J Biomol NMR (2010) 47:85–99
DOI 10.1007/s10858-010-9407-y

ARTICLE

A probabilistic approach for validating protein NMR chemical shift assignments

Bowei Wang · Yunjun Wang · David S. Wishart

Received: 4 March 2010 / Accepted: 5 March 2010 / Published online: 6 May 2010

Abstract It has been estimated that more than 20% of the proteins in the BMRB are improperly referenced and that about 1% of all chemical shift assignments are mis-assigned. These statistics also reflect the likelihood that any newly assigned protein will have shift assignment or shift referencing errors. The relatively high frequency of these errors continues to be a concern for the biomolecular NMR community. While several programs do exist to detect and/or correct chemical shift mis-referencing or chemical shift mis-assignments, most can only do one, or the other. The one program (SHIFTCOR) that is capable of handling both chemical shift mis-referencing and mis-assignments, requires the 3D structure coordinates of the target protein. Given that chemical shift mis-assignments and chemical shift re-referencing issues should ideally be addressed prior to 3D structure determination, there is a clear need to develop a structure-independent approach. Here, we present a new structure-independent protocol, which is based on using residue-specific and secondary structure-specific chemical shift distributions calculated over small (3–6 residue) fragments to identify mis-assigned resonances. The method is also able to identify and re-reference mis-referenced chemical shift assignments. Comparisons against existing re-referencing or mis-assignment detection programs show that the method is as good or superior to existing approaches. The protocol described here has been implemented into a freely available Java program called "Probabilistic Approach for protein Nmr Assignment Validation (PANAV)" and as a web server (http://redpoll.pharmacy.ualberta.ca/PANAV) which can be used to validate and/or correct as well as re-reference assigned protein chemical shifts.

Keywords NMR · Protein chemical shift · Chemical shift assignment · Chemical shift assignment validation · BioMagResBank (BMRB)

B. Wang
Shanghai American School Pudong, 201201 San Jia Gang, Pudong, Shanghai, People's Republic of China

Y. Wang
Mesolight LLC, 4607 W 61st St., Little Rock, AK 72209, USA

D. S. Wishart (✉)
Departments of Computing Science and Biological Sciences, University of Alberta, Edmonton, AL T6G 2E8, Canada
e-mail: david.wishart@ualberta.ca

Introduction

Thanks to continuing improvements in NMR instrumentation, NMR assignment protocols and NMR software, the chemical shift assignment of proteins and peptides has become progressively faster and easier. Indeed, the BMRB now contains more than 3,500,000 protein chemical shifts from nearly 5,700 fully assigned polypeptides (Seavey et al. 1991; Ulrich et al. 2008). Having such a collection of complete or near-complete spectral assignments is giving biomolecular NMR spectroscopists an unprecedented opportunity to compare, explore and decipher the rich structural and dynamic information encoded protein chemical shifts. Indeed, it has been through the BMRB and related chemical shift databases that the correlation between protein secondary structure and chemical shifts was first identified (Wishart et al. 1991; Spera and Bax 1991; de Dios et al. 1993). Likewise, chemical shift databases like the BMRB have helped facilitate the development of better protocols for shift-based secondary structure identification (Wishart et al. 1992, 1994; Metzler et al. 1993; Gronenborn and Clore 1994), the detailed exploration of nearest-neighbor effects

on protein backbone chemical shifts (Wang and Jardetzky 2002a) as well as the development of new and improved methods to predict ^{1}H, ^{13}C and ^{15}N chemical shifts from 3D structure coordinates (Xu and Case 2001; Neal et al. 2003; Wang and Jardetzky 2004; Shen et al. 2008; Kohlhoff et al. 2009).

However, chemical shifts are prone to numerous kinds of reporting and measurement errors. The problem with chemical shift measurement is particularly acute in biomolecular NMR where up to 20% of all chemical shift assignments are improperly referenced and where up to 40% of all proteins have at least one assignment error (Zhang et al. 2003). This situation is certainly understandable given the large number of chemical shifts that must be measured (hundreds to thousands), the variety of chemical shift types (^{1}H, ^{13}C, ^{15}N, ^{31}P), and the incredible range of solvent conditions (pH, temperature, salts, organic solvent mixtures) that experimentalists must use. As a result, protein chemical shifts are perhaps the most precisely measurable but the least accurately measured parameters in NMR spectroscopy. Since the structural and dynamic information contained in chemical shifts is subtle, inaccurate or incorrectly referenced chemical shift measurements and assignment errors can easily blur or distort an exquisitely detailed picture of a biomolecule.

Over the past several years a number of programs have been described to help identify and correct improperly referenced or mis-assigned protein chemical shifts (see Table 1 for a summary and comparison). For instance, the BMRB uses its own programs as well as the Assignment Validation Software Suite (AVS; http://psvs-1_3.nesg.org/htdocs/avs.html) developed by Moseley et al. (2004) to identify potential mis-assignments. These programs compare the BMRB-derived distribution of diamagnetic ^{1}H, ^{13}C and ^{15}N shifts for all 20 amino acid residues to the assigned amino acid shifts submitted by depositors. Chemical shifts that are more than 5 standard deviations away from the residue-specific mean values are flagged and reported as possible assignment errors. While these methods can identify significant outliers or potential typographical errors, they are not able to identify chemical shift referencing errors, nor can they accommodate "differently referenced" chemical shift assignments. To get around this problem Zhang et al. (2003) used the chemical shift calculation program ShiftX (Neal et al. 2003) to predict ^{1}H, ^{13}C and ^{15}N shifts based on the 3D structure coordinates of the assigned protein. By comparing the predicted shifts to the observed shifts they were able to accurately identify chemical shift reference offsets as well as potential mis-assignments. They also produced a continuously updated database of re-referenced chemical shift assignments called RefDB—which now has 2108 re-referenced protein assignments. However, this "post hoc" approach to re-referencing and re-assignment requires that a 3D structure for the target protein be available to assess the correctness of assigned chemical shifts. Given that chemical shift assignments are typically made before the structure is determined, new "ad hoc" approaches needed to be developed. As a result, several new methods were developed which make use of the estimated (via ^{1}H shifts) or predicted (via sequence) secondary structure content of the target protein (Wang and Wishart 2005; Wang et al. 2005; Ginzinger et al. 2007; Wang and Markley 2009). These programs have all been shown to accurately identify mis-referenced and properly re-reference protein chemical shifts deposited in the BMRB.

However, these particular NMR re-referencing programs do not identify mis-assigned chemical shifts. Given that nearly 40% of protein entries deposited in the BMRB appear to have at least one assignment error (Zhang et al. 2003), there is a clear need to develop more robust "ad hoc" or sequence-based software to detect assignment errors (as well as referencing errors) prior to structure determination. Here, we present a new sequence-based protocol that is able to detect assignment errors and correct chemical shift referencing errors. The method is based on using the residue-specific and secondary structure-specific chemical shift distributions originally calculated by Wang

Table 1 Summary and comparison of different chemical shift re-referencing and mis-assignment detection programs

Program	Detects or performs shift re-referencing	Detects gross assignment errors	Detects subtle assignment errors	Flags and suggests correct shift assignments	Distinguishes assignment errors from referencing errors	Requires 3D structure
CheckShift (Ginzinger et al. 2007)	Yes	No	No	No	No	No
AVS (Moseley et al. 2004)	No	Yes	No	No	No	No
LACS (Wang and Markley 2009)	Yes	No	No	No	No	No
PSSI (Wang and Wishart 2005)	Yes	No	No	No	No	No
ShiftCor (Zhang et al. 2003)	Yes	Yes	Sometimes	No	Yes	Yes
PANAV (This paper)	Yes	Yes	Yes	Yes	Yes	No

and Jardetzky (2002b). By comparing the joint, residue-specific chemical shift probabilities from a group of resonances corresponding to a segment of sequentially connected residues against the observed chemical shifts for the same residue segment, it is possible to identify (and correct) chemical shift referencing errors as well as chemical shift assignment errors. To validate this approach we assessed its performance using a variety of real and artificial test data sets. We also compared this method's re-referencing capabilities to CheckShift (Ginzinger et al. 2007) LACS (Wang and Markley 2009) and it's mis-assignment detection capabilities to the AVS suite (Moseley et al. 2004). Overall, the program performed as well or better than these existing programs. The protocol described here has been implemented into a freely available java program called PANAV (Probabilistic Approach for protein Nmr Assignment Validation), which is available for download at http://redpoll.pharmacy.ualberta.ca. It is also available as a web server (http://redpoll.pharmacy.ualberta.ca/PANAV). PANAV can be used to re-reference as well as validate and correct assigned protein chemical shifts.

Methods

The basic concept behind PANAV lies in the fact that each of the 6 types of backbone atoms (N, CA, CB, CO, HA and HN) for each of the 21 amino acid types (including cystine and cysteine) within each of the 3 major secondary structure classes (helix, coil and sheet) have a reasonably unique chemical shift distribution. These Gaussian-like distribution functions have been previously calculated by several groups (Wang and Jardetzky 2002b; Zhang et al. 2003). Combining (i.e. taking the joint probability of) these chemical shift distribution functions over all the backbone atoms within a residue as well as within clusters of sequentially connected residues would be expected to create a unique chemical shift signature for a given multi-residue segment. Comparing this predicted chemical shift signature (over a group of 3 residues, say) with the observed chemical shifts over the same 3 residues from a newly assigned protein allows one to identify potential outliers and to assign a probability that they are outliers. In some respects the PANAV algorithm mimics the earliest stages of the manual assignment process where spectroscopists use their knowledge of residue-specific chemical shifts, inter-residue spin connectivities and the clustered nature of protein secondary structures to make educated guesses about which shifts belong to which residues. The theory behind PANAV is explained in more detail below.

For an observed chemical shift δ_n, where n corresponds to N, CA, CO, CB, HN, or HA, the probability that it "belongs to" or is correctly assigned to one of the 21 amino acids (i) can be evaluated from a Gaussian distribution:

$$G_i(\{\delta_n\}) = \frac{1}{\sqrt{2\pi}\sigma_{n,i}} \exp\left[-\frac{\left(\delta_n - \overline{\delta_{n,i}}\right)^2}{2\sigma_{n,i}^2}\right]$$

where, $\overline{\delta_{n,i}}$ is the average chemical shift and $\sigma_{n,i}$ is the standard deviation for each of the 21 amino acids. For a given set of resonances $\{\delta_n\}$ ($n = \delta_N, \delta_{CA}, \delta_{CO}, \delta_{CB}, \delta_{HN}, \delta_{HA}$) belonging to an amino acid spin system, its probability of belonging to or being correctly assigned to one of the 21 amino acids is given as the product of the individual chemical shift probabilities:

$$p_i(\{\delta_n\}) = \prod_n G_i(\delta_n).$$

In calculating this probability the averaged ^{1}H, ^{15}N and ^{13}C (^{13}CO, ^{13}Cα, ^{13}Cβ) chemical shifts and standard deviations for the 21 amino acids previously reported by Wang and Jardetzky (2002b) are used in the calculation. The program calculates the p_i based on the chemical shifts for all three secondary structure types, and takes the one with the highest probability as being the assigned spin system $\{\delta_n\}$ for amino acid i.

For a group of resonances or amino acid spin systems $\{\delta_n\}_1\{\delta_n\}_2 \ldots \{\delta_n\}_m$ corresponding to a segment containing m sequentially connected residues, the relative probability $p(\{\delta_n\}_m)$ of this segment belonging to this segment or being correctly assigned to this segment can be calculated by:

$$p(\{\delta_n\}_m) = \prod_m p_m(\{\delta_n\}).$$

From these joint probability formulas for individual chemical shifts, spin systems (i.e. residues) and sequential spin systems (i.e. residue fragments), it is possible to determine the likelihood that a chemical shift, an entire residue or even a collection of residues has been properly assigned. For a protein with assigned backbone chemical shifts, PANAV selects fragments containing 3, 4, 5, or 6 residues and successively samples these fragments over the entire length of the protein (i.e. by moving 1 residue at a time from the N-terminus to the C-terminus). The HN, HA, CA, CB, CO and N chemical shifts for each testing fragment are used as the input shifts to calculate the probability that each atom, residue and/or residue fragment is correctly assigned. The calculated probabilities for each fragment are normalized so that the highest value is set to 1.0. An assignment is identified as questionable if the calculated highest probability is either: (1) not at the originally assigned position or (2) larger than the probability at the originally assigned position plus a tolerance (set to 0.1 in this study).

Probabilistic Approach for protein Nmr Assignment Validation (PANAV) accepts both BMRB and SHIFTY

formatted files as input and produces several kinds of output, including suggested chemical shift reference offsets, flagged mis-assignments, percentages of potential assignment errors and a global assignment quality score. Specifically PANAV identifies if the protein needs chemical shift re-referencing and if so, it provides the chemical shift offsets for each nucleus (CA, CB, CO and N). Additionally the regions where potentially incorrect assignments have been made are marked in red and written to a text file called "Potential Mis-assignments". Potential typographical errors (see below) are marked with a "D" in the same text file. The number of mis-assignments and typographical errors (deviant shifts) for the query protein are also calculated. Finally, an overall "Confirmed Assignment" score (called the CONA score) for the entire protein is given. CONA is defined as the number of confirmed assigned fragments divided by total number of selected fragments.

$$\text{CONA} = \frac{\#\text{ of selected fragments} - \#\text{ non confirmed fragments}}{\#\text{ of selected fragments}} \times 100$$

Global CONA scores are calculated over 3, 4, 5 and 6 residue fragment lengths, meaning that up to four global CONA scores can be calculated for a given protein. Prior to the calculation of the CONA score(s), the PANAV program automatically calibrates mis-referenced chemical shifts using a modification on a protocol previously developed by us (Wang and Wishart 2005). The calculation or correction of mis-referenced chemical shifts is obviously critical to the identification of any mis-assigned shifts. In the present study, the cut-offs for identifying and automatically recalibrating mis-referenced chemical shifts were set to 1.0 ppm for CA, CB, and CO; and 1.5 ppm for N chemical shifts, respectively. In other words, if the PANAV-calculated chemical shift offset exceeds these cutoff values, the chemical shifts are re-referenced accordingly.

To reduce errors in the re-referencing process and to identify probable typographical errors, "deviant" chemical shifts are automatically flagged and excluded in the re-referencing calculation. "Deviant" chemical shifts are defined as those that deviate significantly from the majority of assigned shifts (by 6 standard deviations) and "suspicious" chemical shifts are defined as those that differ from the reference-corrected standard values by more than 4 standard deviations. Deviant shifts are most often due to typographical errors (missing or added digits or missing decimals), switched assignments between "similar" nuclei (e.g. N and NH, CB and CA etc.), or other common text entry errors. Given the fact that some BMRB entries have the reference offsets above 40 ppm (e.g. BMRB entries 4,150 and 5,179), it is important to distinguish the deviant/suspicious assignments from mis-referenced assignments. PANAV uses the following strategy to distinguish deviant assignments from mis-referenced assignments. First, PANAV uses as set of "standard" nucleus-specific chemical shifts—175.7, 56.6, 34.4, 119.3, 7.93, and 4.41 ppm for CO, CA, CB, N, H, and HA shifts calculated by averaging the random coil chemical shifts over all 21 amino acids for each nucleus using a database of chemical shifts (Wang and Jardetzky 2002b). Second, for a given BMRB entry, PANAV calculates the average CO, CA, CB, N, H, and HA chemical shifts over the entire protein sequence. Both averaging processes exclude the Gly's CA shifts, as well as the CB shifts of Ala, Ser and Thr, since these shifts differ significantly from all other amino acids. The difference between the two sets of data are then used to "crudely" estimate the reference offsets ($^{\text{off}}\delta_n$) and subsequently to identify deviant assignments. For a given $\delta_{n,i}$ (where n corresponds to N, CA, CO, CB, HN, or HA, and i corresponds to each of the 21 amino acids), if $\delta_{n,i} + {}^{\text{off}}\delta_n$ is greater than 6 standard deviations from the values reported by Wang and Jardetzky (2002b), it is flagged as a deviant assignment. All deviant assignments identified in this step are excluded from the reference offset calculation process. After removing the deviant assignments, PANAV calculates the reference offsets more precisely using the protocol described by Wang and Wishart (2005). PANAV only performs a reference calibration if the calculated offsets are larger than the cutoff values (1.0 ppm for CO, CA, CB; and 1.5 ppm for N shifts). PANAV then conducts a 3-, 4-, 5-, and 6- residue scan to further validate those assignments. PANAV also detects and flags suspicious assignments—those assignments that are more than 4 standard deviations from the expected value. Figure 1 outlines the overall assignment checking and re-referencing process used in PANAV while Fig. 2 provides an example of the output from the PANAV program. Because CA and CB chemical shifts tend to have the greatest impact on whether or not a fragment is ultimately "confirmed", only fragments with 50% or more of assigned CA and CB chemical shifts can be taken into consideration during the PANAV scoring process.

To help clarify some of the definitions concerning deviant, suspicious, mis-assigned and mis-referenced chemical shifts we have generated an example of a hypothetical 25-residue peptide with a variety of deviant, suspicious, mis-assigned and mis-referenced chemical shift values along with the corrections that PANAV is capable of making. This example is shown in Fig. 3.

To assess the performance of PANAV in terms of both its ability to detect/correct chemical shift mis-referencing and chemical shift assignments we assembled several real and synthetic test data sets. All chemical shift assignment data were taken from the BioMagResBank (BMRB). The synthetic test data for assessing reference offset corrections

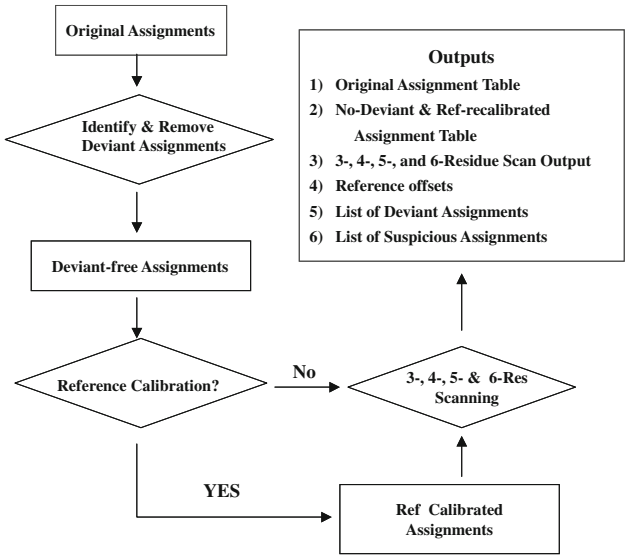

Fig. 1 An outline of the protocol described in this study for assignment validation process. The reference offsets are determined using the protocol previously described by Wang and Wishart (2005)

was assembled from 9 proteins (see Table 2) corresponding to 1,229 residues. Each of the proteins was given artificial reference offsets by adding/subtracting 0.5 ppm, 1.0 ppm, 2.5 ppm, 5.0 ppm and 10 ppm to all the ^{13}C (CA, CB, and CO) and ^{15}N shifts. PANAV's performance was compared against two other chemical shift re-referencing programs: CheckShift (Ginzinger et al. 2007) and LACS (Wang and Markley 2009).

The synthetic test data for identifying chemical shift mis-assignments was assembled from 2 entries (bmr15268 and bmr6212) which had: 1) no suspicious or mis-assigned ^{13}C shifts as assessed by AVS, 2) no ^{13}C chemical shift with a difference (observed-predicted) greater than 5.0 ppm as detected by SHIFTCOR (Zhang et al. 2003), and 3) properly referenced ^{13}C and ^{15}N shifts as shown by SHIFTCOR and CheckShift. From these two "perfect" entries, we created five synthetically mis-assigned entries by randomly switching similar chemical shift assignments with each other. The performance of PANAV in detecting mis-assignments was then compared against two other shift error detection programs: SHIFTCOR (Zhang et al. 2003) and AVS (Moseley et al. 2004). In addition to these synthetic data sets, several specific examples from real data sets are presented, compared and discussed.

Results and discussion

Detection of chemical shift reference offsets

Probabilistic Approach for protein Nmr Assignment Validation (PANAV) uses a modification of the method first described by Wang and Wishart (2005), for chemical shift re-referencing. There are at least two other methods—CheckShift (http://checkshift.services.came.sbg.ac.at) (Ginzinger et al. 2007), and LACS (Wang and Markley 2009), that can also be used to calibrate mis-referenced the protein chemical shifts without the use of a 3D structure. To quantitatively evaluate the performance of these programs in detecting and suggesting chemical shift reference offsets, we selected 9 proteins from the BMRB that are predicted to have relatively small (within 0.05, 0.10, 0.10, and 0.3 ppm for CA, CB, CO, and N shifts, respectively) reference offsets as calculated by SHIFTCOR (i.e. by comparison to their 3D structure-derived shifts) and as indicated by their adherence to IUPAC chemical shift referencing protocols in their BRMB data file. Each of these entries was then deliberately mis-referenced by adding 0.5, 1.0, 2.5, 5.0 or 10 ppm to all the ^{13}C (CA, CB, and CO) and ^{15}N shifts. Therefore a total of $9 \times 5 = 45$ synthetically mis-referenced protein datasets were created. Table 2 summarizes the average reference offsets and standard deviations of these mis-referenced entries as detected by PANAV, LACS, and CheckShift.

As seen from this table all three programs predict slightly different reference offsets for the original (uncorrected) data. Also note that both LACS and CheckShift are programmed to always predict the same offset for CA and CB shifts, whereas PANAV does not make this assumption. Using the reference offset calculated from the original data (the column listed under 0.0 ppm) as a "correction" it is clear that all three programs are able to calculate the synthetic offsets to within 0.02 ppm of their actual value. This is well within the digital precision with which ^{13}C and ^{15}N chemical shifts can be measured. The only minor distinction between the programs is that PANAV typically has a smaller spread (or standard deviation) in the offsets it calculates, indicating that it is slightly more consistent or more precise than either LACS or CheckShift. However, as it will be shown later, we also found that LACS and CheckShift are not able to handle BMRB entries that have the extremely large (above 40 ppm) reference offsets.

Given that PANAV (as well as LACS and CheckShift) appear to be quite robust in finding problems with chemical shift mis-referencing, we decided to run PANAV against the entire BMRB to assess the general status of chemical shift referencing in protein NMR. More specifically, PANAV was run on the 2,421 entries in the BMRB that had sequence lengths >50 amino acids and a sufficient proportion of (>80%) of assigned chemical shifts to perform a robust chemical shift reference correction. In total we identified 243 entries where the CA shifts were offset by more than 1.0 ppm, 238 entries where the CB shifts were offset by more than 1.0 ppm, 200 entries where the CO shifts were offset by more than 1.0 ppm and 137 entries

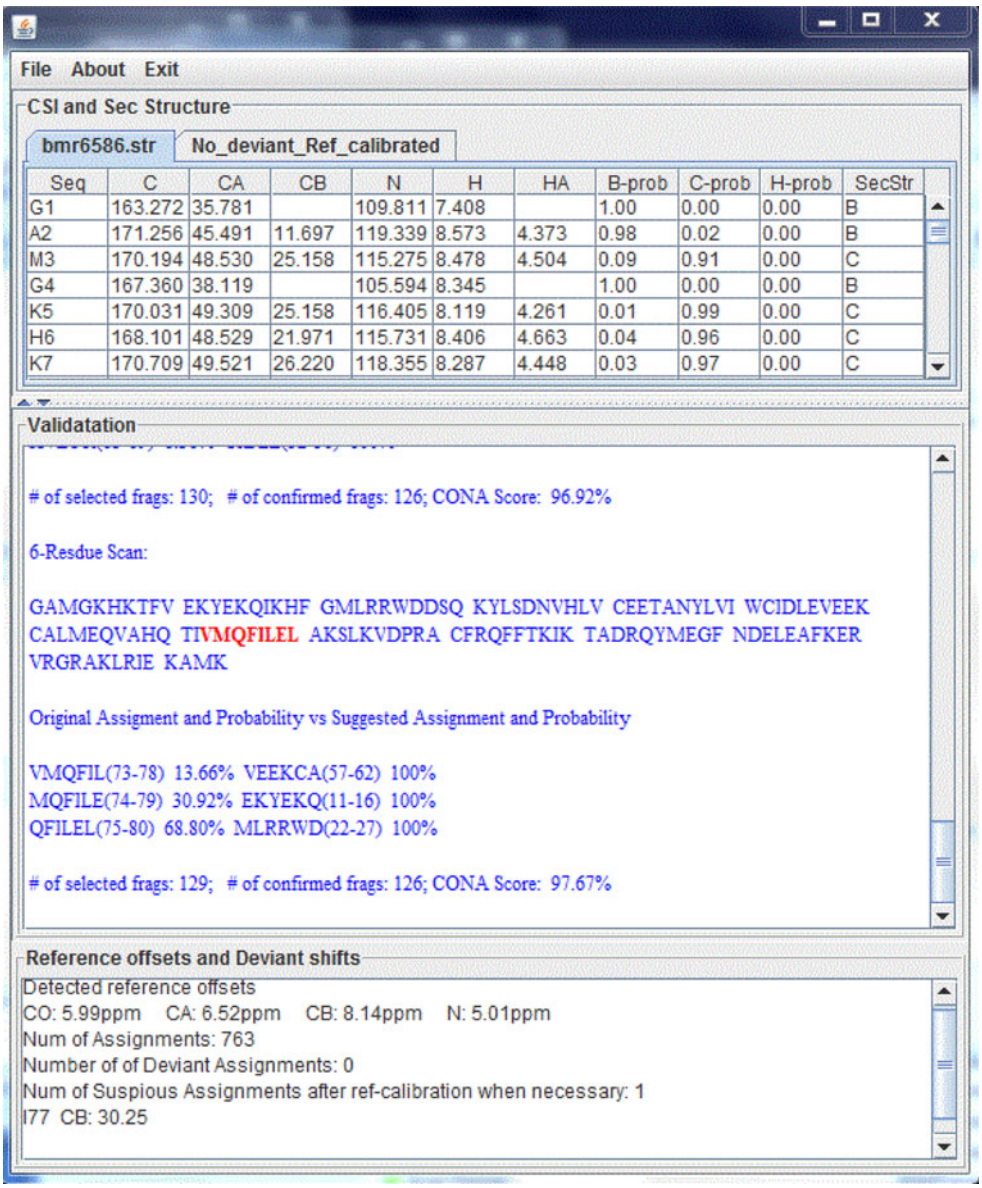

Fig. 2 A screenshot example of PANAV's output for a BMRB entry (bmr6586) with a number of chemical shift referencing problems and suspicious assignments. The *top panel* in the figure contains a table of the protein's original chemical shifts. The *central panel* highlights suspicious or mis-assigned residues (in *red*) and indicates the probability that these residues are assigned correctly (before and after re-assignment). The *lower panel* shows the N and C reference *offsets*, as well as the deviant, and suspicious assignments as detected by PANAV

where the ^{15}N shifts were offset by more than 1.5 ppm. Overall, 19.67% of the entries in the BMRB appear to be mis-referenced (i.e. they required at least one chemical shift reference offset correction). The list of proteins requiring reference offset corrections and the extent of the offsets is available at the PANAV website. This proportion ($\sim$20%) is roughly the same as has been reported by Zhang et al. 2003. Evidently, chemical shift referencing is still a significant problem for the biomolecular NMR community.

Chemical shift assignment validation

To assess PANAV's assignment validation capabilities we performed 4 separate tests. The first test involved assessing the total number of proteins in the BMRB with potential mis-assignments and looking at how PANAV performed in identifying serious or gross mis-assignments. The second test involved identifying mis-assignments in a collection of deliberately, but subtly mis-assigned proteins and comparing the performance of PANAV to AVS and SHIFTCOR. The third test involved looking at pairs of identical proteins in the BMRB where deposited assignments disagree with each other and using PANAV to identify which ones are likely mis-assigned. The fourth test involved looking at the long-term trends in assignment correctness or assignment quality based on the year of deposition in the BMRB.

At the time of this writing, there were 5,649 protein entries available in the BMRB. Among them, 2,421 entries were selected for further processing based on the fact that they had sequence lengths >50 amino acids and that they

A

NUM	RES	Strc	HA	CA	CB	CO	N^{ref}	NH
1	A	C	4.46	53.2	18.9	177.0	**173.5**	8.67
2	G	H	3.77	47.5	----	175.2	**156.1**	8.43
3	**L^{mis}**	H	**4.54**	**54.1**	**43.2**	**174.0**	**174.5**	**9.71**
4	E	H	**41.1**^{D}	59.2	29.4	179.1	**169.0**	7.83
5	M	H	4.06	57.8	32.3	178.9	**168.1**	6.99
6	A	H	3.93	54.7	18.5	180.3	**171.4**	7.31
7	K	H	3.29	58.9	32.1	178.4	**169.2**	**4.44**^{S}
8	S	H	4.04	61.3	63.1	176.0	**165.5**	7.97
9	L	H	3.85	57.5	41.9	178.6	**169.6**	8.12
10	A	H	4.02	55.2	18.2	179.7	**171.3**	8.65
11	G	C	3.98	45.5	**19.3**^{D}	175.5	**159.1**	8.22
12	P	C	4.41	63.4	31.9	177.1	**----**	----
13	D	C	4.68	54.2	40.7	176.8	**169.6**	8.46
14	I	B	4.51	**6.01**^{D}	39.8	174.2	**172.8**	9.12
15	V	B	5.13	60.8	33.7	174.8	**171.9**	9.83
16	I	B	4.93	59.9	40.1	174.9	**173.0**	8.97
17	F	B	4.88	56.7	41.6	173.3	**177.3**	**84.4**^{D}
18	G	C	3.96	44.6	----	174.1	**158.7**	8.01
19	D	C	4.59	55.2	40.9	176.7	**169.9**	8.28
20	W	B	4.81	56.4	31.5	175.4	**172.2**	8.97
21	**L^{mis}**	B	**3.91**	**58.1**	**41.6**	**178.2**	**169.6**	**7.59**
22	I	B	4.92	60.2	39.9	174.0	**172.8**	8.98
23	T	B	4.53	61.1	70.8	174.0	**166.7**	9.39
24	V	B	4.29	60.9	33.6	174.0	**171.3**	8.72
25	G	C	3.99	45.3	----	175.3	**159.4**	8.10

B

NUM	RES	Strc	HA	CA	CB	CO	N	NH
1	A	C	4.46	53.2	18.9	177.0	123.5	8.67
2	G	H	3.77	47.5	----	175.2	106.1	8.43
3	L	H	3.91	58.1	41.6	178.2	119.6	7.59
4	E	H	4.11	59.2	29.4	179.1	119.0	7.83
5	M	H	4.06	57.8	32.3	178.9	118.1	6.99
6	A	H	3.93	54.7	18.5	180.3	121.4	7.31
7	K	H	3.29	58.9	32.1	178.4	119.2	4.44
8	S	H	4.04	61.3	63.1	176.0	115.5	7.97
9	L	H	3.85	57.5	41.9	178.6	119.6	8.12
10	A	H	4.02	55.2	18.2	179.7	121.3	8.65
11	G	C	3.98	45.5	----	175.5	109.1	8.22
12	P	C	4.41	63.4	31.9	177.1	----	----
13	D	C	4.68	54.2	40.7	176.8	119.6	8.46
14	I	B	4.51	60.1	39.8	174.2	122.8	9.12
15	V	B	5.13	60.8	33.7	174.8	121.9	9.83
16	I	B	4.93	59.9	40.1	174.9	123.0	8.97
17	F	B	4.88	56.7	41.6	173.3	123.3	8.44
18	G	C	3.96	44.6	----	174.1	108.7	8.01
19	D	C	4.59	55.2	40.9	176.7	119.9	8.28
20	W	B	4.81	56.4	31.5	175.4	122.2	8.97
21	L	B	4.54	54.1	43.2	174.0	124.5	9.71
22	I	B	4.92	60.2	39.9	174.0	122.8	8.98
23	T	B	4.53	61.1	70.8	174.0	116.7	9.39
24	V	B	4.29	60.9	33.6	174.0	121.3	8.72
25	G	C	3.99	45.3	----	175.3	109.4	8.10

Fig. 3 A hypothetical peptide with complete backbone assignments. Shown in *bold* for part A) are examples of mis-assignments or misplaced decimal places (*superscript* "mis"), deviant assignments (*superscript* "D"), suspicious assignments (*superscript* "S") and chemical shift referencing errors (*superscript* "ref"). In particular, the assignments for Leu3 have been erroneously swapped with Leu21 (mis-assignment). Likewise the HA shift for Glu4, the CA shift for Ile14, the CB shift for Gly11 and the NH shift for Phe17 have typographical errors (deviant assignments), the NH shift for Lys7 is flagged as suspicious, while the chemical shifts for all N nuclei have been mis-referenced by 50 ppm. The corrected assignments for this peptide are shown below in B)

had a sufficient proportion (>80%) of assigned CA and CB chemical shifts. CA and CB shifts typically exhibit the greatest residue and structure-specific dispersion among all 6 backbone nuclei and so they are particularly important to the optimal performance of PANAV. They also have the greatest effect on the CONA score(s).

In the first test used to assess PANAV's performance we ran the program on all 2,421 BMRB entries using 3-, 4-, 5-, and 6-residue fragment scans, respectively, with the reference offset correction set to "on". As noted previously, PANAV's reference offset correction is always run prior to running its CONA evaluation. Using 1.0 ppm for CO, CB, and CA nuclei and 1.5 ppm for N nuclei as the cut-off re-referencing values, 20% or 476 out of the 2,421 selected BMRB entries, were found to need chemical shift re-referencing. In addition, we found that 594 proteins had at least 1 deviant shift (i.e. obvious typographical errors), with 25 proteins having deviant CO shifts, 120 proteins having deviant CA shifts, 310 proteins having deviant CB shifts, 87 proteins having deviant N shifts, 65 proteins having deviant HN shifts, and 151 proteins having deviant HA shifts. In other words, 24.5% of the proteins in the BMRB appear to have at least one gross chemical shift assignment error. In addition to these obvious assignment errors, it appears that more than 330 proteins exhibited global CONA scores below 98% (using a 6-residue fragment scan), which suggests that 14% of proteins in the BMRB have somewhat more subtle mis-assignments. Table 3 lists the 80 BMRB entries that had the lowest global CONA scores (below 95% using a 6 residue fragment scan) together with the identified number of "deviant" and "suspicious" assignments.

Examination of the entries listed in Table 3 revealed that most of these proteins had significantly more "deviant" and "suspicious" chemical shifts (i.e. typographical errors) than those with high CONA scores. For instance, the average percentage of the deviant and suspicious assignments (relative to the total number of assignments) for these 80 proteins is 0.60%. In comparison, the percentage of deviant and suspicious assignments for the entire set of 2,420 BMRB entries that we selected is just 0.36%. Since the deviant chemical shifts are excluded in the calculation of CONA scores, the low CONA scores for these entries indicate they probably have other assignment problems that cannot be as readily identified by simple visual inspection. This observation that low CONA scores and high numbers of deviant/suspicious shifts seem to be correlated suggests that sloppy text entry is likely a good indicator of sloppy assignment practices. Surprisingly, many relatively small proteins (which presumably would have better quality NMR spectra) exhibited problems with both deviant shifts and low CONA scores. In contrast, we examined the three largest proteins in the BMRB: bmr6416, bmr10053, and bmr5471, which have 501, 517, and 731 assigned residues, respectively. Despite their large size, these three entries exhibited exceptionally high global CONA scores (>98% for the 6-residue fragment test).

Obviously the identification of gross mis-assignments is a relatively simple task that can often be done by simply comparing observed chemical shifts to standard residue shift tables. These residue-specific shift tables have been published elsewhere (Wishart et al. 1991; Wishart and Nip

Table 2 Outputs* from PANAV, LACS, and CheckShift for selected BMRB entries with artificial reference offsets

Artificial offsets	0.0 ppm	−0.5 ppm	−1.0 ppm	−1.5 ppm	−2.5 ppm	−5.0 ppm	−10.0 ppm
PANAV output							
CO	0.16 (0.27)	−0.34 (0.28)	−0.84 (0.28)	−1.34 (0.28)	−2.32 (0.28)	−4.83 (0.28)	−9.82 (0.29)
CA	0.03 (0.11)	−0.47 (0.11)	−0.97 (0.11)	−1.47 (0.11)	−2.46 (0.13)	−4.96 (0.12)	−9.96 (0.13)
CB	−0.02 (0.16)	−0.51 (0.16)	−1.01 (0.16)	−1.51 (0.16)	−2.53 (0.17)	−5.00 (0.17)	−10.05 (0.19)
N	0.15 (0.53)	−0.34 (0.53)	−0.85 (0.53)	−1.35 (0.53)	−2.36 (0.53)	−4.86 (0.53)	−9.84 (0.53)
LACS output							
CO	0.27 (0.44)	−0.23 (0.44)	−0.73 (0.44)	−1.23 (0.44)	−2.23 (0.44)	−4.73 (0.44)	−9.73 (0.44)
CA	−0.08 (0.13)	−0.58 (0.13)	−1.08 (0.13)	−1.58 (0.13)	−2.58 (0.13)	−5.08 (0.13)	−10.08 (0.13)
CB	−0.08 (0.13)	−0.58 (0.13)	−1.08 (0.13)	−1.58 (0.13)	−2.58 (0.13)	−5.08 (0.13)	−10.08 (0.13)
N	−0.17 (0.53)	−0.67 (0.53)	−1.17 (0.53)	−1.67 (0.53)	−2.67 (0.53)	−5.17 (0.53)	−10.09 (0.41)
CheckShift output							
CO	0.37 (0.28)	−0.13 (0.28)	−0.63 (0.28)	−1.13 (0.28)	−2.13 (0.28)	−4.63 (0.28)	−9.63 (0.28)
CA	0.28 (0.28)	−0.22 (0.28)	−0.72 (0.28)	−1.22 (0.28)	−2.22 (0.28)	−4.72 (0.28)	−9.74 (0.29)
CB	−0.22 (0.14)	−0.72 (0.14)	−1.22 (0.14)	−1.72 (0.14)	−2.72 (0.14)	−5.22 (0.14)	−10.22 (0.14)
N	−0.22 (0.80)	−0.72 (0.81)	−1.22 (0.81)	−1.72 (0.80)	−2.72 (0.80)	−5.22 (0.80)	−10.22 (0.81)

* The average value and standard deviation (in brackets) for BMRB entries 10138, 10139, 15249, 16006, 4094, 5516, 6071, 7055, and 7086

1998; Zhang et al. 2003) and several are maintained at the BMRB. This comparison is essentially what is done using the AVS software (Moseley et al. 2004). A much more challenging task is to identify subtle mis-assignments, where a 2 or 3 residue fragment with a similar sequence has been accidentally swapped with another 2 or 3 residue fragment. These types of mis-assignments can and have been detected via SHIFTCOR (Zhang et al. 2003), but as mentioned before, SHIFTCOR requires a 3D structure to detect these problems.

To test how PANAV performs and how it compares to AVS and SHIFTCOR in detecting subtle mis-assignments, we chose to analyze two protein BMRB entries: the apolipoprotein C-III protein (bmr15268) and F6 subunit from ATP synthase (bmr6212), which appear to have near "perfect" assignments (i.e. no apparent mis-assignments or referencing problems as measured by multiple programs). From these two "perfect" real entries, we created five artificially mis-assigned entries by switching similar values in chemical shifts with each other. These mis-assigned entries are labeled in a similar manner to the way protein mutations are labeled. For instance bmr15268_X#_Y# indicates that entry bmr15268 has had the assignments for residue #X swapped with the assignments of residue #Y. PANAV, AVS, and SHIFTCOR were then used to assess these artificially mis-assigned proteins to see if they could detect the mis-assignments. Table 4 shows the results.

Inspection of this table shows that PANAV is able to detect mis-assignments much more consistently than either SHIFTCOR or AVS. For instance, PANAV easily detects the mis-assigned entry bmr15268_E4_K24 whereas AVS and SHIFTCOR do not. Bmr15268_E4_K24 was generated by switching the CA and CB chemical shifts for E4 and K24 in the original bmr15268 entry. These two amino acids were selected because of their very similar CA and CB shifts (differences of only 3.0 and 3.2 ppm, respectively) and the close resemblance between the 4-residue fragments they were selected from AEDA (3-6) and AKDA (23-26). PANAV not only calculates a lower CONA score for 3- and 4- residue scans in response to the switch, but also successfully detects that AKDA (23-26) should be switched to the 3–6 position (Fig. 4a). Looking at the other artificially mis-assigned entries, including bmr6212_QKL(8-10)_REY(16-18), and bmr6212_EYQQ(34-37)_KLKQ(45-48), we once again see that PANAV detects these mis-assignments while AVS and SHIFTCOR do not.

Interestingly, PANAV's 3-residue scan suggests 4 possible corrections to make to bmr6212_QKL(8-10)_REY(16-18) (Fig. 4b). Although 3 of the suggested corrections are not valid, they do provide valuable information to the user. In particular, these four alternate suggestions involve amino acids in the area where the artificial switch was made, indicating that these suggestions are not without reason and should serve to draw attention to this region. For instance, the first suggestion points out that VVE(70-72) and KFE(67-69) are in fact strikingly similar to VQK(7-9) and EYR(17-19) respectively, and thus deserve a recheck of whether or not an error was possibly made in the original assignment. PANAV's 4-residue scan suggests 2 possible corrections to make to bmr6212_EYQQ(34-37)_KLKQ(45-48)—one of these suggested corrections is "valid" (Fig. 4c).

The last entry in Table 5, bmr6212_QKLF(8-12)_REYR(16-19) was the only entry in which both SHIFTCOR

Table 3 List of BMRB entries with low CONA scores (after reference calibration when necessary)

BMRB	6-Residue CONA (%)	Total Assignments*	Deviant Assignmts	Suspicious Assignmts	BMRB	6-Residue CONA (%)	Total Assignments*	Deviant Assignmts	Suspicious Assignmts
16221	70.21	251	4	3	6242	92.71	550	2	3
6299	77.27	360	1	0	6713	92.71	440	1	0
6565	79.14	931	1	7	6549	92.72	1142	1	7
5935	84.58	862	2	9	6080	92.75	620	1	3
15986	85.23	346	1	3	15956	92.78	553	0	0
6031	85.32	608	1	3	7032	92.78	556	0	2
4437	85.71	530	0	4	5365	92.86	461	0	7
5690	85.71	346	0	1	4371	93.00	617	1	3
4391	86.89	314	2	6	5895	93.02	509	1	1
16060	89.29	233	1	1	4908	93.04	610	1	2
15763	89.39	368	0	0	15996	93.06	839	0	4
6812	90.38	628	1	10	7243	93.26	518	0	0
15633	90.77	362	0	2	11019	93.55	348	0	1
15635	90.77	372	0	1	6732	93.55	673	5	4
15636	90.77	372	0	1	4795	93.58	467	0	0
15637	90.77	372	0	1	6238	93.62	296	0	1
5797	90.79	444	3	7	15776	93.90	443	0	0
4973	90.82	587	2	2	5757	93.98	304	2	0
6529	90.83	656	0	2	4710	94.02	416	0	0
15758	90.91	355	0	1	6213	94.03	713	4	3
5506	91.27	579	2	3	5956	94.12	408	0	1
5871	91.30	645	0	1	4958	94.23	512	0	2
4981	91.36	496	0	3	15162	94.24	652	2	1
5954	91.67	491	1	3	15726	94.32	970	1	1
5075	91.74	607	0	3	7325	94.33	677	2	5
6582	91.79	713	1	6	4349	94.39	585	1	4
15529	91.89	277	0	0	5462	94.50	927	1	6
6628	91.98	776	0	6	6506	94.55	333	2	0
4893	92.16	620	0	3	4519	94.62	759	0	2
6873	92.17	560	0	1	15932	94.67	1239	2	8
5008	92.31	436	3	0	4708	94.74	443	0	0
7291	92.31	530	3	1	15136	94.78	496	0	0
7319	92.31	1311	0	5	15896	94.78	653	0	2
15036	92.39	549	0	0	15897	94.78	650	0	2
5826	92.42	628	0	3	5545	94.79	564	0	0
5359	92.45	321	1	1	6845	94.83	350	0	0
15680	92.50	1253	0	16	15462	94.87	464	0	0
5175	92.59	422	3	6	16113	94.87	518	1	1
7302	92.68	718	1	3	4078	94.92	710	0	4
5116	92.71	355	0	1	15090	95.00	459	0	1

* Indicates total number of CO, CA, CB, N, H, HN, and HA assignments

and AVS were able to detect a mis-assignment—albeit only partially. This likely has to with the fact that one of the mis-assignments had an unusually large chemical shift difference (F12-R19 = 9.0 ppm), While both SHIFTCOR and AVS marked F12 and R19 as suspicious assignments, only PANAV was able to detect the other 6 chemical shifts that were also switched. This example clearly demonstrates how PANAV is not only able to detect switches with large chemical shift differences such as 9.0 ppm, but also has the ability to detect more subtle mis-assignments that AVS and SHIFTCOR are unable to detect.

Table 4 PANAV, AVS and ShiftCor results on BMRB entries with artificial mis-assignments

BMRB Accession #	Chemical shift difference for switched residues (ppm)		PANAV CONA score*					AVS	ShiftCor
	CA	CB	3-Res	4-Res	5-Res	6-Res	D-S*		
Apolipoproteiin C-III									
15268			98.70 (76/77)	98.68 (75/76)	100 (75/75)	100 (74/74)	0(0)	0 (0)	0
15268_E4_K24	4E-24 K = −3.0	4E-24 K = −3.2	96.10 (74/77)	98.68 (75/76)	100 (75/75)	100 (74/74)	0(0)	0(0)	0
15268_FMQ(11–13)_YMK(15–17)	11E-16Y = −0.29	11E-16 K = −0.1	98.70 (76/77)	98.68 (76/76)	100 (75/75)	100 (74/74)	0(0)	0(0)	0
	12 M-17 M = 0.6	12E-17 K = 0.4							
	13Q-18 K = 0.8	13E-18 K = −3.5							
ATP synthase F6 subunit									
6212			98.65 (73/74)	100 (73/73)	100 (72/72)	100 (71/71)	0(0)	0(0)	0
6212_QKL(8–10)_REY(16–18)	8Q-16R = −1.1	8Q-16R = −1.3	94.59 (70/74)	100 (73/73)	100 (72/72)	100 (71/71)	0(0)	0(0)	0
	9 K-17E = −0.1	9 K-17E = 2.9							
	10L-18Y = −3.0	10L-18Y = 3.3							
6212_EYQQ(34–37)_KLKQ(45–48)	34E-45 K = −1.0	34E-45 K = −3.1	94.59 (70/74)	97.26 (71/73)	100 (72/72)	100 (71/71)	0(0)	0(0)	0
	35Y-46L = 1.7	35Y-46L = −3.5							
	36Q-47 K = 1.7	36Q-47 K = −3.4							
	37Q-48Q = 0.2	37Q-48Q = 0.2							
6212_QKLF(8–11)_REYR(16–19)	8Q-16R = −1.1	8Q-16R = −1.3	93.24 (69/74)	97.26 (71/73)	97.22 (70/72)	100 (71/71)	1(1)	2(0)	2
	9 K-17E = −0.1	9 K-17E = 2.9							
	10L-18Y = −3.0	10L-18Y = 3.3							
	11F-19R = 2.0	11F-19R = 9.0							

D-S #: number of deviant and suspicious (in brackets) shifts as determined by PANAV

AVS #: sum of the mis-typed and suspicious (in brackets) shifts detected by AVS

* Numbers in parentheses indicate the number of fragments with confirmed assignments and the total number of fragments selected

Fig. 4 A screenshot showing PANAV's validation results for two proteins with artificial mis-assignments (bmr15268 and bmr6212). Potentially mis-assigned regions are marked in *red*. As seen in **a**, PANAV indicates that for bmr15268_E4_K24, residues 23–26 should be switched to the 3–6 position. In **b** PANAV's 3-residue scan suggests 4 possible corrections to make to bmr6212_QKL(8-10)_REY(16-18) with most of these suggestions centering around residues 7–10 and 15–18 (the location of the mis-assignment). In **c** PANAV's 4-residue scan suggests 2 possible corrections to make to bmr6212_EYQQ(34-37)_KLKQ(45-48) with most of the problems being identified around residues 34-37. In d) PANAV's 4-residue scan of bmr6212_QKLF(8-11)_REYR(16-19), it suggests 3 possible corrections to make, with 2 being valid

While it is clear that PANAV performs very well under controlled circumstances, we also felt it was important to assess its relative performance using real examples where deviant assignments and mis-assignments have very likely been made. For this test a total of 9 BMRB entries, with significant assignment or chemical shift referencing problems, were manually selected and each was run through PANAV, AVS, LACS, and CheckShift. The output from these runs is summarized in Table 5.

The first entry, bmr4150, is an interesting example where the CB chemical shifts have actually been mis-referenced by ~40 ppm, as detected by both PANAV and CheckShift. If we run bmr4150 through AVS (which does not check for referencing errors) we see that it—incorrectly—indicates that 88 residues appear to have been mis-assigned, including all of bmr4150's CB shifts. Alternately, running bmr4150 through PANAV, we see that PANAV automatically re-references the mis-referenced CB shifts and it indicates that only 26 residues in this protein appear to have been mis-assigned, with most of these misassignments being "deviant" shifts or typographical errors. In fact, bmr4150's 6-residue CONA score is actually 100% after removing the deviant assignments and performing a proper chemical shift reference recalibration.

The second entry in Table 5 is bmr5179. This protein appears to have a CB reference offset of at least 48 ppm as determined by PANAV. Interestingly, neither LACS nor CheckShift were able to process this entry, likely because the CB shifts were so discordant from the CA shifts. After performing the reference correction, PANAV indicates that only 12 of the residues in this protein may have been mis-assigned, which is significantly less than that obtained from AVS (78). After removing the deviant assignments and performing a proper reference recalibration its 6-residue CONA score actually climbs to a respectable 98.63%.

The third entry in Table 5 is bmr6585, which has reference offsets of ~6.0, ~6.6, ~7.7, and 4.9 ppm for CO, CA, CB, and N shifts as detected by PANAV. LACS and CheckShift also generate very similar chemical shift reference offsets for this entry. After performing the reference offset correction, PANAV indicates this protein has no deviant assignments and its 6-residue CONA core is a very respectable 98.45%, indicating it probably has no mis-assignments. In contrast, because AVS is not able to perform chemical shift reference offset correction, AVS calculates that this entry has 58 mis-assigned and 103 suspicious shifts. These examples serve to emphasize that

Table 5 Output from PANAV and AVS on selected BMRB entries

BMRB ID	PANAV						AVS [b]Mis-typed and Suspicous Assignments	LACS Reference Offset (ppm)	CheckShift Reference Offset (ppm)
	[a]Deviant and Suspicious Assignments	Reference Offset (ppm)	CONA score (%)						
			3-Res	4-Res	5-Res	6-Res			
4150	26 (1)	CB: −40.07	87.85	93.40	96.19	100	88(0)	No result	CB: −39.37
5179	12(2)	CB: −47.61	84.21	90.67	95.95	98.63	78 (4)	No result	No result
6586	0(1)	CO: 5.99	79.55	93.13	96.92	97.67	58(103)	CO: 6.05	CO: 6.49
		CA: 6.52						CA: 7.17	CA: 6.89
		CB: 8.14						CB: 7.17	CB: 7.49
		N: 5.01						N: 4.03	N: 4.41
6299	1(0)	Not needed*	76.47	76.12	77.27	77.27	0(1)	Not needed	CO:1.17
6565	1(7)	Not needed	69.09	68.67	70.12	79.14	5(1)	CO: 1.05	Not needed
ARD subunit 1									
7103	0(0)	Not needed	92.53	97.73	98.86	100	0(2)	Not needed	Not needed
4313	3(2)	Not needed	79.29	88.69	91.18	96.41	1(6)	Not needed	Not needed
E9 DNase-Im9									
4293	0(1)	CA: 1.19	90.62	92.97	96.03	98.41	0(1)	Not needed	CB:1.22
		CB: 1.26							
4352	0(3)	CA: 1.33	82.95	89.92	96.09	97.67	0(2)	CA: 1.13	Not needed
		CB: 1.28						CB: 1.13	

Not needed: the detected offsets are less than 1.0 ppm for CO, CA, and CB shifts and 1.5 ppm for N shifts

No result: no output from the program

[a] Number of deviant and suspicious (in brackets) CO, CA, CB, N, H, HA assignments

[b] Number of mis-typed and suspicious (in brackets) CO, CA, CB, N, H, HA assignments

mis-referenced proteins can be mistaken for mis-assigned proteins if the validation program is not sufficiently robust.

The fourth entry in Table 5 is bmr6299. CheckShift indicates that this entry has a relatively minor CO reference offset problem of 1.17 ppm whereas LACS and PANAV predict that this entry has no chemical shift referencing problems. When we run this entry through PANAV and AVS we see only one deviant assignment as identified by PANAV and one suspicious assignment as identified by AVS. However, PANAV also finds that this entry has extremely low CONA scores, 76.47, 76.12, 77.27, and 77.27% for its 3-, 4-, 5-, and 6-residue scans. In particular, unlike most other BMRB entries, the CONA score for this protein seems not to increase with the length of the residue scan, indicating the existence of a large number of subtle mis-assignments. The same situation is also found for bmr6565, which doesn't have a large number of deviant or suspicious assignments but also has extremely low CONA scores. Unfortunately we were not able to obtain the original spectral data for these entries and so it is impossible to know where and why these mis-assignments may have occurred.

Another way of detecting proteins with subtle mis-assignments can be done by looking at protein pairs in the BMRB that have been independently assigned and deposited by two different groups. While dozens of such "redundant" assignment pairs exist in the BMRB, we identified 2 pairs of BMRB entries, bmr4131-bmr7103 and bmr4293-bmr4352, that have the same primary sequence but exhibit significantly different assignments in certain localized regions (i.e. a different CONA score). Both PANAV and AVS predict bmr4131 and bmr4352, which have lower CONA scores, to have more mis-assigned or suspicious chemical shifts than their counterparts. For instance with bmr4131, both PANAV and AVS identify Ile163 and Ile132 to be suspiciously assigned. The CA chemical shifts for Ile163 are listed as 53.2 ppm (for bmr4131) and 60.4 ppm (for bmr7103). 60.4 ppm is much closer to the known Ile CA chemical shifts (given its secondary structure) and is likely the correct assignment. In another example, the assigned CB shifts of residue Ile132 is 29.2 ppm (for bmr4131) and 43.8 ppm (for bmr7103) respectively. Again, 43.8 ppm is much closer to the expected CB shift of Ile (given its secondary structure) and is likely the correct assignment.

We noticed that, for each above selected and several other pairs, the later the entry was reported/deposited in the BMRB (e.g. bmr7103 and bmr4131 were deposited in 2006 and 1998, respectively), the higher the CONA scores. To investigate the influence of year of deposition with the frequency of mis-assignments, we selected 990 entries with protein sizes ranging from 80 to 120 amino acids from the original set of 2422 BMRB entries. Figure 5 shows the average 3-, 4-, 5-, and 6-residue global CONA score plotted versus the year of deposition/submission of these entries. As shown in this figure, the confirmed assignment rates increase with respect to the year of deposition, with the most recent entries having the highest CONA scores. Evidently improved NMR technology, combined with the greater use of doubly and triply labeled proteins, as well as the use of semi- or fully automated assignment software has progressively improved the accuracy of chemical shift assignments. It is also of interest to note that BMRB entries deposited during 2001–2003 appear to have anomalously low CONA scores.

Other influences and limitations on CONA scores

Most of the previous sections have focused on assessing the performance of PANAV relative to other re-referencing and mis-assignment detection programs. Here we shall discuss some of the influences and limitations of CONA scores. One of the potential limitations of PANAV concerns the fact its probabilistic nature means that proteins with longer sequences will tend to exhibit a higher probability of finding segment similarities within proteins, thereby increasing the chance for mis-assignments. Plotting the CONA score as a function of the protein size (or number of residues) shows that the CONA score increases with the length of the fragment, but decreases with the protein size (data not shown). Choosing a longer scanning segment such as a 6-residue fragment ameliorates this effect. Also, calculating the joint probability using more amino acids utilizes more parameters and thus produces more robust results than those calculated by using just a few amino acids. We have also noticed that reference offset re-calibration has a very significant impact on the CONA score. Table 6 lists BMRB entries that exhibit significant improvement in CONA scores after reference calibration.

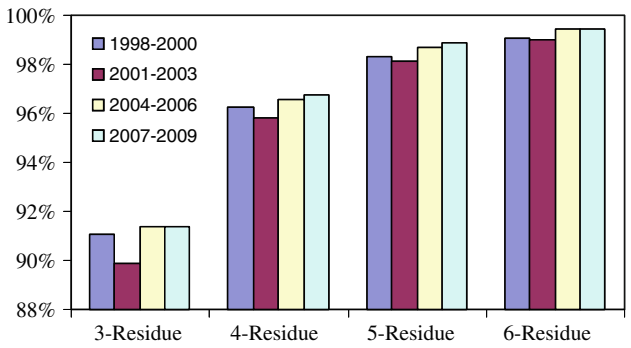

Fig. 5 Average CONA (CONfirmed Assignment) scores using 3-, 4-, 5-, and 6-residue scans versus the year of deposition/submission. The BMRB entries used for this analysis have a protein length ranging from 80 to 120 amino acids

Table 6 BMRB entries that exhibited significant improvement in CONA scores after reference calibration

BMRB	CONA sore (%) before and after (in brackets) ref calibration				Reference offsets (ppm)			
	3-Residue	4-Residue	5-Residue	6-Residue	CO	CA	CB	N
4150	14.02 (87.85)	18.87 (93.40)	27.62 (96.19)	17.31 (100.00)			−40.07	1.57
6586	21.21 (79.55)	32.06 (93.13)	37.69 (96.92)	49.61 (97.67)	5.99	6.52	8.14	5.01
6080	39.85 (79.70)	52.94 (83.09)	70.15 (89.55)	75.36 (92.75)		2.58	3.18	
7351	54.86 (86.81)	64.08 (95.77)	73.76 (99.29)	82.86 (100.00)	2.63	2.70	2.80	
7034	56.10 (87.80)	67.82 (91.95)	80.00 (97.65)	84.09 (100.00)		2.32	3.09	
6531	64.13 (96.74)	72.04 (100.00)	80.43 (100.00)	84.78 (100.00)	2.60	2.35	3.23	
7414	44.50 (74.35)	61.26 (89.53)	75.79 (95.79)	84.13 (98.94)	3.17	2.94	3.16	
15820	58.00 (87.00)	65.35 (92.08)	78.00 (96.00)	83.33 (98.04)	2.55	2.28	3.16	
5678	55.26 (75.00)	60.26 (80.13)	73.83 (87.92)	81.33 (96.00)	2.09	1.72	2.85	
5324	54.38 (91.88)	72.22 (94.44)	77.02 (95.65)	84.18 (97.47)		3.03	4.02	
6612	50.50 (82.18)	63.46 (89.42)	81.37 (97.06)	85.05 (98.13)		1.93	3.00	−1.57
4953	54.46 (88.12)	72.28 (100.00)	83.00 (100.00)	87.00 (100.00)	3.45	2.53	2.75	
15148	54.10 (87.43)	61.11 (93.33)	77.22 (95.56)	86.67 (99.44)		1.74	3.08	
6532	54.55 (92.05)	68.89 (92.22)	80.90 (98.88)	88.64 (100.00)		2.19	3.10	
11019	62.30 (95.08)	76.19 (85.71)	79.03 (93.55)	82.26 (93.55)	2.57	2.83	3.10	
5547	42.97 (83.13)	62.25 (94.38)	80.65 (96.77)	89.52 (99.19)		2.45	4.36	
5179	27.63 (84.21)	40.00 (90.67)	63.51 (95.95)	89.04 (98.63)			−47.61	
5558	64.44 (91.11)	73.91 (95.65)	84.44 (97.78)	91.30 (100.00)		2.32	2.56	
15070	57.07 (84.39)	68.60 (91.79)	81.64 (97.10)	90.34 (99.03)		1.57	3.72	
4445	64.63 (98.78)	78.05 (98.78)	85.19 (100.00)	91.36 (100.00)		1.94	2.42	
15315	62.26 (90.57)	77.22 (94.94)	88.54 (99.36)	91.67 (99.36)	3.18	2.83	2.54	
7035	72.55 (90.20)	75.93 (90.74)	86.79 (96.23)	92.31 (100.00)		1.97	1.56	

Conclusion

PANAV represents a new sequence-based, probabilistic method that is uniquely able to detect chemical assignment errors (mis-assignments, deviant assignments and suspicious assignments) as well as correct chemical shift referencing errors in both peptides and proteins. While other programs such as CheckShift and LACS are able to correct chemical shift referencing errors, they are not able to detect chemical shift assignment errors. Likewise, programs such as AVS and certain utilities at the BMRB are able to detect assignment errors, they are not able to perform chemical shift re-referencing. PANAV is uniquely capable of performing both operations—without prior knowledge of the protein's 3D structure. As illustrated in this paper, there are situations where errors in chemical shift referencing can create the "illusion" that chemical shift assignment errors have been made. Likewise, errors in chemical shift assignments or in data entry can lead to potential errors in the calculation of chemical shift reference offsets. Therefore, having the capacity to do both kinds of operations is critical to catching and correcting both kinds of errors. In addition to its dual corrective capabilities, we have also shown that PANAV is able to handle a much wider number of re-referencing problems than either LACS or CheckShift and it is able to detect much more subtle mis-assignment errors than either AVS or SHIFTCOR. By applying PANAV to a large number of BMRB entries we have shown that chemical shift mis-referencing problems continue to be widespread in the biomolecular NMR community, with about 20% of all BMRB entries being improperly referenced. Likewise, it appears that at least 40% of all BMRB entries have at least one mis-assignment. Hopefully the development and widespread use of programs such as PANAV could go a long way towards detecting and eliminating persistent problems of chemical shift mis-referencing and chemical shift mis-assignment that, unfortunately, seem to plague the field of protein NMR.

Acknowledgments DSW wishes to acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) and the Alberta Prion Research Institute (APRI).

References

de Dios AC, Pearson JG, Oldfield E (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an ab initio approach. Science 260:1491–1496

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

Gronenborn AM, Clore GM (1994) Identification of N-terminal helix capping boxes by means of 13C chemical shifts. J Biomol NMR 4:455–458

Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895

Metzler WJ, Constantine KL, Friedrichs MS, Bell AJ, Ernst EG, Lavoie TB, Mueller L (1993) Characterization of the three-dimensional solution structure of human profiling: 1H, 13C and 15 N NMR assignments and global folding pattern. Biochemistry 32:13818–13829

Moseley HN, Sahota G, Montelione GT (2004) Assigment validation software suite for the evaluation and presentatio of protein resonance assignment data. J Biomol NMR 28:341–355

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15 N chemical shifts. J Biomol NMR 26:215–240

Seavey BR, Farr EA, Westler WM, Markley JL (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1:217–236

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Sugnarapu KK, Ignatchenko A, Arrowsmith CH, Syperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

Spera S, Bax A (1991) Empirical correlation between protein backbone conformatio and Ca and Cb 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

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

Wang Y, Jardetzky O (2002a) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084

Wang Y, Jardetzky O (2002b) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

Wang Y, Jardetzky O (2004) Predicting 15 N chemical shifts in proteins using the preceding residue-specific individual shielding surfaces from phi, psi i-1 and chi 1 torsion angles. J Biomol NMR 28:327–340

Wang L, Markley JL (2009) Empirical correlation between protein backbone 15 N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J Biomol NMR 44:95–99

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

Wishart DS, Nip AM (1998) Protein chemical shift analysis: a practical guide. Biochem Cell Biol 76:153–163

Wishart DS, Sykes BD (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J Biomol NMR 4:171–180

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222:311–333

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651

Xu XP, Case DA (2001) Automated prediction of 15 N, 13Callpha, 13Cbeta and 13C' chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195