Journal of Biomolecular NMR (2006) 34: 41–56
DOI 10.1007/s10858-005-5358-0

Article

PASA – A program for automated protein NMR backbone signal assignment by pattern-filtering approach

Yizhuang Xu[a,b], Xiaoxia Wang[a], Jun Yang[a], Julia Vaynberg[a] & Jun Qin[a,*]
[a]*Structural Biology Program, NB20, The Lerner Research Institute, The Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH, 44195, USA;* [b]*Department of Chemistry, Peking University, Beijing, 100871, P.R. China*

Received 26 July 2005; Accepted 9 November 2005

Key words: automated assignment, backbone assignment, NMR, sequential assignment, triple resonance experiments

Abstract

We present a new program, PASA (Program for Automated Sequential Assignment), for assigning protein backbone resonances based on multidimensional heteronuclear NMR data. Distinct from existing programs, PASA emphasizes a per-residue-based pattern-filtering approach during the initial stage of the automated $^{13}C^{\alpha}$ and/or $^{13}C^{\beta}$ chemical shift matching. The pattern filter employs one or multiple constraints such as $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift ranges for different amino acid types and side-chain spin systems, which helps to rule out, in a stepwise fashion, improbable assignments as resulted from resonance degeneracy or missing signals. Such stepwise filtering approach substantially minimizes early false linkage problems that often propagate, amplify, and ultimately cause complication or combinatorial explosion of the automation process. Our program (http://www.lerner.ccf.org/moleccard/qin/) was tested on four representative small-large sized proteins with various degrees of resonance degeneracy and missing signals, and we show that PASA achieved the assignments efficiently and rapidly that are fully consistent with those obtained by laborious manual protocols. The results demonstrate that PASA may be a valuable tool for NMR-based structural analyses, genomics, and proteomics.

Introduction

Sequential resonance assignment is a key step towards protein structure determination and dynamics studies by NMR spectroscopy. For small unlabeled proteins/peptides, early studies have established that sequential assignment may be accomplished by through-space sequential NOE connectivity along the backbone of a protein, followed by side chain spin system verification using through-bond TOCSY and COSY-type experiments (Wüthrich, 1986). Development of multidimensional triple resonance experiments on isotope-labeled proteins (Bax and Grzesiek, 1993) now permits more unambiguous through-bond sequential assignment and such approach has been successfully applied to many proteins/domains up to the size of 723 residues (Tugarinov et al., 2002). However, the resonance assignment using manual protocols is a laborious and time-consuming process especially for larger proteins and structural genomics/proteomics initiatives, and hence large efforts have been made over the past decade to develop automated or computer-assisted methods for sequential assignment (Friedrichs et al., 1994; Hare and Prestegard, 1994; Meadows et al., 1994; Zimmerman et al., 1994, 1997; Morelle et al., 1995; Lukin et al., 1997; Leutner et al., 1998; Gronwald

*To whom correspondence should be addressed. E-mail: qinj@ccf.org

et al., 1998; Moseley et al., 1999; Atreya et al., 2000; Güntert et al., 2000; Hyberts and Wagner, 2003; Coggins and Zhou, 2003; Slupsky et al., 2003; Malmodin et al., 2003; Hitchens et al., 2003; Baran et al., 2004; Jung and Zweckstetter, 2004). The methods reported so far, albeit having different features, are largely based on three algorithms: (i) Best-first approach that selects the best sequential correlation out of one or more choices (Friedrichs et al., 1994; Zimmerman et al., 1994, 1997; Hyberts and Wagner, 2003; Jung and Zweckstetter, 2004). This approach is effective for small-medium sized proteins with well-resolved spectra, however, errors may be introduced at early stages of the assignment due to various reasons such as limited digital resolution in ^{13}C dimension or resonance overlapping. Best-first approach alone might be suitable for small proteins <25 kD (Zimmerman et al., 1997). (ii) Global search method (Lukin et al., 1997; Leutner et al., 1998; Hitchens et al., 2003; Jung and Zweckstetter, 2004). Global search defines a pseudo-energy for each possible solution of assignment to describe how good the solution fits the matching of chemical shift of adjacent spin systems and types of spin systems. The final assignment is a solution with global minimum energy. To find global minimum point, different methods are adopted including simulated annealing, genetic algorithm, threshold accepting algorithms etc. MONTE, as an example, may be effective on large proteins, which has been tested on a deuterated 45 kDa homodimeric protein (Hitchens et al., 2003).While this approach overcomes some weaknesses of the best-first approach, the energy functions for minimization can be trapped at local minima thus inducing ambiguity. (iii) Exhaustive search method (Atreya et al., 2000; Güntert et al., 2000; Coggins and Zhou, 2003; Jung and Zweckstetter, 2004). The idea for this approach is to enumerate all of the possible choices for a given protein and then eliminate improbable outcomes based on amino acid types, sequence-based fragment filtering, etc, which eventually yields a single correct assignment. Recently, two algorithms including exhaustive search algorithm (e.g., Coggins and Zhou, 2003) and MARS, which combines best-first, exhaustive search, and global search algorithms (Jung and Zweckstetter, 2004) have been shown to be able to deal with large proteins up to 723 residues (MSG). However, degeneracy-induced multiple connectivity problems at initial stage of exhaustive search may propagate and amplify for larger proteins yielding enormous amount of possible choices. The problem may become even more severe if the data sets are incomplete due to missing peaks ultimately resulting in so-called "combinatorial explosion" or "program freeze" before the filtering step is employed.

To alleviate this problem, we have developed a new algorithm by introducing a robust pattern filtering at very early stage of the sequential alignment, i.e., rather than performing filtering after global exhaustive searches that may accumulate large number of possible assignment choices due to degeneracies and missing peaks, we apply pattern-filtering at each step of $^{13}C^{\alpha}$ and/or $^{13}C^{\beta}$ matching by combining all possible constraints such as distinct $^{13}C^{\alpha}/^{13}C^{\beta}$ shifts for amino acids, side chain spin system information, selective labeling, etc. This approach was found to substantially minimize the degeneracy and/or missing signal-induced complication/combinatorial explosion problems. The program also has other distinct features such as using intermediate results as filters, early identification of distinct sites, etc. The program has been successfully tested on four representative proteins with various degrees of degeneracy and missing signals and the results demonstrate that it is a highly efficient and robust tool that may be valuable for NMR-based structural genomics, proteomics, and other applications.

Methods

Data set requirements

The efficiency of any successful sequential assignment program depends upon the completeness of intra and inter $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals. Typically, these signals are provided by HNCACB (providing both intra and inter $^{13}C^{\alpha}/^{13}C^{\beta}$ signals) and CBCA(CO)NH (providing only inter $^{13}C^{\alpha}/^{13}C^{\beta}$ signals) experiments (Bax and Grzesiek, 1993). However, these experiments, especially HNCACB, become less effective when protein sizes become larger, which often results in incomplete data set and ambiguity. On the other hand, HNCA, albeit providing only intra and inter Cα signals, is much more sensitive backbone correlation experiment (Bax and Grzesiek, 1993) that would provide more complete sequential linkages. Hence, to be more effective for large number of proteins, our program emphasizes

the data obtained by HNCA for initial sequential linkages, and utilizes other experiments at the same time such as HNCACB/CBCACONH to preclude false linkages. In addition, our program has the flexibility to incorporate all other useful experimental information such as intra/inter C′ from HNCO/HNCACO, side chain spin systems from C(CO)NH/HC(CO)NH (Bax and Grzesiek, 1993), and selective labeling assignment, etc.

Theory

Pattern filter

Considering a manual assignment process starting from a distinct spin system Gly. If one finds three possible spin systems from HNCACB, which can be connected to the Gly, an experienced spectroscopist would try to eliminate two of the three possible linkages by using primary sequence and other available information such as amino acid type or side chain spin system. We can apply the same strategy on a computer program by introducing a pattern filter at each step of the sequential mapping. The pattern filter can include one or a combination of the following available constraints: (i) intra/inter residue $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift ranges for each amino acid type. A thorough statistical analysis has been performed on the published chemical shifts of $^{13}C^{\alpha}/^{13}C^{\beta}$ from the BioMagResBank (Seavey et. al., 1991) (Figure S1, supplementary material), and the ranges of the chemical shifts for $^{13}C^{\alpha}/^{13}C^{\beta}$ for the 20 amino acid residues are summarized in Table 1. Based on this table, if the chemical shift of $^{13}C^{\alpha}$ and/or $^{13}C^{\beta}$ of a spin system is outside the defined chemical shift ranges for amino acid *X*, the program will eliminate the possibility of the spin system for *X*. (ii) Side chain information as derived from HC(CO)NH and C(CO)NH. This filter is useful even if the side chain information is not complete. For example, if we see a methyl peak in HC(CO)NH and/or C(CO)NH, we can immediately exclude the possibilities of Gly, Ser, Phe, etc. Also for example, if three peaks are observed for a spin system, we can exclude those amino acid residues containing only one or two aliphatic carbons. (iii) Intermediate assignment information. Briefly, any spin-system, if already unambiguously assigned, can be used but once. Also, any site in the target protein can be occupied by only one spin system. (iv) Amino acid type information obtained from specific experiments such as selective labeling, amino acid type identification (Dötsch et al., 1996a, b) and 2D experiments such as MUSIC (Schubert et al., 1999, 2001a, b). (v) Sequential NOEs, which may be used at the final stage to distinguish some degenerate $^{13}C^{\alpha}/^{13}C^{\beta}$ pairs that cannot be resolved by other filtering tools (see results section below).

To express our filtering idea mathematically in a computer program, we define a term called pattern for each spin system: *A* pattern is a string composed of *n* binary codes (0 or 1), where *n* is the number of amino acid residues in the target protein/fragment. The binary code at each site represents the possibility of the spin system to the corresponding site of target protein. The code at site i of spin system *A* is defined as $C(i, A)$. If *A* can be put at site i, $C(i, A)$ is 1; otherwise, $C(i, A)$ is 0. Consider a target protein composed of *N* amino acid residues and a possible linkage from A_1 to A_2: If there is no integer $k(1 \le k \le N-1)$ that satisfies $C(k, A_1)=1$ and $C(k+1, A_2)=1$, A_1 and A_2 cannot be put into any adjacent positions. The possibility of the linkage from A_1 to A_2 is then excluded by the pattern filtering criterion (Figure 1). This criterion can be similarly extended and applied to a chain composed of multiple spin systems. Such per-residue-based filtering is clearly advantageous to prevent accumulation of multiple connectivity problems induced by degeneracy or missing signals during the sequential matching process. Figure 2 illustrates a chain of spin systems $(A_1, A_2, \ldots, A_i, \ldots, A_n)$, where linkage between any two adjacent spin systems (A_j, A_{j+1}) may occur. This chain can be discarded if there is no positive integer *k* that satisfies $C(k+i-1, A_i)=1$ for every integer $i(i >= 1$ and $i <= N-n+1)$.

$$R = \frac{\text{Average number of } C(i,A) = 1 \text{ over all the observable spin systems}}{\text{Number of non-Pro sites in target protein}} \quad (1)$$

Table 1. Ranges of $^{13}C^{\alpha}/^{13}C^{\beta}$ shifts for 20 amino acid residues*

aa	$^{13}C^{\alpha}$ (ppm)		$^{13}C^{\beta}$ (ppm)		Number of chemical shifts	Number of Violation[a]	Percentage of violation (%)[b]
	min	max	min	max			
A	46	64	9	29(26)	9102	8(18)	0.09(0.20)
C	49	69(66)	22 (24)	52(50)	1887	1(18)	0.05(0.95)
D	48	63	32 (34)	50	300	10(15)	0.14(0.21)
E	49	64	23 (24)	38(36)	9480	11(49)	0.12(0.52)
F	49	66(64)	33 (34)	47	4613	16(49)	0.35(1.06)
G	40	50	–	–	9674	20	0.21
H	49	6564)	22 (24)	42(36)	2600	14(51)	0.54(1.96)
I	52	72	30 (34)	46	6506	9(98)	0.14(1.51)
K	48	64	25	42(38)	8920	8(88)	0.09(0.99)
L	48	63	32 (34)	52(50)	10303	11(19)	0.11(0.18)
M	49	64	25	42(39)	2611	6(38)	0.23(1.46)
N	46 (48)	63	32 (34)	47	5122	11(43)	0.21(0.84)
P	52 (56)	69(68)	25 (27)	40(36)	4688	9(58)	0.19(1.24)
Q	49	63	23 (24)	39(36)	4960	6(28)	0.12(0.56)
R	48	64	23 (24)	39(36)	5771	11(46)	0.19(0.80)
S	51	67(65)	56 (58)	72(68)	7105	9(96)	0.13(1.35)
T	51 (52)	73	62 (65)	88	6608	20(73)	0.30(1.10)
V	54	71	25	40(38)	8109	10(25)	0.12(0.31)
W	48	66(64)	23 (24)	39(36)	1415	5(16)	0.35(1.13)
Y	49	66(64)	29 (34)	47	3591	6(44)	0.17(1.23)

[a]Number of assigned amino acid residues locating outside the above ranges.
[b]See Figure S1 for statistical analysis. The average rates of assigned amino acid residues outside the above ranges are about 0.19%. Most of these violations are likely caused by wrong offsets, typos (e.g. Cβ of Ala = 45 ppm in Figure S1), and paramagnetic effects. The ranges were set conservatively based on Figure S1, however, our experience (see text) indicated that narrower ranges with $<2\%$ violations (see numbers in the parentheses) substantially increase the speed of the calculations without affecting the accuracy of all the tested proteins.

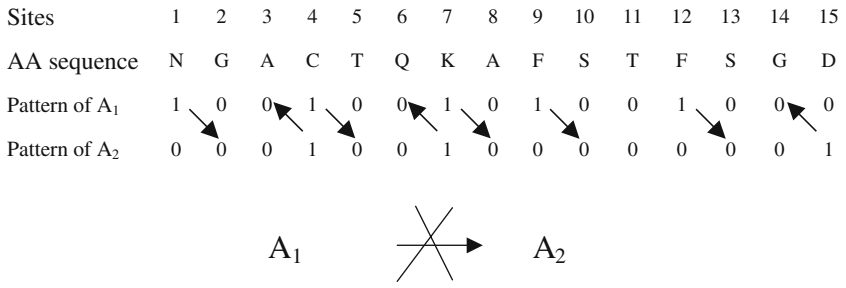

Figure 1. Schematic diagram showing how pattern-filtering criterion is used to preclude a false linkage using a 15 residue fragment as an example. Suppose that two spin systems A1 and A2 can be potentially placed in any adjacent sites in this fragment. Using patter filtering, the program may eliminate the possibilities of A1 at sites 2, 3, 5, 6, 8, 10, 11, 13, and 14, and A2 at sites 1–3, 5–6, 8–14 (codes = 0). After this filtering, the possibility of A1–A2 linkage is precluded since no adjacent positions can satisfy both $C(k, A_1) = 1$ and $C(k+1, A_2) = 1$ (see the diagonal arrows).

To quantitatively evaluate the effect of the pattern-filtering, we introduce a characteristic ratio (R) for the pattern as defined by Equation (1). Here R can be regarded as an average probability for those $C(i, A) = 1$. R is always <1 and its magnitude depends on how many experimental data are incorporated into the pattern. R ranges between 0.08 and 0.45 for all the proteins we have studied. If a chain A_1–A_n starting from site k contains false connection(s) but still passes pattern filtering, the probability for such scenario will be R^n. Since $R < 1$, the probability of survival for the

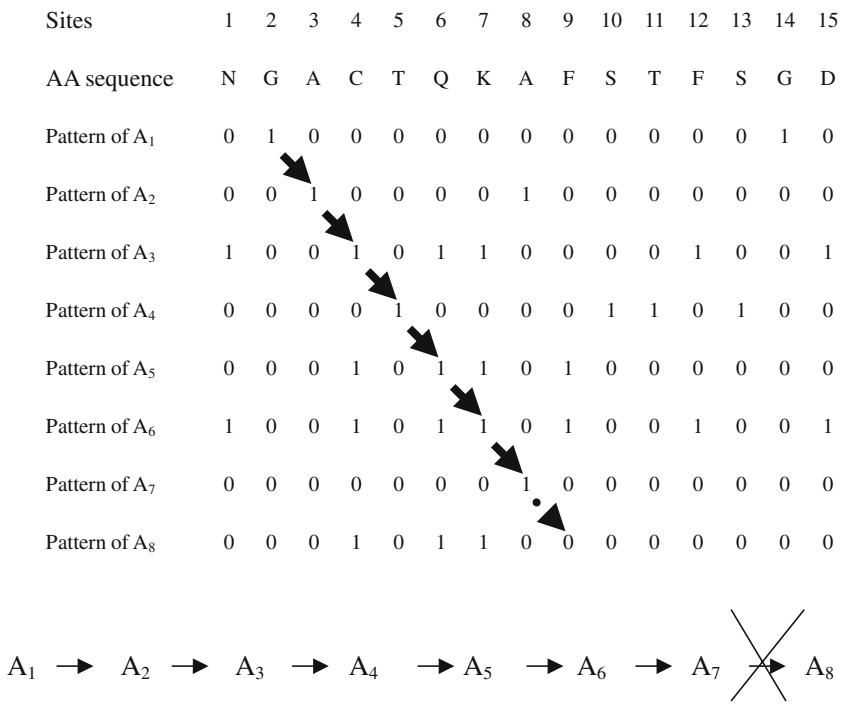

Figure 2. Schematic diagram showing how pattern-filtering criterion is used to preclude a long chain of multiple spin systems containing a false linkage. Considering eight spin systems (A1–A8) that can be potentially aligned into the 15 residue fragment shown in Figure 1. Matching of intra and inter chemical shifts followed by pattern filtering results in a linkage from A1 to A7, however, the A1–A8 chain is excluded if a mismatching of patterns occurs between A7 and A8 (see the dotted arrow from site 8 to site 9).

falsely-linked chain after pattern filtering would approach zero rapidly as n increases. For example, if $R = 0.3$ and $n = 6$, the probability of the wrong linkage would be 0.07% and hence the survived assignment of such fragment is most likely the correct one. This unique property allows unambiguous assignment for longer chains at initial stages. In other words, the longer the chain, the more powerful the pattern filter would impact on the assignment. Assignment of the shorter chains ($n < 6$) can be readily obtained at later stages of the automation process after incorporating longer chain assignments into the new pattern filter. This was vigorously examined in our study and was found to be extremely effective.

Exhaustive search with "growth and filter" cycles

The automated assignment calculation is performed by exhaustive spin system search in a "growth and filter" fashion. The growth step can start from a given site i. All the spin systems that can be placed on the site are selected based on the filtering criteria. The candidate spin systems for $i+1$ site are then found via matching the intra-residue/inter-residue chemical shift(s). Immediately after this growth step, pattern filtering is performed. If multiple candidate spin systems exist, this step will eliminate some degeneracy-induced false connections. The program then moves to the next growth-filter step and goes on until a chain of possible linkages is formed.

Let's consider a potential combinatory explosion problem mathematically and how our program may alleviate it. Considering that a chain ($A_1, \ldots, A_n$) starting from site i has passed the pattern-filtering and a new spin system A_{n+1} is matched to A_n based on the intra- and inter-C^{α}/C^{β} shifts, this new linkage is acceptable when $C(i+n, A_{n+1})$ is 1. As described above, we know that the probability for $C(i+n+1, A_{n+1}) = 1$ or accepting chain $A_1, \ldots, A_n, A_{n+1}$ is R. However, there may be on average M candidates of A_{n+1} that match intra-residue and inter-residue chemical shifts, which will lead to $M \times R$ (MR) possible chains that satisfy the requirement of pattern filtering. If we try to assign X spin systems starting from A_n, the program will calculate X times, which will yield

possibly $(MR)^x$ chains. If MR < 1, all the possible chains will die out as X increases except the one with correct linkage that will always satisfy the requirement of pattern filtering. However, if multiple connectivity problems occur (MR > 1 or >>1), $(MR)^x$ will become enormous, which then leads to the so-called combinatorial explosion. Hence a feasible algorithm must keep the MR value as low as possible. Clearly, there are two ways to improve the performance of the exhaustive search algorithm: (i) Reduction of R value. This can be obviously achieved by introducing various amino acid type information into patterns, which is the key feature of our program. (ii) Reduction of M value. One obvious approach is to adopt strict but reasonable tolerance during the chemical shift matching of $^{13}C^{\alpha}$ and/or $^{13}C^{\beta}$. Figure 3 shows the variation of M value during the matching of $^{13}C^{\alpha}$ shifts for a 15 kDa CH_2 domain of ILKBP (Tu et al., 2001) (see application below). Different values of tolerance were given in Figure 3 and it is clear that larger value of tolerance results in remarkable increase of M value and potential combinatory explosion. Hence, selection of proper tolerance value to limit the M value without losing real connectivity is critical for establishing a feasible algorithm. To achieve this, one should first minimize the subtle spectral differences among different triple resonance data. For example, it is preferable to collect all correlation experiments on the same sample and instrument. Secondly, one can minimize digital resolution-induced error by performing the matching of $^{13}C^{\alpha}$ and/or $^{13}C^{\beta}$ peaks within the same spectrum (e.g., HNCA or HNCACB). In this manner, the influence of the digital resolution and reference is kept minimal. Thirdly, one can use a dual tolerance for well-resolved peaks and overlapping peaks, respectively. We conducted a statistic analysis of $^{13}C^{\alpha}$ shifts from the assignment results performed in our lab to find out the differences of intra/inter $^{13}C^{\alpha}$ peaks between adjacent spin systems. Figure 4 shows that the differences can be classified into two types: Type I containing well-resolved intra/inter $^{13}C^{\alpha}$ peaks where the differences between the adjacent $^{13}C^{\alpha}$ peaks are within 0.1 ppm (Figure 4a). Type II containing intra- and/or inter–$^{13}C^{\alpha}$ peaks that overlap with other peaks (Figure 4b). Type II would clearly introduce errors when chemical shifts are obtained using NMR software such as PIPP (Garrett et al., 1991). We found that the maximum error was within 0.6 ppm for $^{13}C^{\alpha}$ among all the datasets we have examined. Based on this, we may set up a dual tolerance for our datasets: 0.12 ppm when no overlapping occurs on

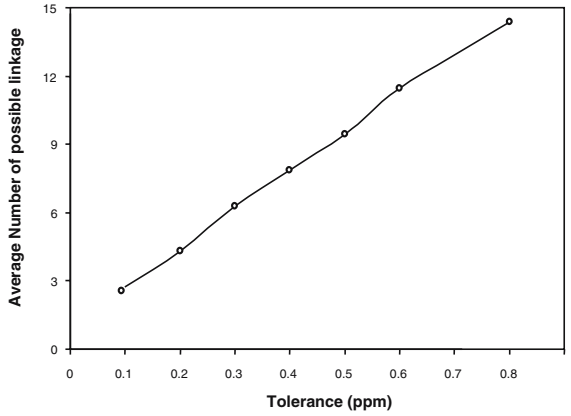

Figure 3. Variation of average number of possible linkages (M) with different tolerance values based on the data of C matching for ILKBP CH_2 domain.

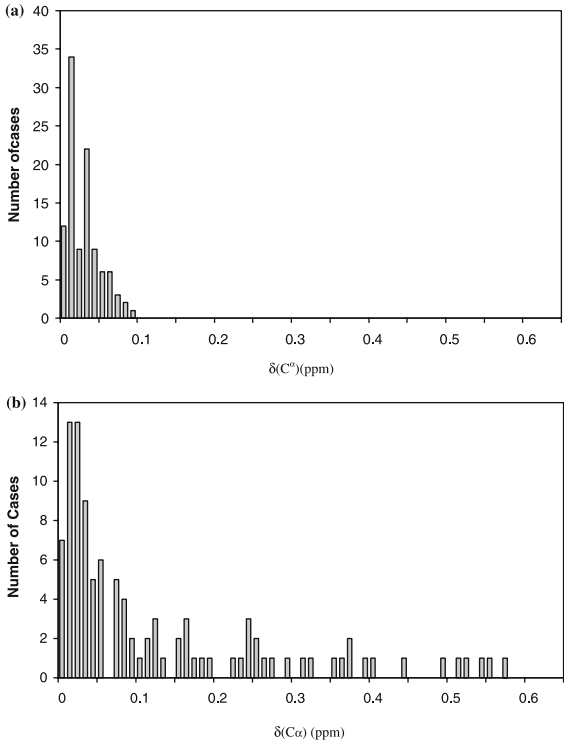

Figure 4. Histograms of C shift differences based on PIPP-based peak-picking. (a) The differences between residue i and i–1 C peaks. Neither i nor i–1 C peak overlap with other peaks. (b) The differences between residue i and i–1 C peaks. Either i or i–1C peak overlaps with other peaks.

intra-residue/inter-residue $^{13}C^{\alpha}$ peaks and 0.6 ppm when there is overlap. The dual tolerance values are of course variable depending on user's experimental parameters and hence our program provides an interface to change the tolerance based on user's specific experimental conditions.

Special considerations in the program

The above features constitute the main thrust of our algorithm. Additional considerations were made to further polish the program, which are especially effective for larger proteins with severe degeneracy and missing signals.

(a) Use of intermediate results to assist the automated assignment: As briefly mentioned above, the program uses the intermediate results to assist the further assignment process. This consideration was based on the fact that the assignment process is carried out in a residue-by-residue manner. Our criteria are as follows: (i) any spin-system, if already unambiguously assigned, can be used but once; (ii) any site in the target protein can be occupied by only one spin system. The restriction provides us with another powerful tool to eliminate the degeneracy problems in ^{13}C chemical shifts. Two special treatments are thus adopted in the program. First, any linked chain will be discarded if the same spin system appears twice in it. Second, the unambiguously assigned spin systems can be used as additional filters, i.e., when a spin system A is assigned to the site i of the target protein, $C(i, A)$ will be set 1 and $C(j, A)$ will be as 0 when $j \neq i$. Also, for any spin system B other than A, $C(i, B)$ will set as 0. The R value can decrease considerably after the above treatments.

(b) Exhaustive search from distinct sites: Experienced spectroscopists often start assignment from some amino acid residue with distinct chemical shifts such as Gly, Ala, Ser, and Thr. The unusual chemical shifts of these residues often have less degeneracy problems. The program is thus designed to allow the users to begin the assignment from some special sites occupied by Gly, Ala, Ser, Thr or the sites where the degeneracy on ^{13}C is not severe. Once some assignments are obtained, the degeneracy problem will be alleviated to some extent. As a result, the assignment will propagate from less degeneracy to high degeneracy part thus reducing the degeneracy-induced complication.

(c) Linkage with Proline: Linkage with proline is often regarded as a tough task. Due to the lack of amide proton, the connection between proline and other amino acid residue cannot be established via the conventional matching. However, the chemical shifts of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ in proline i can still be observed from the inter-residue chemical shift of $i+1$ spin system. Moreover, based on Table 1, any false connection between a proline and its $i+1$ spin system can be excluded if their inter-residue Cα and/or Cβ shifts are outside the distribution profile of the corresponding proline.

(d) Missing entire spin system: This is a critical issue. In some cases, entire spin system(s) for non-proline amino acid residues do not show up because of line-broadening or fast amide/water exchange problem. This phenomenon raises the risk of wrong assignment during the automation process. Considering a polypeptide segment $(T_1, \ldots, T_{j-1}, T_j, T_{j+1}, \ldots, T_n)$ where the correct assignment should be $(A_1, \ldots, A_{j-1}, A_{j+1}, \ldots, A_n)$ and T_j is a breaking point (a site with missing NH). When performing the assignment without knowing the missing spin system T_j, another spin system B_i may be falsely linked with A_{j-1}. Since many assignments are performed in a stepwise fashion and intermediate results are used for further assignment, the false linkage may propagate along the logic chain and cause the collapse of the entire assignments. To minimize the risk of such false linkages, we take the following steps: (i) Inputting as much filtering information as possible. We emphasize the combined use of intra-residue and inter-residue information, which can increase the possibility of identifying breaking points (sites with code = 0 for all the observable spin systems) at early stage. Some distinct sites such as Ala followed by Gly may be identified by the program and used as distinct patterns to help to validate the segmental assignment and identify some breaking points. (ii) Dividing the calculation into several levels. In each level, PASA will scan the entire target protein by performing assignment independently on each polypeptide segments that do not contain any breaking points identified by the program. At the end of each level, cross validations will be made on the assignment results from different polypeptide segments. In the assignment results, each site must have only one spin system. If a spin system is shared by multiple sites in different segments, this must result from

missing spin systems and we call this the assignment conflict, which will be rejected and the relevant site is then treated as potential breaking point for an additional round of calculation to resolve ambiguity. The conflict-free assignment from different segments are then combined and used as intermediate filter for the next level. (iii) The program will only accept the assignment from long polypeptide segment at the initial stage of the automation (e.g., $n > 6$ as discussed above). This is based on the fact that the longer the assignment chain one obtains, the less possibility (R^n) the results contain false linkages (see the theory above). If a long polypeptide from site M to site N contains a breaking point at site T, the calculation is expected to stop at site T, which suggests a breaking point. However, caution has to be taken since we found that the calculation sometimes stops a few residues after T due to severe degeneracy (in all the proteins we have tested, this happened a few times with maximally 4 residues passing after T). To minimize this problem, PASA is designed to have option to only conservatively accept a portion of the assigned long chain ($n > 12$) by eliminating six residues at its C-terminus. The discarded C-terminal sites can be used for cross validation to find conflict assignment. Note that we found that this special treatment is only necessary at the initial levels of calculation for large proteins with severe degeneracy and missing signals, e.g., MSG, the largest protein assigned by NMR (Tugarinov et al., 2002) (see below).

(e) Isotope effect on ^{13}C shifts of extensively deuterated samples: The ^{13}C shifts for extensively deuterated samples, especially for $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ used for matching, are often not identical to non-deuterated ones. The isotope-induced ^{13}C shift changes can range between 0 and 0.5 ppm and sometimes may be larger than 1.0 ppm. However, our chemical shift ranges in Table 1 based on the statistical analysis cover both non-deuterated and deuterated ^{13}C shifts that have been deposited into BMRB. The table is suitable for more than 98% of chemical shifts of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ no matter whether a protein is deuterated or not. For example, in addition to the largest protein MSG whose shifts were all deuterated, the other four large proteins whose backbones have been assigned and deposited into BMRB were all based on the deuterated samples (Gardner et al., 1998; Garrett et al., 1999; Tugarinov et al., 2002; Rajesh et al., 2005) and none of their shifts (total 1753 including 127 $GlyC^{\alpha}$) were violated according to Table 1 except one if we use wider chemical shift ranges in Table 1. If narrower chemical shift ranges defined in Table 1 (see numbers in parantheses in Table 1) were used, eight C^{β} were violated but the rate is only 0.49% that is smaller than the average violation rate 2% (see Table 1). For the few violated C^{α} and C^{β} shifts, PASA can treat the corresponding sites as breaking points (see above missing spin system section). A series of strategies have been designed to protect the breaking points from the misassignment so that the users can use dataset with wider chemical shift range to get correct assignment (see above). Of course, some manual intervention may be used at the later stages to identify these few violations.

It is important to point out that the isotope effects on ^{13}C shifts may cause problem for matching if shift data from different samples are used. Therefore, it is recommended to use a single sample for all backbone and side chain ^{13}C correlation experiments for all the matching. On the other hand, since we always use the inter-/intra data from the same spectrum for chemical shift matching, the chemical shift differences between deuterated/partially deuterated and protonated samples should in practice have little impact to each other as long as we can use the same amide proton and nitrogen to assign different set of peaks to the same spin system (amide ^{1}H and ^{15}N shifts are not affected by deuteration).

(f) overlap in the H–HN plane: In theory, the number of HSQC peaks excluding those for side chain NH2s should correspond to the number of the amino acids except prolines and the first amino acid residue at the N termini. However, for large proteins where the number of HSQC peaks is less than expected, there may be severe overlaps for both HN and ^{15}N resonances. The user may then examine the possible H–HN overlap in HSQC spectrum by searching extra C' peak(s) in HNCO spectrum, and/or C^{α} peak(s) (mostly at $50 \sim 65$ ppm, except for Gly) in sensitive HNCA/CBCA(CO)NH spectra. If distinct multiple peaks are observed, the user may project the corresponding 3D data into different planes and go over slices to assign peaks to their own spin system. For example, in HNCACB data, a projection to H–C plane will be useful to distinguish overlapping peaks along nitrogen dimension in HSQC

spectrum; a projection to N–C plane will be useful to distinguish overlapping peaks along proton dimension in HSQC spectrum. If necessary, one can also project the data to H–N plane to further resolve the overlap.

The program and its execution

The program written in *C* language was developed in a modular fashion, where each module has one or several well-defined tasks. All calculations were performed on a PC LINUX UNIX system (2.4 GHz) as well as a slow Indigo II SGI UNIX Workstation (250 MHz). The flowchart for the program operation is illustrated in Figure 5, which is composed of data preparation stage and iterative assignment stage. There are three tasks to be accomplished at the first stage: (a) HSQC-based peak-picking (intra and inter $^{13}C^{\alpha}/^{13}C^{\beta}$) from the correlation spectra. A table containing the intra and inter $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks is generated for each HN. The missing $^{13}C^{\alpha}/^{13}C^{\beta}$ peaks are marked as 0 so that the program can treat them properly. (b) Pattern is generated for each spin system according to its intra/inter $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ peaks. As options, some other experimental information such as side chain peaks from C(CO)NH/HC(CO)NH and/or specific amino acid type identification experiments may be utilized for the generation of patterns. That is to say, patterns for filtering are automatically created at the beginning stage of the calculation. Some unique patterns may be found at this stage, e.g., Gly–Ala, Ser–Thr, etc., which can be used by program in the iterative calculation process. (c) Construction of an assignment table in which each spin system is mapped to a site in the target protein and the corresponding site for each spin system is set as 0 in initial state. Tables 2–3 display examples of input data format and output data, respectively.

The iterative calculation process involves six steps and begins in an interactive fashion. At step 1 (Figure 5), the users are required to input starting and ending sites for the target polypeptide segment. The starting and ending sites are just site numbers in the target protein, e.g., the 3rd and 18th position from *N* terminal of the target protein. In most cases, the polypeptide segment is a fragment sandwiched by prolines. At step 2, the program will find all spin systems that may be mapped onto the starting site. Step 3 is a growing step where possible linkages are set up at the C termini of the chain by $^{13}C^{\alpha}$ matching. At step 4, the program performs the chemical shift matching for intra- and inter–$^{13}C^{\beta}$ and/or C′,H^{α} to remove degeneracy-induced false linkages during Step 2. At step 5, pattern filtering is performed to filter out additional false linkages. After this, the program will return to step 3 to continue the growth-filter cycle calculation until the ending site that is defined by the user. At step 6 the program will check if a unique assignment has been obtained. If so, the result will be written into an assignment table so that the program may use it as an intermediate filter during the next cycle of the calculation. Step 1–6 is repeated for other segments separated by prolines so that the entire protein sequence will be screened during the first cycle of the calculations. These six steps, executed by one command, require little manual intervention except for the need to manually enter the starting and ending sites at Step 1. The second iterative cycle of calculation will be performed if only partial assignment of the protein is obtained in the cycle. The iterative cycles of calculation goes on until no further assignment can be made. Obviously, more cycles of calculation will be needed to achieve the assignment for proteins with more degeneracy and missing spin systems (see the application blow). Note that if a breaking point, i.e., a site with missing entire spin system, was found during a cycle of calculation, this point will be treated as a pseudo-proline, e.g., a new starting or ending site, in the next cycle of calculation.

Results and discussion

We have tested the program on four target proteins whose spectral features vary widely including the degree of chemical shift degeneracy and missing peaks. As described below, these tests have provided vigorous assessment for the robustness of our program. The assignments of the first three target proteins have been manually performed in our laboratory in a parallel fashion whereas the last one, MSG has been reported recently (Tugarinov et al., 2002). The results of these tests are described below:

(1) Assignment on Nck-2 SH3-3 domain involved in interacting with PINCH LIM4 domain during cell spreading and migration (Velyvis et al., 2003; Vaynberg et al., 2005). This is the most straightforward test due to the excellent chemical

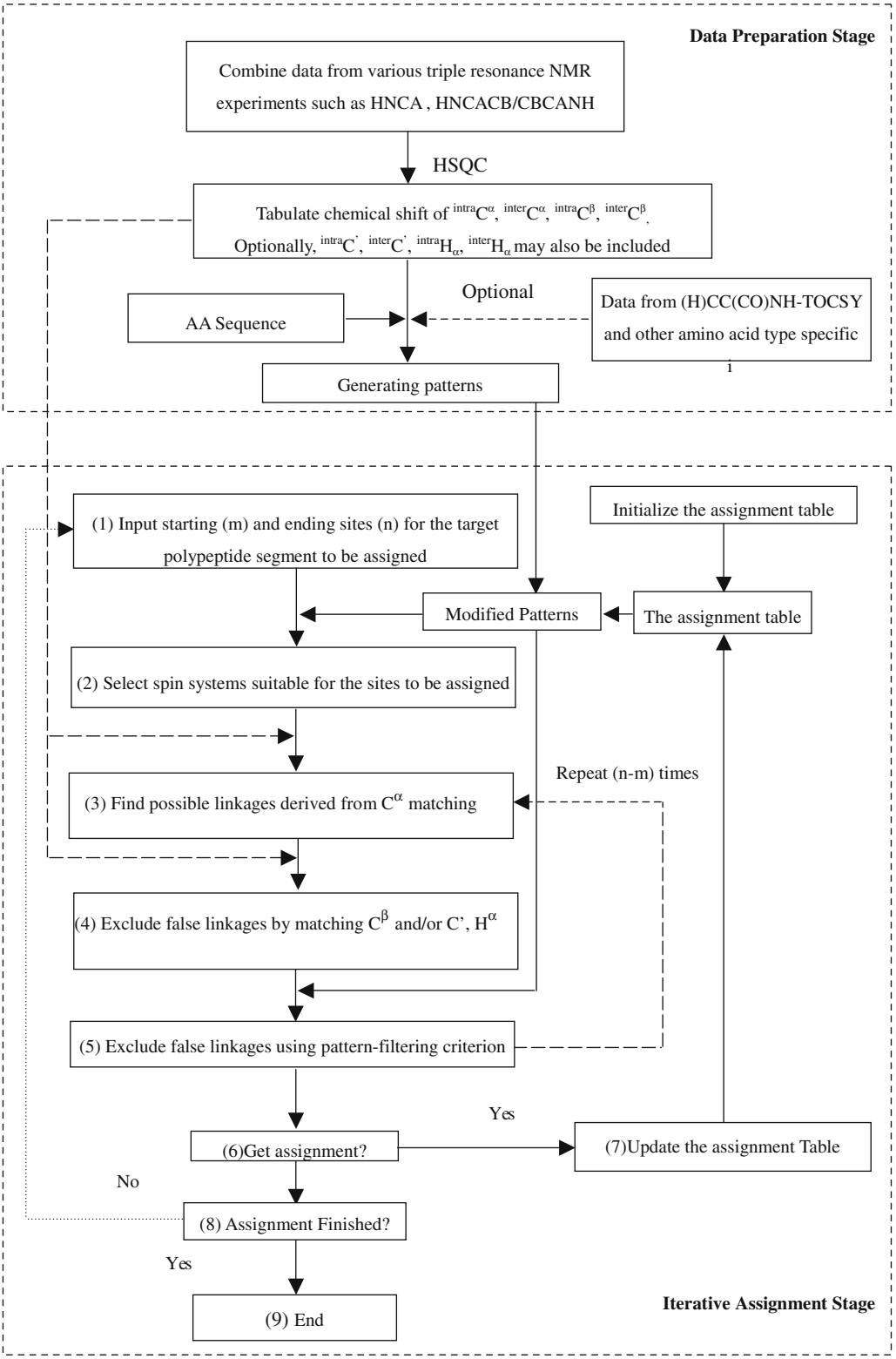

Figure 5. Diagram of the PASA assignment procedure.

Table 2. Examples of PASA input data

No	SS[a]	$^{Intra}C^{\alpha}$	$^{Inter}C^{\alpha}$	$^{Intra}C^{\beta}$	$^{Inter}C^{\beta}$	$O(^{intra}C^{\alpha})^{b}$	$O(^{inter}C^{\alpha})^{b}$
Input data 1 – Chemical shifts							
.....							
10	13	54.940	44.270	45.090	0.0000	1	0
11	14	53.910	36.290	57.350	34.780	0	1
12	15	54.540	18.750	51.070	39.100	0	1
13	16	60.700	39.520	62.210	32.390	1	1
14	17	53.950	37.020	61.350	69.740	1	0
15	18	62.270	32.490	54.480	32.490	1	1
16	19	56.520	34.120	52.790	33.020	1	1
17	20	58.670	29.800	63.130	32.260	1	1
18	21	51.080	39.120	53.900	36.330	1	1
19	22	53.300	31.930	56.610	34.560	1	0
20	23	56.280	42.560	62.940	32.370	0	1
21	25	63.780	31.890	58.440	35.070	1	1
22	27	62.880	32.380	53.700	34.460	1	1
23	28	58.330	29.510	60.520	70.270	1	1
24	29	53.930	41.250	57.640	64.480	1	1
25	30	58.000	33.200	53.830	31.680	1	1
.....							
Input data 2 – Sequence of a target protein Nck2 SH3 domain[c]							
QGSRVLHVVQTLYPFSSVTEEELNFEKGETMEVIEKPENDPEWWKCKNARGQVGLVPKNYVVVLSDGPALH							

[a]Spin system ID.
[b]Overlaping code for intra and inter residue C^{α} peaks: if no overlapping occurs on the C^{α} peak, the code is 1; if overlapping occurs, the code is 0.
[c]Although the sequence is entered once, the calculation is performed by fragments sandwiched by prolines with corresponding output (see Table 3).

shift dispersion and completeness of the data with only one missing spin system. The $^{15}N/^{13}C$ labeled SH3 domain is composed of 71 residues including 5 prolines. Except for the first N-terminal residue and 5 prolines, there should be 65 sites. 64 spin systems (both intra and inter $^{13}C^{\alpha}/^{13}C^{\beta}$) were obtained from HNCACB. The single missing spin system (the breaking point) was readily identified at the data preparation stage (pattern formation) as site 2 (Gly) by the program based on HNCACB and C(CO)NH/HC(CO)NH, i.e., HNCACB only revealed four out of five Gly spin systems and the $i-1$ spin systems of all the observable Glys do not match the $i-1$ (Gln) of the site 2 Gly.

Because of the excellent chemical shift dispersion and data quality, the automated assignment was carried out in a single data set, HNCACB as also done in the manual assignment. Briefly, the sequence was divided into six polypeptide segments separated by five prolines and the assignment for each segment was performed sequentially by the program. After the first cycle of the calculation during which the program scanned all six polypeptide segments, unique assignment results were obtained for 42 out of 64 sites. 22 remaining sites had multiple choices due to degeneracy. Using 42 assignments as an intermediate result (filtering), the program identified additional 19 sites during the second level of the calculation. The remaining three sites were readily assigned after using the 61 sites as an intermediate filter during the third level of calculation. The three levels took less than two minutes on the Indigo II SGI workstation and seconds for the PC yielding the same result as obtained by the manual protocol (Vaynberg et al., 2005).

(2) Assignment on a point mutant of β3 integrin cytoplasmic domain. The second test was on a β3 integrin cytoplasmic domain mutant where a serine at position 752 was mutated into a proline (termed S752P-β3 here). The S752P mutation leads to *Glanzmann's thrombasthenia*, a severe congenital

Table 3. Output data for the assignment on Nck-2 SH3 (fragment 14–37)

Site	aa	SS[a]	$^{intra}C^{\alpha}$	$^{inter}C^{\alpha}$	$^{intra}C^{\beta}$	$^{inter}C^{\beta}$	$O(^{intra}C^{\alpha})$[b]	$O(^{inter}C^{\alpha})$[b]	$\Delta(C^{\alpha})$[c]	$\Delta(C^{\beta})$[c]
.....										
14	P	0								
15	F	46	58.160	61.980	43.600	32.470	1	1		
16	S	63	55.630	58.150	64.250	43.570	1	1	0.01	0.03
17	S	43	56.690	55.600	64.760	64.10	0	0	0.03	0.15
18	V	31	61.980	56.720	32.070	64.810	1	0	0.03	0.05
19	T	66	60.550	61.980	70.270	32.110	1	1	0.00	0.04
20	E	28	58.330	60.520	29.510	70.270	1	1	0.03	0.00
21	E	41	57.840	57.840	29.920	29.920	0	0	0.49	0.41
22	E	64	55.020	57.720	30.600	30.600	1	1	0.12	0.68
23	L	53	54.620	54.620	43.690	30.660	0	0	0.40	0.06
24	N	65	52.460	54.490	40.800	43.630	1	1	0.13	0.06
25	F	34	56.160	52.420	40.790	40.790	0	1	0.04	0.01
26	E	7	53.800	56.140	31.680	40.820	1	1	0.02	0.03
27	K	3	58.000	53.830	33.200	31.680	1	1	0.03	0.00
28	G	42	45.540	57.980	0.000	33.190	0	0	0.02	0.01
29	E	48	57.170	45.510	31.730	0.000	0	1	0.03	−1.0[d]
30	T	33	61.470	57.170	69.710	31.730	1	1	0.00	0.0
31	M	17	53.950	61.350	37.020	69.740	1	0	0.12	0.03
32	E	44	54.320	54.320	32.590	37.000	0	0	0.37	0.02
33	V	18	62.270	54.480	32.490	32.490	1	1	0.16	0.10
34	I	16	60.700	62.210	39.520	32.390	1	1	0.06	0.10
35	E	47	56.620	60.610	34.610	39.540	1	1	0.09	0.02
36	K	22	53.300	56.610	31.930	34.560	1	0	0.01	0.05
37	P	0								
.....										

[a]Spin system Id from HSQC. Notice the continuous non-proline sequence sandwiched by Pro14 and Pro37 (see Table 2 for the complete sequence of SH3 domain).
[b]Overlapping code for intra and inter residue C^{α}peaks: if no overlapping occurs on the C^{α}peak, the code is 1; if overlapping occurs, the code is 0.
[c]$\Delta(C^{\alpha}) = |^{intra}C^{\alpha} - ^{inter}C^{\alpha}|, \Delta(C^{\beta}) = |^{intra}C^{\beta} - ^{inter}C^{\beta}|$.
[d]Invalid $\Delta(C^{\beta})$ because there is no $^{intra}C^{\beta}$, $^{inter}C^{\beta}$ peak in the corresponding spin systems.

bleeding disorder (Chen et al., 1992). The S752P-β3 was dissolved in membrane-mimic DPC micelles and the sample preparation and data collection are the same as those for wild type β3 (Vinogradova et al., 2004). Although S752P-β3 contains only 47 amino acid residues with 2 prolines (44 assignable sites), its HSQC exhibited significant resonance degeneracy due to large portion of unstructured regions and lack of tertiary interactions. Hence, the program combined three datasets, HNCACB, HNCA, CC(CO)NH for the automated assignment. The calculation was performed on three polypeptide segments separated by the two prolines. The assignment took four levels of calculation in which 31, 37, 39, and 41 total assignments were progressively obtained after each level (note previous level assignments were used as the intermediate filters). Three out of 44 sites, which formed a segment of E732–E733–R734, could not be unambiguously assigned because of severe resonance overlap in all bond correlation data, i.e., $^{13}C^{\alpha}/^{13}C^{\beta}/^{13}C^{\gamma}$ and the corresponding proton shifts at $i-1$ positions (all Glu) are identical. With the help of ^{15}N-edited NOE spectra, the R734 was assigned because of a strong NOE from its amide proton to A735 NH. In this case, 42 of 44 sites were unambiguously assigned leaving E732–E733 assignments inter-changeable. The entire process took less than five minutes for the SGI workstation and less than a minute for the PC and the results

were fully consistent with those obtained by the manual protocol.

(3) Assignment on CH_2 domain of ILKBP involved in mediating cell adhesion (Tu et al., 2001). This is a quite challenging test case. Although the CH_2 domain contains only 129 amino acid residues including 6 prolines (~15 kDa), degeneracy was more severe than the last two cases. Moreover, significant amount of $^{13}C^{\beta}$ and side chain signals were missing due to non-specific monomer-dimer exchange and lower temperature (15 °C) experiments (to slow down protein degradation) that broaden the signals. For example, although 120 out of 122 expected spin systems were recognized from HSQC, only 65 out of 115 expected inter-residue $^{13}C^{\beta}$ signals were identified by the combined analysis of HNCACB and CBCA(CO)NH. In addition, two intra-residue $^{13}C^{\beta}$ peaks were absent in HNCACB. In spite of the above problems, by combining three experiments, HNCACB, HNCA and CBCA(CO)NH, 97% assignment was successfully obtained following nine levels of PASA calculations. During each level except the first one, the program automatically uses assignments obtained in previous levels as intermediate filters. At the end of the ninth level, 118 unique assignments out of 122 total expected sites were obtained. Out of four remaining sites, two sites belong to missing NHs (missing entire spin systems), which were identified by the program (breaking points) during the 2nd and 5th level calculations. The other two sites are surrounded by prolines and their assignments had to be confirmed by NOESY at later stages. Because of severe signal loss and degeneracy, the total time for the nine levels of calculation was much longer (~1.5 h on the PC) than Nck2 SH3 domain and S752P-β3. Interestingly, we found that the calculation only took less than 3 min on PC with the same results if we use narrower ranges of the chemical shifts for 20 amino acids (see the numbers in parentheses in Table 1). Clearly, the boundary of the filtering plays an important role in the program efficiency. Although using narrower ranges as defined in table 1 may have some risk of excluding some assignments with unusual chemical shifts, it substantially improves the efficiency and speed of the automated calculations. If users are limited by computational resource and power, narrower chemical shift ranges are recommended for the automated assignment.

(4) Assignment on Malate synthase G (MSG). To further evaluate the program, we performed automated sequential assignment on MSG, the largest protein whose backbone resonances have been recently reported (Tugarinov et al., 2002). The automated assignment of MSG was based on the data reported in the literature with the chemical shifts directly downloaded from the BioMagResBank(BMRB) (Seavey et. al., 1991). The protein contains 723 amino acid residues but only 654 sites were observable (Tugarinov et al., 2002). In other words, except for the first N-terminal amino acid residue and 31 prolines, there are 37 amino acid residues whose NH signals did not show up.

The PASA input data contained the following sets as derived from all reported experimental information: (i) Intra and inter-residue $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shifts for all 654 observable spin systems; (ii) Intra and inter residue C′ chemical shifts. This subset data served as a special pattern filter since it provided an alternative sequential matching to resolve degeneracy derived from $^{13}C^{\alpha}/^{13}C^{\beta}$ matching; (iii) Ala sites and the residues next to each Ala, which were obtained by special pulse sequence experiment; (iv) Val, Leu, Ile sites and residues next to these specific amino acids, which were obtained by selective ^{13}C labeling experiments. The information for (i), (iii), and (iv) was incorporated into the patterns for the observed spin systems. The tolerance for matching intra/inter-residue chemical shifts was as follows: 0.2 ppm for $^{13}C^{\alpha}$, 0.4 ppm for $^{13}C^{\beta}$ and 0.15 ppm for C′. This tolerance was specifically used for MSG by Jung et al. (2004) based on experimental parameters that are different from ours.

The calculation took ~2 h on the PC using the wider chemical shift ranges for 20 amino acids in table 1 and only 30 minutes using narrower ranges of the chemical shifts (see the numbers in parentheses in Table 1). 648 of 654 observable spin systems were uniquely assigned and 28 of 37 breaking points were identified for the wider chemical shift ranges. If using the narrow chemical shift range, the program could assign 649 of 654 spin systems and 31 of 37 breaking points. The assignments were fully consistent with those deposited by Tugarinov et al 2002 in the Protein NMR Databank. When we examined the calculation process, we found that wrong assignment may

be produced when the program pass through unidentified breaking points. However, since PASA adopted a series of conservative strategies, the wrong assignment can be confidently captured during the calculation process. These conservative strategies include (a) cutting tail on C terminus at the assignment results of each polypeptide segment; (b) rejecting assignment result from short polypeptide segments at the early stages of calculation, (c) performing independent segmental assignment on each polypeptide segment and carrying out cross-validation among the results from different polypeptide to detect conflict assignment, and (d) identifying breaking points and performing independent assignments based on patterns. Our results showed that despite the large size of the protein with 37 missing spin systems, PASA was very efficient to minimize the risk of propagating degeneracy- and missing spin system-induced wrong assignments.

At the end of calculation, the un-assigned spin systems had multiple choices and could not be further assigned even if all the assignments were used as intermediate filter. For example, one unassigned Glycine spin system based on its Cα shift could be placed either between P72 and P74 (no NHs) or between E359 and I361 (E359 and I361 have been identified as breaking points by PASA). The E359 and I361 belong to a segment E359-G360-I361 whose entire spin systems were missing as identified by Tugarinov et al. (2002). The unassigned Glycine spin system results in an unidentified breaking point at G360. Most other unassigned spin systems are from isolated sites sandwiched by two breaking points. Further, these spin systems themselves could also belong to breaking points, which explain why we cannot unambiguously locate the sites for some of the missing spin systems. In other words, there could be a short chain whose entire spin systems are all missing, which results in continuous breaking points. Hence, the lack of sufficient linkage information and inadequate amino acid type information prevented further assignment for these remaining residues and identification of the remaining breaking points. In comparison with the result using narrow range of chemical shift for the 20 amino acids, the wide chemical shift range prevents unique assignment on site S537 and identification of breaking points at A2, T431 as well as S638. However, these sites may eventually be assigned with the help of side chain NOEs during structure refinement.

(5) Evaluation of the pattern filtering. To evaluate the power of the pattern filtering, we performed the calculations on the above proteins by turning off all the filters in PASA except for the chemical shift range constraints of Gly, Ala, Ser, and Thr (researchers normally use them to guide the manual assignment process). In the cases of small Nck-2 SH3 domain and S752P-β3 that had good chemical shift dispersion and minimal signal loss, we found that the program could still run, albeit at slower speed (2–3 times slower), to achieve the correct assignments. However, when testing on ILKBP CH2 domain and MSG respectively, the calculations could not run through even for the first level. Combinatorial explosion occurred for both proteins due to the enormous amount of possible assignment choices that were accumulated and amplified during the initial chemical shift matching. For CH2 domain, the combinatorial explosion appears to be mainly due to the missing C^{β} signals (~44%) since the calculation can be completed, albeit much slower, after we simulated the missing C^{β} shifts into the input data. For MSG, the combinatorial explosion is obviously caused by severe degeneracy and 37 missing entire spin systems. These tests demonstrate that the pattern filtering is very powerful for assigning proteins with significant chemical shift degeneracy and missing signals.

Conclusion

Using a new pattern-filtering algorithm, we have developed an automated assignment protocol (PASA) that may have potential to examine large and spectrally complicated proteins. The effectiveness of this program is evident as shown by its successful application on four representative proteins with the size up to 723 residues (see table 4 for the summary of the assignments). It is clear that the per-residue-based pattern filtering is very efficient to minimize the risk of combinatorial explosion induced by degeneracy and missing peaks during the initial global search. We expect this program will be a valuable tool for a variety of NMR applications including mechanism-based structural analyses as well as structural genomics/proteomics.

Table 4. Calculation results of four representative target proteins by PASA

Proteins	Data types	# of residues	# of Pro	Observable spin systems	Assigned Spin systems	Completeness of the assignments
β3 integrin cytoplasmic domain (S752P mutant)	Real data	47	2	44	42	95%
Nck-2 SH3-3 domain	Real data	71	5	64	64	100%
CH_2 domain of ILKBP	Real data	129	6	120	118	98%
Malate synthase G	Data downloaded from BMRB	723	31	654	649	99%

Supplementary material to this paper is available in electronic format at http://dx.doi.org/10.1007/s10858-005-5358-0

Acknowledgements

We thank Sujay Ithychandra, Michael Ford, Kwaku Dayie, and Olga Vinogradova for useful discussions. This study was supported by NIH grants to Jun Qin and Yan Xu, and a UICC American Cancer Society International Fellowship for Beginning Investigators, NSF grant (30371604, 50203001), 973 Project (2002CCA01900) from China to Yizhuang Xu. We are grateful for the kind help and support by Prof. Yan Xu of the Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic Foundation. The program PASA and its manual will be freely available on http://www.lerner.ccf.org/moleccard/qin/.

References

Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.
Baran, M.C., Huang, Y.P.J., Moseley, H.N.B. and Montelione, G.T. (2004) *Chem. Rev.*, **104**, 3541–3555.
Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.
Chen, Y.P., Djaffar, I., Pidard, D., Steiner, B., Cieutat, A.M., Caen, J.P. and Rosa, J.P. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10169–10173.
Coggins, B.E. and Zhou, P. (2003) *J. Biomol. NMR*, **26**, 93–111.
Dötsch, V., Oswald, R.E. and Wagner, G. (1996a) *J. Magn. Reson. B*, **110**, 304–308.
Dötsch, V., Matsu, H. and Wagner, G. (1996b) *J. Magn. Reson. B*, **112**, 95–100.
Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.
Gardner, K.H., Zhang, X., Gehring, K. and Kay, L.E. (1998) *J. Amer. Chem. Soc*, **120**, 11738–11748.
Garrett, D.S., Powers, R., Gronenborn, A.M. and Clore, G.M. (1991) *J. Magn. Reson.*, **95**, 214–220.
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.
Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.
Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.
Hitchens, T.K., Lukin, J.A., Zhan, Y.P., McCallum, S.A. and Rule, G. (2003) *J. Biomol. NMR*, **25**, 1–9.
Hyberts, S.G. and Wagner, G. (2003) *J. Biomol. NMR*, **26**, 335–344.
Jung, Y.S. and Zweckstetter, M. (2004) *J. Biomol. NMR*, **30**, 11–23.
Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.
Lohr, F. and Ruterjans, H. (2002) *J. Magn. Reson.*, **156** (1), 10–18.
Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.
Malmodin, D., Papavoine, C.H.M. and Billeter, M. (2003) *J. Biomol. NMR*, **27**, 69–79.
Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.
Morelle, N., Brutscher, B., Simorre, J-P. and Morelle, M.D. (1995) *J. Biomol. NMR*, **5**, 154–160.
Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Rajesh, S., Heddle, J.G., Kurashima-Ito, K., Nietlispach, D., Shirakawa, M., Tame, J.R. and Ito, Y. (2005) *J. Biol. NMR*, **32**, 177 .
Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (1999) *J. Magn. Reson.*, **141**, 34–43.
Schubert, M., Oschkinat, H. and Schmieder, P. (2001a) *J. Magn. Reson.*, **148**, 61–72.
Schubert, M., Oschkinat, H. and Schmieder, P. (2001b) *J. Magn. Reson.*, **153**, 186–192.
Seavey, B., Farr, E., Westler, W. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.
Slupsky, C.M., Boyko, R.F., Booth, V.K. and Sykes, B.D. (2003) *J. Biomol. NMR*, **27**, 313–21.
Tu, Y., Huang, Y., Zhang, Y., Hua, Y. and Wu, C. (2001) *J. Cell Biol.*, **153** (3), 585–598.
Tugarinov, V., Muhandiram, R., Ayed, A. and Kay, L.E. (2002) *JACS*, **124**, 10025–10035.

Vaynberg, J., Fukuda, T., Chen, K., Vinogradova, O., Velyvis, A., Tu, Y., Ng, L., Wu, C. and Qin, J. (2005) *Mol. Cell*, **17** (4), 513–523.

Velyvis, A., Vaynberg, J., Yang, Y., Vinogradova, O., Zhang, Y., Wu, C. and Qin, J. (2003) *Nat. Struct. Biol.*, **10**, 558–564.

Vinogradova, O., Vaynberg, J., Kong, X.M., Haas, T.A., Plow, E.F. and Qin, J. (2004) *Proc. Natl. Acad. Sci.*, **101**, 4094–4099.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acid.*, Wiley, New York, NY.

Zimmerman, D., Kulikowski, C., Wang, L.L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y.P., Feng, W.Q., Tashiro, M., Shimotakahara, S., Chien, C.Y., Powers, R. and Montelione, G.T. (1997) *JMB*, **269**, 592–610.