues in the different loop regions are generally degenerate. This results in more than one pathway for the assignment along the polypeptide chain. Both situations can result in erroneous assignments. However, TATAPRO helps in overcoming these problems, as discussed below.

Automatically picked peak lists were obtained using the software Felix97 (Molecular Simulations Inc., San Diego, CA) from the four 3D triple resonance spectra, CBCANH, CBCA(CO)NH, HNCO and HN(CA)CO. Peaks were picked with a low threshold in all the spectra to avoid missing real peaks with low intensity, particularly in CBCANH and HN(CA)CO spectra. However, peaks from CBCA(CO)NH and HNCO experimental spectra were picked at a higher threshold, because of their inherent higher sensitivity. Thus, for 134 residues, about 3500 peaks were picked in the CBCANH spectrum and 1144 peaks in the CBCA(CO)NH spectrum. The corresponding figures for HNCO and HN(CA)CO spectra were 172 and 299, respectively.

Starting with a tolerance limit of 0.01 ppm along the $^1H^N$ dimension and 0.05 ppm along the ^{15}N dimension, chemical shifts obtained from these spectra were grouped and re-arranged to form a master_list containing rows of $^1H_i^N$, $^{15}N_i$, $^{13}C_i^{\alpha}$, $^{13}C_i^{\beta}$, $^{13}C_i'$, $^{13}C_{i-1}^{\alpha}$, $^{13}C_{i-1}^{\beta}$ and $^{13}C_{i-1}'$ chemical shifts. Wherever the ($^{13}C_i$, $^{15}N_i$, $^1H_i^N$) peaks were not found within this tolerance limit, the tolerance was gradually increased until a set of cross peaks for $^{13}C^{\alpha}$ or/and $^{13}C^{\beta}$ was seen. Next, by beginning at the first row in the master_list, which belonged to a Gly residue ($\delta(^{13}C_i^{\beta}) = 0.0$), sequence specific resonance assignment was carried out using 0.5 ppm as the tolerance limit for $^{13}C^{\alpha}$ chemical shifts, 0.2 ppm for $^{13}C^{\beta}$ chemical shifts and 0.025 ppm for $^{13}C'$ chemical shift. Once the percentage of assigned residues reached around 75%, these tolerance limits were automatically increased to 1.0 ppm, 0.4 ppm and 0.05 ppm, respectively. Both these tolerance limits and the number of residues to be assigned in a single run can be interactively altered by the user, depending on the requirement. Following this procedure, about 95% of the residues could be assigned sequence specifically. The exceptions were M1, A2, E3, I9, N56, Y62, G76 and E111. In the final stage of the algorithm, residues N56, Y62, G76 and E111, each of which were flanked by two assigned polypeptide stretches, were assigned unambiguously. Cross peaks ($^{13}C_i$, $^{15}N_i$, $^1H_i^N$) for M1, A2 and E3 were not observed in any of the aforementioned triple resonance spectra, while cross peaks for I9 could not be picked in the CBCANH spectrum at the chosen threshold. In the present case, there were three mismatches corresponding to V17 ($\delta(^{13}C_i^{\beta}) = 22.5$ ppm), L57 ($\delta(^{13}C_i^{\beta}) = 32.60$ ppm) and K94 ($\delta(^{13}C_i^{\beta}) = 39.7$ ppm). Polypeptide stretches containing these mismatches could still be assigned uniquely. Figure 3 shows the order in which different polypeptide stretches of amino acid residues were assigned during subsequent runs.

To test the robustness of the program, rows in the master_list were deleted randomly (up to 30%)

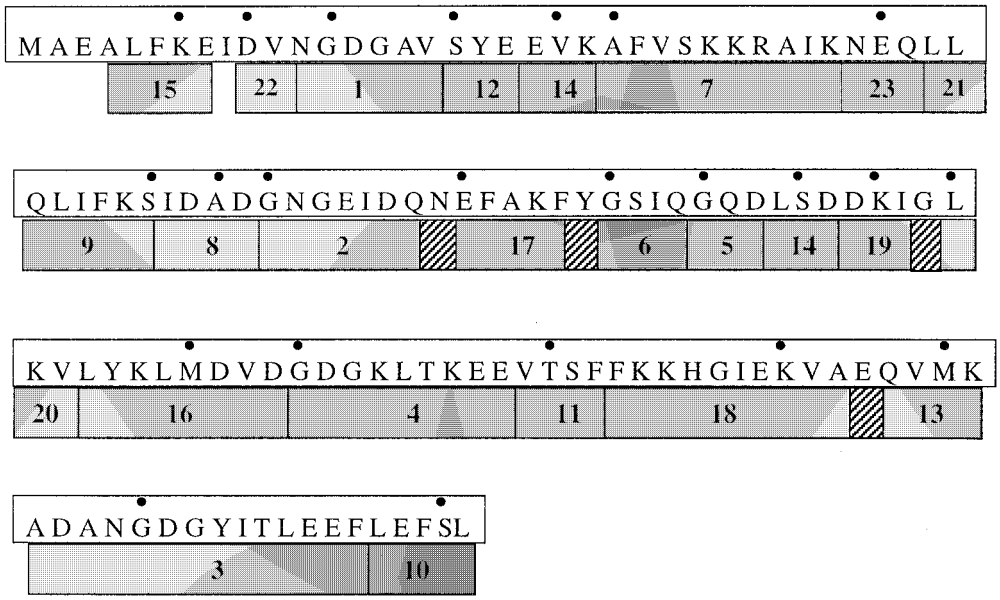

Figure 3. Polypeptide stretches assigned during subsequent runs for *Eh*-CaBP. The numbers in the boxes indicate the order in which these polypeptide stretches have been assigned using TATAPRO. The starting residue within a stretch is indicated by a black dot above it. The hashed boxes indicate lone residues flanked by two assigned polypeptide stretches.

and assignments were carried out, as discussed above. However, the tolerance limits for ^{13}C chemical shifts were increased after about 50% of the residues got assigned. The percentages of assignment obtained after the deletion of 15% and 30% of the rows were 85% and 77%, respectively (Table 2). When a large number of rows are deleted, only short stretches of amino acid residues are assigned (3–4 residues). This results in multiple mapping of assig_array onto the pps_array for such a stretch, leading to a decline in the percentage of unambiguous assignments. Further, we have observed that keeping a high ^{13}C chemical shift tolerance limit (> 0.5 ppm) in the beginning, results in erroneous assignments, as it leads to incorrect assignment pathways. This is more likely when a large number of rows of chemical shifts are deleted from the master_list. Thus, to start with, it is recommended to keep the ^{13}Cchemical shift tolerance as low as possible (about 0.25 ppm for $^{13}C^{\alpha}$, 0.15 ppm for $^{13}C^{\beta}$ and 0.1 ppm for $^{13}C'$). The tolerance limits can be increased automatically when a large number of residues ($\sim$75%) are assigned, as the chances of assignment proceeding along an incorrect pathway then reduce considerably.

Assignments using published data

BMRB data for four other proteins was used to test the reliability of our approach. These proteins, namely, drosophila numb phosphotyrosine-binding domain complexed with a phosphotyrosine peptide (17.8 kDa), a fragment of fibroblast collagenase (18.7 kDa), *Borrelia burgdoferi* OspA (28 kDa) and *Escherichia coli* maltose binding protein (42 kDa), either have a large number of mismatches (Table 2) or have a high degree of polypeptide redundancies within their primary sequence (Supplementary material, Table 1). Further, two of these proteins, namely, fibroblast collagenase and *Escherichia coli* maltose binding protein, possess a large number of Pro residues, restricting the assignable polypeptide stretches to shorter fragments, which in turn can lead to multiple mapping. Thus, resonance assignments in these four proteins constituted a rigorous evaluation for the reliability of our algorithm.

Ssr_assignments in the test proteins were carried out as in the case of *Eh*-CaBP. However, since the data sets were perfect, relatively narrow tolerance limits corresponding to 0.3 ppm for $^{13}C^{\alpha}$ chemical shifts, 0.2 ppm for $^{13}C^{\beta}$ chemical shifts and 0.025 ppm for $^{13}C'$ chemical shifts were chosen for the test proteins. First, all the rows in the master_list created for each protein were retained. This resulted in 100% ssr_assignment of the residues for which assignments have been reported. In the case of drosophila numb phosphotyrosine-binding domain, there are nine Cys

residues with all their $\delta(^{13}C_i^{\beta}) < 36$ ppm, which reflects the reduced state of these residues. In view of this, all the Cys residues were initially given a code **4 0**, which resulted in 100% unambiguous resonance assignments. In order to evaluate the robustness of our approch, the process was repeated with all the Cys residues assigned a code **5 0**, as mentioned earlier. It is interesting that this did not affect the percentage of assignments.

To further verify the reliability of the program, several rows from the master_list were randomly deleted (up to 30%) and the assignments were repeated. The percentage of assignments obtained in each case is shown in Table 2. With 15% random deletion of rows from the master_list, about 90% resonance assignment is achieved in the test cases. For proteins having a large number of Pro residues (which constitute a break during assignments in our algorithm), the percentage of assignments obtained after 30% deletion of peaks declined to 65%. On the other hand, it is observed that a large number of polypeptide segmental redundancies within the primary sequence or a large number of mismatches do not affect the percentage of assignments obtained, establishing the reliability of this approach.

Conclusions

The approach adopted here for resonance assignments resembles to some extent the one proposed by Friedrichs et al. (1994). However, there are subtle methodological differences and improvements. (i) Our algorithm is based on a deterministic approach (as opposed to probabilistic). (ii) Only four triple resonance experiments are required as input to our algorithm. (iii) We classify the 20 amino acid residues into eight different categories, with each category having a characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift range. This approach, which has been found to be more useful and reliable, is facilitated by a two-digit code. (iv) In the event of an unexpected degeneracy in all the $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ chemical shifts, our approach explores along all possible pathways to maximize the stretch of amino acid residues that can be assigned. (v) No manual intervention is required to check the grouping of peaks or to check the residues which do not satisfy their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift range. The latter is checked and reported to the user automatically. The assignments of test proteins in the molecular weight range of 15–42 kDa, wherein assignments up to 75% are achieved even after 30% random deletion of peaks, establish the reliability of the program. Further, the program is shown to be robust to internal sequence homology and unusual chemical shifts. Thus, TATAPRO can be succesfully used for the assignment of large-size proteins.

Acknowledgements

The facilities provided by the National Facility for High Field NMR, supported by the Department of Science and Technology (DST), the Department of Biotechnology (DBT), the Council of Scientific and Industrial Research (CSIR), and the Tata Institute of Fundamental Research, Mumbai, are gratefully acknowledged.

References

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.
Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.
Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.
Choy, W.Y., Sanctuary, B.C. and Zhu, G. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 1086–1094.
Clubb, R.T., Thanabal, V. and Wagner, G. (1992a) *J. Magn. Reson.*, **97**, 213–217.
Clubb, R.T., Thanabal, V. and Wagner, G. (1992b) *J. Biomol. NMR*, **2**, 203–210.
Clubb, R.T. and Wagner, G. (1992) *J. Biomol. NMR*, **2**, 389–394.
Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.
Gardner, K.H. and Kay, L.E. (1998) *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 357–406.
Gardner, K.H., Zhang, X., Gehring, K. and Kay, L.E. (1998) *J. Am. Chem. Soc.*, **120**, 11738–11748.
Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sönnichsen, F.D. and Sykes, B.D. (1998) *J. Biomol. NMR*, **12**, 395–405.
Grzesiek, S. and Bax, A. (1992) *J. Am. Chem. Soc.*, **114**, 6291–6293.
Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.
Hare, B.J. and Prestegard, H. (1994) *J. Biomol. NMR*, **4**, 35–46.
Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.
Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**, 496–514.
Li, K.-B. and Sanctuary, B.C. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.
Li, S.C., Zwahlen, C., Vincent, S.J., McGlade, C.J., Kay, L.E., Pawson, T. and Forman-Kay, J.D. (1998) *Nat. Struct. Biol.*, **5**, 1075–1083.
Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.
Loria, J.P., Rance, M. and Palmer, A.G. (1999) *J. Magn. Reson.*, **141**, 180–184.
Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.
Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) *Biochemistry*, **32**, 13818–13829.
Montelione, G.T., Rios, C.B., Swapna, G.V.T. and Zimmerman, D.E. (1999) In *Biological Magnetic Resonance, Volume 17: Structure, Computation and Dynamics in Protein NMR* (Eds., Krishna, R. and Berliner, L.), Plenum Press, New York, NY, pp. 81–130.
Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Moy, F.J., Pisano, M.R., Chandra, P.K., Urbano, C., Killar, L.M., Sung, M.L. and Powers, R. (1997) *J. Biomol. NMR*, **10**, 9–19.
Olson Jr., J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.
Pham, T.-N. and Koide, S. (1998) *J. Biomol. NMR*, **11**, 407–414.
Sahu, S.C., Atreya, H.S., Chauhan, S., Bhattacharya, A., Chary, K.V.R. and Govil, G. (1999) *J. Biomol. NMR*, **14**, 93–94.
Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13585–13590.
Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wüthrich, K. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.
Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Zimmerman, D.E., Kulikowski, C., Wang, L.L., Lyons, B.A. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.
Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.