



RefDB: A database of uniformly referenced protein chemical shifts

Haiyan Zhang, Stephen Neal & David S. Wishart*

Faculty of Pharmacy & Pharmaceutical Sciences, University of Alberta, Edmonton, AB, T6G 2N8, Canada

Received 12 August 2002; Accepted 27 December 2002

Key words: chemical shift, DSS, NMR, protein, referencing

Abstract

RefDB is a secondary database of reference-corrected protein chemical shifts derived from the BioMagResBank (BMRB). The database was assembled by using a recently developed program (SHIFTX) to predict protein ^1H , ^{13}C and ^{15}N chemical shifts from X-ray or NMR coordinate data of previously assigned proteins. The predicted shifts were then compared with the corresponding observed shifts and a variety of statistical evaluations performed. In this way, potential mis-assignments, typographical errors and chemical referencing errors could be identified and, in many cases, corrected. This approach allows for an unbiased, instrument-independent solution to the problem of retrospectively re-referencing published protein chemical shifts. Results from this study indicate that nearly 25% of BMRB entries with ^{13}C protein assignments and 27% of BMRB entries with ^{15}N protein assignments required significant chemical shift reference readjustments. Additionally, nearly 40% of protein entries deposited in the BioMagResBank appear to have at least one assignment error. From this study it is evident that protein NMR spectroscopists are increasingly adhering to recommended IUPAC ^{13}C and ^{15}N chemical shift referencing conventions, however, approximately 20% of newly deposited protein entries in the BMRB are still being incorrectly referenced. This is cause for some concern. However, the utilization of RefDB and its companion programs may help mitigate this ongoing problem. RefDB is updated weekly and the database, along with its associated software, is freely available at <http://redpoll.pharmacy.ualberta.ca> and the BMRB website.

Introduction

Chemical shifts are perhaps the most precisely measurable but the least accurately measured parameters in NMR spectroscopy. This curious state of affairs has arisen because, unlike most spectroscopic measurements, chemical shifts are relative. As such, chemical shifts are prone to numerous kinds of reporting and measurement errors. The problem with chemical shift measurement is particularly acute in biomolecular NMR. Indeed, the large number of chemical shifts that must be measured (hundreds to thousands), the variety of chemical shifts (^1H , ^{13}C , ^{15}N , ^{31}P), and the incredible range of solvent conditions (pH, temperature, salts, organic solvent mixtures) – all contribute to the problem. A further complication has been the historic reliance on many different chemical shift stan-

dards or chemical shift measurement protocols – many of which are now obsolete or widely considered to be irreproducible. The problems with chemical shift standardization have been discussed at length in a number of recent articles (Wishart and Sykes, 1994a; Iwadate et al., 1999; Cornilescu et al., 1999; Wishart and Case, 2001) and several suggestions or widely-agreed upon standards have been advocated (Wishart et al., 1995; Wishart and Sykes, 1994a; Maurer and Kalbitzer, 1996; Markley et al., 1998).

A key point raised by these authors has been the fact that biomolecular chemical shifts, in particular, contain a tremendously rich source of structural and dynamic information. However, the structural and dynamic information contained in chemical shifts is subtle and, consequently, inaccurate or incorrectly referenced chemical shift measurements can easily blur or distort an exquisitely detailed picture of a biomolecule.

*To whom correspondence should be addressed. E-mail: david.wishart@ualberta.ca

The BioMagResBank (Seavey et al., 1991) was established in 1991 to help address some of the problems and inconsistencies in biomolecular chemical shift reporting. Over the past 10 years the BMRB has given biomolecular NMR spectroscopists a superb opportunity to systematically assemble, compare and interpret chemical shifts. It has been through the BMRB, for instance, that a number of important chemical shift trends have been identified (Spera and Bax, 1991; Wishart et al., 1991, 1992; Metzler et al., 1993; Gronenborn and Clore, 1994) and a variety of chemical shift theories or prediction/assignment schemes have been refined (Osapay and Case, 1994; Wishart and Nip, 1998; Beger and Bolton, 1997; Le and Oldfield, 1994). Throughout its 10-year history the BioMagResBank has served as a superb historical archive as it has meticulously recorded the ever-changing trends in chemical shift measurement and reporting. Because the BMRB is an archival database (it accepts 'as-is' data directly from depositors) it depends crucially on the integrity and accuracy of its depositors. However, given the nature of chemical shift assignments and the variability of chemical shift reporting, it has been difficult to develop a rigorous set of protocols to validate the chemical shifts being deposited into the BMRB. As a result the BMRB likely contains a number of chemical shift assignments which have been improperly referenced or incorrectly assigned (Williamson et al., 1995; Iwadate et al., 1999; Wishart and Case, 2001). Indeed, a preliminary survey conducted in 2000 suggested that up to 20% of ^{13}C shifts and 30% of ^{15}N shifts are incorrectly referenced (Wishart and Case, 2001). This result is of some concern and it leads immediately to a number of important questions: What is the true magnitude of these referencing problems? What nuclei are most frequently or significantly affected? Do these affect the chemical shift trends or theories that have been developed from BMRB data? Can a corrected set of shifts be assembled? Can a chemical shift validation suite or protocol be developed for the BMRB?

Here we wish to report on the development of a set of software tools and a complementary chemical shift database (RefDB) containing a subset of BMRB chemical shifts that have been properly re-referenced according to the IUPAC/IUB conventions (Wishart et al., 1995; Markley et al., 1998). We also demonstrate how these analysis tools can be used not only to correctly reference chemical shifts, but to identify potential mis-assignments, to flag typographical errors, to detect spectral folding problems and zero-in on the

location of potential structural differences or structure refinement errors. We also show how the RefDB shifts can be used to generate a more refined set of secondary shifts for all 20 amino acids, a tabulation which may be of some use in secondary structure analysis and empirical chemical shift calculations.

Materials and methods

RefDB was prepared using a combination of three different computer programs. The first program (SHIFTX) calculates backbone ^1H , ^{13}C and ^{15}N chemical shifts from protein 3D coordinate data. The second program (SHIFTCOR) compares the calculated shifts with the observed shifts, evaluates any statistically significant differences and performs the necessary chemical shift corrections. The third program (UPDATE) automatically retrieves newly deposited BMRB data along with any corresponding PDB data. UPDATE also directs the data to SHIFTCOR and appends the 'corrected' chemical shift file to the RefDB database. A more detailed description of each program follows.

SHIFTX

SHIFTX uses a semi-empirical approach to calculate ^1H , ^{13}C and ^{15}N protein chemical shifts. The program employs both a combination of empirically derived chemical shift hypersurfaces (Spera and Bax, 1991; Le and Oldfield, 1994; Wishart and Nip, 1998) and classically calculated ring-current, electric field, nearest neighbor and hydrogen bond effects (Wagner, 1983; Wishart et al., 1991; Osapay and Case, 1991). The hypersurfaces, which relate ^1H , ^{13}C and ^{15}N chemical shifts to backbone dihedral angles, were derived from the chemical shift assignments of 37 fully assigned and highly resolved ($< 2.1 \text{ \AA}$) X-ray structures in a manner similar to Iwadate et al. (1999). Ring current effects were calculated using the method of Haigh and Mallion (1979), whereas electric field and hydrogen bonding effects were calculated using methods similar to Osapay and Case (1991) and Wagner (1983). Nearest neighbor effects and local side chain effects were derived through a specialized data-mining program and incorporated into SHIFTX as empirical correction factors or as chemical shift 'hypersurfaces'. Nucleus-specific constants were calculated for ring-current, electric field, nearest neighbor and hydrogen bond effects.

Table 1. Data used to train and test SHIFTX

PDB ID	Protein name	Resolution (Å)	BMRB accession / Reference
2ALP	Alpha-lytic protease (L. enzymogenes)	1.70	(Davis et al., 1997) ^a
1GZI	Antifreeze Protein (Ocean pout)	1.80	(Sonnichsen, F.D.) ^b
1A6K	Myoglobin (Sperm whale)	1.10	4061
1A2P	Barnase (B. amyloliquefaciens)	1.50	975
4ICB	Calbindin D9K, minor A form (Pig)	1.60	390
1CLL	Calmodulin (Drosophila)	1.70	547
1ROP	ColE1 Repressor protein (E. coli)	1.70	4072
1CEX	Cutinase (F. solani)	1.00	4101
3EZM	Cyanovirin-N (Nostoc ellipsosporum)	1.50	(Bewley et al., 1998) ^a
2CPL	Cyclophilin-A (Human)	1.63	(Ottiger et al., 1997) ^a
1HCB	Carbonic anhydrase I (Human)	1.60	4022
1DMB	D-maltodextrin-binding protein (E. coli)	1.80	4354
1ICM	Fatty Acid Binding Protein (Rat)	1.20	(Hodsdon et al., 1995)
1HFC	Fibroblast Collagenase (Human)	1.56	4064
4FGF	Fibroblast Growth Factor (Human)	1.60	4091
1BKF	FK506 Binding Protein (Human)	1.60	4077
1HVR	HIV Protease (HIV)	1.80	(Yamazaki et al., 1996) ^a
4I1B	Interleukin 1 β (Human)	2.00	1061
3LZT	Lysozyme (Chicken)	0.92	4562
1LZ1	Lysozyme (Human)	1.50	5130
1ONC	P-30 Protein (Northern leopard frog)	1.70	4371
5PTI	Pancreatic trypsin inhibitor (Bovine)	1.00	46, 262, 485
1F3G	Phosphocarrier protein III glc (E. coli)	2.10	(Pelton et al., 1991) ^a
1ACF	Profilin I (A. castellanii)	2.00	(Archer et al., 1994) ^a
1HKA	Pyrophosphokinase (E. coli)	1.50	4299
5P21	RAS P21 (Human)	1.35	(Campbell-Burk, 1997)
1RUV	Ribonuclease A (Bovine)	1.30	4031
2RN2	Ribonuclease H (E. coli)	1.48	(Yamazaki et al., 1991)
1RGE	Ribonuclease S (S. aureofaciens)	1.15	4259
2RNT	Ribonuclease T1 (Aspergillus oryzae)	1.80	(Weisemann et al., 1993)
1SVN	Savinase (Bacillus lentus)	1.40	(Foghe et al., 1995) ^a
1SNC	Staphylococcal Nuclease (S. aureus)	1.65	(Cornilescu et al., 1999) ^a
1MKA	Thiol ester dehydrase (E. coli)	2.00	(Copie et al., 1996) ^a
2TRX	Thioredoxin (E. coli)	1.68	(Chandrasekhar, 1997)
1ERT	Thioredoxin - reduced (Human)	1.70	(Qin et al., 1996) ^a
1TOP	Troponin C (Chicken)	1.78	4401
1UBQ	Ubiquitin (Human)	1.80	(Wang et al., 1995) ^a

^aIndicates these shifts are from the re-referenced TALOS database.

^bPersonal communication.

The database used to calibrate the ¹H, ¹³C and ¹⁵N shift calculations consisted of 37 diamagnetic proteins assembled from an extensive literature and BMRB database search. These proteins, their BMRB and PDB accession numbers as well as their corresponding X-ray resolution are listed in Table 1. In preparing this database every effort was made to find those proteins which (a) spanned a variety of structural classes (all

α , all β , mixed α/β), (b) were well-structured, (c) had no ‘shift-significant’ ligands or paramagnetic moieties, and (d) had high resolution X-ray structures (< 2.1 Å). Myoglobin (BMRB 4061), which does have a heme group, was included in this training/testing set because of the availability of an exceptionally high resolution X-ray structure and because we only used its ¹³C α and ¹³C β shifts in training and evaluating

SHIFTX. Note that ^{13}C shifts are essentially insensitive to heme-induced ring current effects (Iwdate et al., 1999; Wishart and Case, 2001) – a fact that was subsequently confirmed by the excellent fit between the calculated and observed ^{13}C shifts for this protein.

To ensure that SHIFTX was capable of calculating ^1H , ^{13}C and ^{15}N shifts that were consistent with IUPAC chemical referencing recommendations (Markley et al., 1998; Wishart et al., 1995) several ‘bootstrapping’ reference validation tests were employed during the development of the program. Initially 6 proteins (cutinase, calmodulin, cole1 Rop, hen lysozyme, ribonuclease A and carbonic anhydrase – see Table 1 for their BMRB accession numbers) that had been assigned by 6 different labs in North America and Europe were selected from the BMRB. In choosing a wide variety of geographically distinct sources, we hoped to reduce the referencing bias that might exist if we selected assignments from only one lab, one type of spectrometer or only one region. Each of these chemical shift assignments was reported, both in the literature and BMRB database submissions, to have been referenced to DSS and liquid ammonia according IUPAC recommendations. The exception was calmodulin which used TSP, instead of DSS for ^{13}C referencing (this was corrected by adjusting the ^{13}C shifts by 0.12 ppm as noted by Wishart and Sykes (1994a)). From this small training set, SHIFTX was initially calibrated in 1998 (Wishart and Nip, 1998). The calibration process consisted of calculating the difference between the observed shifts and SHIFTX-calculated shifts for each nucleus and for each residue. These were then averaged over the total of all residues for a given nucleus to generate a reference ‘offset’. If the differences were only due random or digital resolution errors, these offsets would be expected to average to zero. If these offsets are non-zero and greater than two standard deviations away, then this would suggest a systematic (i.e., referencing) error. Comparisons between the observed and calculated ^1H , ^{13}C and ^{15}N shift offsets, conducted at that time, indicated that the difference between the SHIFTX-calculated ‘zero-point’ reference and the observed ‘zero-point’ reference was no greater than ± 0.03 ppm for ^1H , ± 0.28 ppm for ^{13}C and ± 0.60 ppm for ^{15}N for any one of the 6 proteins. This back-calculation confirmed that all 6 proteins were referenced almost identically (i.e., correctly) or referenced to within the precision of the SHIFTX calculation (see later). In addition, chemical shift data for several other proteins (BPTI, thioredoxin, ribonuclease H – see Table 1) that had been

explicitly referenced to dioxane (for ^{13}C) were also tested and the calculated reference offsets (~ 1.6 ppm) were shown to be essentially identical to those determined from earlier studies (Wishart and Sykes, 1994b). These calculated reference offsets were added to the reported shifts of these ‘non-IUPAC’ referenced proteins and the SHIFTX training database was eventually expanded to 12 entries. After 1998, additional assignment data was added to the SHIFTX training set, including the reference-corrected TALOS data of Cornilescu and Bax (1999). Tests conducted with the TALOS data using the 1998 version of SHIFTX indicated that these new entries were referenced consistently (i.e., within 0.5 ppm of the SHIFTX calculated zero-point reference for ^{13}C and ^{15}N shifts) with the SHIFTX training data. Consequently the TALOS data (<http://spin.niddk.nih.gov/NMRPipe/talos/>) were added to the SHIFTX training set without further adjustment. Later additions eventually brought the database up to its current size of 37 proteins. Using this database, SHIFTX was trained and tested throughout 2001 using standard data mining and parameter optimization protocols. Specifically, the database was divided into equal-sized test and training sets, with every other residue being assigned to the test set. As a further check against over-fitting, all optimization steps were evaluated by testing their results against randomly-chosen samples from the databases. Several (usually twenty) such samples would be generated and evaluated (in terms of correlation or RMSD) against the optimized surfaces. Additional verification was done on a protein-by-protein basis as well as on an amino acid-by-amino acid basis to detect any systematic bias in the fitting functions. The optimization criteria were constructed to: (1) Maximize the correlation coefficient (between observed and calculated chemical shifts), and (2) minimize the RMS error.

In minimizing the RMS error the initial set of reference-corrected chemical shifts were allowed to be further adjusted (i.e., re-referenced on a whole protein basis). Typically these changes were less than 0.02 ppm for ^1H shifts and less than 0.1 ppm for ^{13}C and ^{15}N shifts for any given protein entry. While it might be argued that this kind of RMS optimization creates an artificial (i.e., non-IUPAC) and unrealistically precise ‘zero-point’ standard, it is a necessary part of any optimization process. It is also a process that has been routinely used by several other investigators working on ^{13}C and ^{15}N chemical shift calculation (Iwdate et al., 1999; Xu and Case, 2001). Using the final, fully optimized version of SHIFTX,

another round of reference-validation and checking was conducted. Reference offsets were recalculated for the original set of 6 IUPAC-referenced test proteins and for the original TALOS set of proteins. The results from these tests were essentially identical to the results described earlier, with the TALOS and IUPAC-referenced proteins exhibiting no systematic or statistically significant reference offset bias (i.e., the differences between the SHIFTX-calculated 'zero-point' reference and the observed 'zero-point' reference were no greater than ± 0.07 ppm for ^1H , ± 0.32 ppm for ^{13}C and ± 0.57 ppm for ^{15}N). As a final, external check we also compared the reference corrections reported by Iwate et al. (1999) for their ^{13}C database with reference corrections calculated by SHIFTX for proteins which closely matched entries (both in terms of structure and assignments) in our RefDB data set. Iwate et al. attempted to correct their ^{13}C shifts to a TSP or near-IUPAC standard. A total of 9 proteins were identified (BPTI, interleukin 1 β , staphylococcal nuclease, ribonuclease H, Fk506 binding protein, cyclophilin, interleukin 4, human profilin and HPr). The difference between the two sets of calculated offsets for ^{13}C shifts is 0.06 ± 0.10 ppm – which is statistically insignificant from zero. Hence, on the basis of multiple internal and external tests, we are confident that SHIFTX calculates ^1H , ^{13}C and ^{15}N shifts that, to the level of its computational accuracy, are statistically indistinguishable from IUPAC referenced shifts.

Overall, the SHIFTX program is able to attain a correlation coefficient (r) between observed and calculated shifts in diamagnetic proteins of 0.905 ($^1\text{H}\alpha$), 0.977 ($^{13}\text{C}\alpha$), 0.996 ($^{13}\text{C}\beta$), 0.860 (^{13}CO), 0.896 (^{15}N) and 0.732 (^1HN). The RMS error is 0.23, 1.06, 1.15, 1.18, 2.60, 0.52 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts, respectively. Relative to the width of the respective chemical shift distributions these RMS errors translate to a 'percent error' of 5.8%, 4.2%, 2.0%, 10.8%, 7.5%, 13% for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts, respectively. This percent error is roughly inversely proportional to the correlation coefficient. Compared to previously published shift prediction programs (Osapay and Case, 1994; Iwate et al., 1999; Case, 2000; Xu and Case, 2001) SHIFTX appears to perform as well or, in some cases better, for a larger number and a greater variety of shifts among diamagnetic proteins. SHIFTX reads standard PDB-formatted files and outputs the predicted chemical shifts in a simple tabular form (BMRB or SHIFTY (Wishart et al., 1997) format).

More complete details regarding the performance and structure of SHIFTX will be forthcoming shortly (S. Neal, A. Nip, H. Zhang and D. Wishart, submitted).

SHIFTCOR

SHIFTCOR is an automated shift correction program that uses statistical methods to compare and correct SHIFTX-predicted shifts relative to an input set of observed chemical shifts. SHIFTCOR uses several simple statistical approaches and pre-determined cut-off values to identify and correct potential referencing, assignment and typographical errors. The standard input for the SHIFTCOR program is a set of observed chemical shifts (BMRB or SHIFTY format) and a corresponding PDB file. SHIFTCOR identifies potential chemical shift referencing problems by comparing the difference between the average value of each set ($^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN) of observed and predicted chemical shifts. The difference between these two averages results in a nucleus-specific chemical shift offset or reference correction. To ensure that these average offset values are not unduly biased by certain extreme outliers, the average of the observed shifts is only calculated after excluding potential mis-assignments or typographical errors. Potential mis-assignments are initially identified by looking for predicted chemical shifts that differ from their corresponding observed chemical shifts by approximately 4 standard deviations (i.e., $4 \times$ the RMS error expected for SHIFTX predicted shifts). Specifically the maximal cutoff differences were 0.7, 5.0, 5.0, 5.0, 10.0 and 2.0 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts, respectively. These values were determined after an extensive series of cut-off selection trials and later rounded up or down for ease of recall. Notice that the precise cut-off value differs slightly for each nucleus due to a combination of factors. It is important to note that SHIFTCOR is not capable of detecting or classifying typographical errors (missing or added digits), switched assignments (i.e., Ser for Thr) or other anomalies. These were identified manually (after initially being identified as mis-assignments).

Because SHIFTCOR determines reference corrections on the basis of a statistically averaging procedure is important to know or understand the limits of its precision and accuracy. Intuitively, the reliability of the calculated reference offsets depend on the accuracy of the chemical shift calculation for a given nucleus and the number of shifts that are being averaged for that nucleus. What we would like then is a confi-

```

#N.B. (Observed* = Observed shift + Offset correction)
#
#After reference correction, the following residues still
#have a HA chemical shift difference (obs*-pred) greater than
0.7ppm:
#NUM  AA   CS   Observed*   Predicted
# 17  Q   HA    4.04         5.25
# 76  F   HA    4.40         5.47
# 77  G   HA    5.18         4.19
# 78  Q   HA    4.22         5.29
# 98  L   HA    4.17         2.61
#102  A   HA    4.54         3.50
#104  K   HA    4.69         2.94
#105  F   HA    4.85         3.43
#107  E   HA    5.24         3.79
#109  K   HA    5.23         3.91
#
#After reference correction, the following residues still
#have a CB chemical shift difference (obs*-pred) greater than
5.0ppm:
#NUM  AA   CS   Observed*   Predicted
# 99  S   CB   56.55       62.34
#
#The average CS difference between predicted and observed:
#HA      CA      CB      CO      N      HN
#0.04    0.05    0.29    -0.21   -1.87   0.17
#
#bmr4766.str.corr chemical shifts have been re-referenced with
the following
#offsets (these values have been added to the original
bmr4766.str file):
#HA      CA      CB      CO      N      HN
#N/A     +0.17   +0.17   -0.21   -1.87   N/A
#
#The 95% confidence intervals for the above recommended offsets
are:
#      HA      CA      CB      CO      N      HN
# +/-0.03 +/-0.12 +/-0.16 +/-0.13 +/-0.41 +/-0.07
#
#The Correlation Coefficients between predicted and observed
#chemical shifts are:
#HA      CA      CB      CO      N      HN
#0.542   0.975   0.996   0.865   0.851   0.741
#
#The RMSD between predicted and observed* (reference
#corrected) chemical shifts are:
#HA      CA      CB      CO      N      HN
#0.159   0.608   0.765   0.672   2.004   0.320

```

Figure 1. SHIFTCOR output for bmr4766.str.

dence interval for the population mean (mean of the difference between the calculated and observed shifts) under the situation where the population variance is unknown and the distribution of shift-difference errors is normal. This statistic is best calculated using a student's *t*-test working from the standard deviation (RMSD) of chemical shifts calculated for each nucleus in each individual protein. For SHIFTCOR we have chosen to calculate the 95% confidence interval (roughly 2 standard deviations in the *t* distribution) for every calculated reference offset. Therefore, for an average protein of average length (105 residues in RefDB) and using the average standard deviations calculated by SHIFTX for each nucleus in RefDB, one can be 95% confident that the reference offsets calculated by SHIFTCOR are within ± 0.05 , ± 0.22 , ± 0.23 , ± 0.19 , ± 0.47 , and ± 0.12 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts, respectively. These values only reflect an overall average for RefDB as a whole. SHIFTCOR actually generates and reports protein-specific values for each nucleus (Figure 1). As will be shown later, these confidence intervals reliably capture the observed variation in chemical shift reporting and they also appear to cover for any potential errors or oversights that may occur in the RefDB updating process.

When SHIFTCOR is run, it creates two files. One contains the chemical shift analyses (including lists of potential mis-assignments, estimates of the referencing errors, the estimated error in the calculated reference offset (95% confidence interval), the applied or suggested reference offset, correlation coefficients, RMSD values) and the other contains the corrected BMRB formatted chemical shift file with the SHIFTCOR analysis attached as a header (see Figure 1 for an example).

As can be seen from this figure, the reference or offset corrections applied to the chemical shift file (designated as a *.corr file) are calculated by averaging the $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shift offsets and adding this common value to all $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shifts to this entry. We average these offsets because $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shifts are typically measured in the same experiment, so one would expect their offsets to be identical. Our data indicate that >95% of calculated $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ offsets are within 0.5 ppm of each other (i.e. within the SHIFTX error of the two ^{13}C shift calculations). In situations where the offsets differ by more than 0.5 ppm, SHIFTCOR flags their occurrence. These discordant $^{13}\text{C}\alpha/^{13}\text{C}\beta$ shifts may reflect differential shifting due to deuteration, limited sampling of one nucleus rela-

tive to the other, incorrect instrumental calibration or possibly assignment errors. Unlike $^{13}\text{C}\alpha/^{13}\text{C}\beta$ shifts, ^{13}CO and ^{15}N shift offsets are treated and applied independently as these shifts are typically measured in separate NMR experiments with separate chemical shift referencing calibrations. Note that while SHIFTCOR does report $^1\text{H}\alpha$ and ^1HN shift offsets, these are never applied to the shifts listed in the *.corr file. This is because referencing problems with ^1H data are essentially non-existent or statistically insignificant (Wishart and Case, 2001).

For the identification and presentation of outliers in the SHIFTCOR output, we use reference-corrected observed shifts (denoted as Observed*) instead of uncorrected observed shifts as our comparison standard. This was done in order to handle situations where the reference corrections were quite large (2–4 ppm). In these situations, virtually every ^{13}C and ^{15}N shift would be identified as an outlier if their raw, uncorrected values were compared to the SHIFTX predicted shifts. Given that reference corrections of 2–4 ppm proved to be quite common (>20% of entries), it was determined that the reference-corrected observed (Observed*) shifts would be more useful for flagging real problem assignments or outliers.

UPDATE

UPDATE is a database updating program designed to automatically process newly deposited protein chemical shift data in the BioMagResBank and store the results in the RefDB database (both the web version and a flat file version). It can be divided into six steps (Figure 2). Firstly, UPDATE uses standard web query protocols to identify and download newly deposited chemical shift data in the BioMagResBank. Second, after downloading the BMRB file, UPDATE reads the file description keywords to determine if the file contains a protein or not. Third, the file is checked (using a simple text matching algorithm) to determine if it corresponds to a paramagnetic protein, a denatured protein, a protein dissolved in an organic solvent or a protein with a heme group (with the exception of myoglobin). If the entry passes these tests, UPDATE extracts the sequence from the BMRB file and uses a web-based query to conduct a BLAST sequence search against the PDB. At least 50% of the length of the query (BMRB) sequence (or at least 50 residues, which ever is less) must overlap with the selected PDB file and the overlapped region must have at least 95% sequence identity within the over-

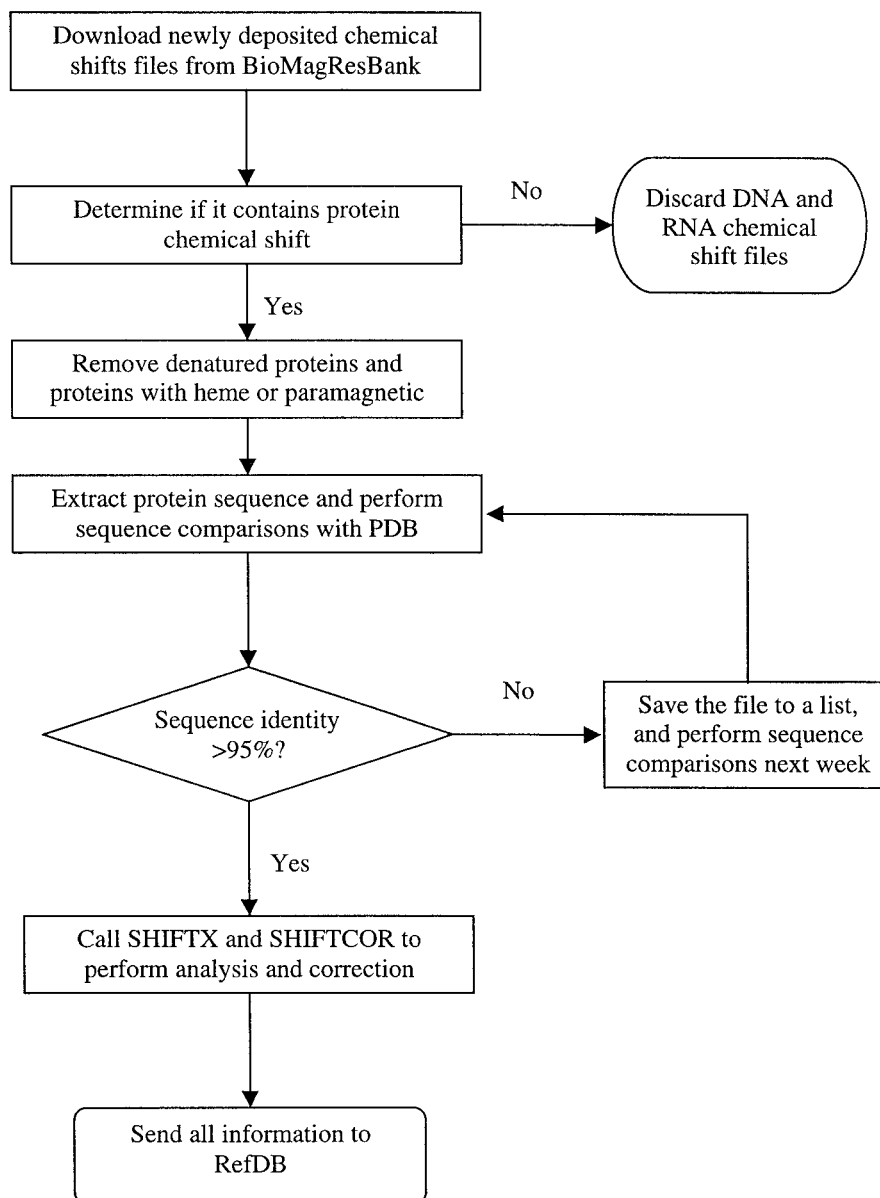


Figure 2. An outline of the UPDATE algorithm.

lapped region in order to be identified as a ‘matching’ PDB file. If a single match is found, it is downloaded and processed. If more than one PDB file is found, the 3D coordinate file that yields the highest average chemical shift correlation coefficient (averaged over all reported shifts) is selected. The use of calculated chemical shift correlation coefficients helps the program automatically eliminate PDB files that may contain structures of chemically modified proteins or those which may have structure-altering ligands that

would be inconsistent or different from the protein sample used to generate the NMR assignments. Ideally an ‘intelligent’ text analysis of the respective BMRB and PDB header files should be used to find the most appropriate match, but the inconsistencies between the word usage and the complexities of biological and chemical nomenclature made this too difficult to code. Hence we chose the correlation coefficient to serve as a numerical proxy to structural similarity or structural identity. Overall, this selection criteria seems to

work well (i.e., it is routinely able to select a PDB file that closely matches the structure for which NMR data was generated). If two or more proteins have the same chemical shift correlation coefficients, the structure with the higher resolution (or the X-ray structure) is chosen. X-ray structures are given precedence over NMR structures because of their intrinsically higher resolution (Vriend, 1990; Laskowski et al., 1993). If the PDB file contains more than one structure (as is the case with many NMR data sets) UPDATE selects only the first structure in the ensemble for processing. Tests using reference offset averages calculated for the entire ensemble (up to 32 structures) indicated that these values were statistically no different than those calculated for a single member (see Results and Discussion). If the X-ray/NMR structure differs in length from the reported assignments, only those residues with 3D coordinates will have their chemical shifts calculated and adjusted.

After the appropriate PDB file has been selected and automatically downloaded, SHIFTX and SHIFTCOR are then called to perform their respective calculations and corrections. UPDATE then appends these corrected data files, along with the corresponding 3D coordinates to the RefDB database. UPDATE also uses calculated chemical correlation coefficients to flag proteins that appear to have serious assignment, alignment or structural problems. Specifically, if a protein is processed that has either a $^1\text{H}\alpha$ correlation coefficient below 0.3, a $^{13}\text{C}\alpha$ correlation coefficient below 0.9, a ^{13}CO correlation coefficient below 0.2 or a ^1HN correlation coefficient below 0.1, then UPDATE sends an email note to the RefDB archivist indicating which protein and which nucleus appears to be causing the problem. The archivist then evaluates the data along with any other information available about the protein and makes a determination on whether to include or remove the offending protein from RefDB. If the protein entry is kept, a double asterisk is added to the BMRB accession number as listed in both the web and flat file version of RefDB. In addition to this periodic manual intervention, the UPDATE program along with RefDB is routinely checked (manually, on a quarterly basis) to ensure it is running correctly and to make periodic changes or corrections to extracted protein names.

RefDB

Currently RefDB contains nearly 500 sets of corrected protein chemical shifts. All of the original chemical shift sets were obtained from the BioMag-

ResBank. Each polypeptide in RefDB is required to contain at least 25 residues and to have an X-ray or NMR structure deposited in the PDB with backbone and side chain coordinates. RefDB does not include proteins dissolved in urea, DMSO or other organic solvents (except TFE) since these solvents can affect the chemical shift referencing in unpredictable ways (Wishart et al., 1995; Wishart and Nip, 1998). Furthermore, polypeptides dissolved in these solvents differ substantially from their native (X-ray) or reference structure. RefDB exists as both a single flat-file (~30 Megabytes) for convenient downloading, and as a web-enabled, queryable database. The RefDB web server is located at <http://redpoll.pharmacy.ualberta.ca> (Figure 3). The web version of RefDB uses a formatted table to list the name of the original BMRB file (hyperlinked to the BMRB site), the name of the corrected or adjusted shift file (hyperlinked to the shift list), the full name of the protein and the PDB accession number of the corresponding 3D structure (hyperlinked to the PDB). The web version of RefDB also supports a local BLAST sequence search (Altschul et al., 1997) as well as a fast boolean keyword query system supported by GLIMPSE (Manber and Wu, 1994; Manber and Bigot, 1998). This allows users to search RefDB via the sequence, partial sequence, protein name, author name, accession number, chemical shift or any other keyword or combination of keywords. All corrected protein chemical shift files archived in RefDB adhere to the BMRB star format, with the SHIFTCOR analysis placed at the top of each file as a set of comments. Individual files can be downloaded separately via the web. RefDB is updated weekly via the UPDATE program.

Results and discussion


At the time of this writing, RefDB consists of 459 different proteins out of a total of 2400 macromolecular entries and ~600 fully (>80% complete) assigned, non-redundant proteins in the BioMagResBank. Of these 459 proteins, 87 contain only ^1H assignments, 67 have ^{15}N and ^1H assignments and 305 proteins have ^1H , ^{13}C and ^{15}N assignments. Of those proteins with reported ^{13}C assignments, 98.7% of these entries have at least $^{13}\text{C}\alpha$ shift assignments, 86.9% have both $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ shift assignments while just 59.3% have ^{13}CO shift assignments. A total of 226 proteins have at least one corresponding X-ray structure while 233 have only NMR derived struc-

Re-referenced Chemical Shifts Database - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print

Address <http://redpoll.pharmacy.ualberta.ca/RefDB/> Go Links



Re-referenced Protein Chemical Shift Database (RefDB)

The Re-referenced Protein Chemical shift Database (RefDB) is a database of carefully corrected or re-referenced chemical shifts, derived from the BioMagRes Bank. The process involves predicting protein ^1H , ^{13}C and ^{15}N chemical shifts using X-ray or NMR coordinate data via SHIFTX and then comparing those predictions to the observed shifts reported in the BMRB (via SHIFTCOR). RefDB provides a standard chemical shift resource for NMR spectroscopists, wishing to derive or compute chemical shift trends in peptides and proteins.

RefDB currently has 459 re-referenced protein chemical shift files.

To download the complete RefDB database as a single file (30 MB), click [here](#)

To view the list of RefDB entries with their corresponding PDB files, click [here](#).

To download the RefDB in SHIFTY format with the secondary structure information for each protein sequence, click the following links.
[RefDB-C.db](#) [RefDB-H.db](#) [RefDB-N.db](#)

Search the RefDB database using a keyword (You may use the boolean search with AND, OR, or '*'.):

Search the RefDB database using a sequence (FASTA format or raw sequence):

[RefDB statistics](#)

Internet

Figure 3. A screenshot of the RefDB database server.

tures. The smallest protein in RefDB is 25 residues (PDB 1K09; bmr5183) and the largest is 731 residues (PDB 1D8C; bmr5471). There are a total of 366,669 re-referenced ^1H , ^{13}C and ^{15}N chemical shift assignments in the RefDB database, of which 188,824 are backbone assignments. Statistics concerning the size and composition of RefDB are updated weekly and posted at the RefDB web site.

Chemical shift referencing – Limits and limitations

The principle purpose of RefDB and its associated programs is to serve as a resource for biomolecular spectroscopists to check, verify and/or correct chemical shift assignments and chemical shift referencing for peptides and proteins. As such it is critical to know that the methods used by RefDB are robust, justified and fully validated. It is also important to understand the inherent limitations and appropriate applications of the database and its chemical shift re-referencing protocols. A key point to remember throughout all of this discussion is that chemical shift referencing and, therefore, chemical shift assignments are inherently imperfect. As pointed out in the work that first described the current IUPAC recommendations for indirect referencing, instrumental differences can contribute to a systematic reference offset difference of at least 0.07 ppm for ^{13}C and ^{15}N shifts (Wishart et al., 1995). Furthermore, the digital resolution for many heteronuclear experiments is such that ^{13}C and ^{15}N chemical shift assignments typically have a precision averaging about ± 0.15 ppm (BMRB reports a range 0.05 to 0.47 ppm) – which, for an average protein of 100 residues, can translate to a systematic ‘referencing error’ of up to ± 0.03 ppm. Likewise, differences in data processing methods, data processing software and intrinsic operator bias can also lead to small (~ 0.1 ppm) systematic variations in ^{13}C and ^{15}N chemical shift assignments. In addition, variations from sample to sample (pH, salt, local structure), or temporal changes within the same sample (temporal changes in pH, temporal changes in structure) along with weak binding of the DSS (or TSP) to the protein of interest can contribute to uncertainties and systematic errors in chemical shift assignments and chemical shift referencing (Lam and Kotowycz, 1977). Overall, it is not unreasonable to suggest that if two individuals in two different laboratories using two different software packages were to analyze and assign the same protein sample using an identical IUPAC referencing protocol, they could end up with ^{13}C and ^{15}N as-

signments that might systematically differ by 0.1 to 0.2 ppm. Indeed, it is our experience that even the same sample analyzed by the same individual on the same instrument, but with data collected at a different time (1–2 months apart) will show small systematic differences in ^{13}C and ^{15}N assignments (± 0.1 ppm) and in ^1H assignments (± 0.02 ppm).

The point of this discussion is to illustrate that instrumental imprecision can lead even the most careful of experimenters to ‘effective’ chemical shift referencing differences on the order of 0.2 ppm for ^{13}C and ^{15}N chemical shifts. Similar reasoning by Cornilescu and Bax (1999) suggested a 0.3 ppm tolerance (assuming an average length of 100 residues) was acceptable. Therefore in addressing the issue of chemical shift referencing it seems that a fair estimate for the experimental or ‘effective’ error in chemical shift referencing, is about ± 0.2 ppm for ^{13}C and ^{15}N chemical shifts and ± 0.02 ppm for ^1H chemical shifts. This means that measured or calculated chemical shift referencing errors of this order cannot be considered statistically or even experimentally significant. On the other hand, it has been argued that chemical shift referencing errors on the order of ± 0.5 ppm for ^{13}C and ^{15}N chemical shifts and ± 0.10 ppm for ^1H chemical shifts are significant as they adversely affect secondary structure identification and subsequent structure refinement (Wishart and Sykes, 1995; Wishart and Case, 2001). Furthermore, based on the measured precision of SHIFTX and SHIFTCOR it appears that reference offsets calculated in RefDB have a precision of approximately ± 0.05 , ± 0.22 , ± 0.23 , ± 0.19 , ± 0.47 , and ± 0.12 ppm for $^1\text{H}\alpha$, $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO , ^{15}N and ^1HN shifts respectively (using a 95% confidence interval). Based on these three lines of reasoning (intrinsic experimental errors, concern over structure identification by chemical shifts and the limitations of SHIFTX calculations), we decided to classify reference offsets with calculated absolute magnitudes > 0.1 ppm for $^1\text{H}\alpha$, > 0.2 ppm for ^1HN , > 0.5 ppm for $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$ and ^{13}CO and > 1.0 ppm for ^{15}N as referencing errors (see Table 2). Referencing errors of this magnitude, if left uncorrected, would be sufficient to mask out structurally important secondary shifts, alter the identification of secondary structures or significantly reduce the accuracy of any chemical shift refinement scheme.

Table 2. The number of proteins and their associated ranges of referencing errors (in ppm) for ^{13}C and ^{15}N chemical shifts. Referencing errors of less than 1 ppm for ^{15}N shifts are not considered significant

Referencing error (ppm)	$^{13}\text{C}\alpha$	$^{13}\text{C}\beta$	^{13}CO	^{15}N
0.5–1.0	18	39	16	(93)
1.0–1.5	18	10	7	47
1.5–2.0	18	17	11	24
>2.0	10	13	6	22
Total	64	79	40	93

Table 3. The maximum range of SHIFTCOR-calculated offsets (in ppm) determined over an ensemble of NMR structures

BMRB ID (# structures)	$^{13}\text{C}\alpha$ (ppm)	$^{13}\text{C}\beta$ (ppm)	^{13}CO (ppm)	^{15}N (ppm)	^1HN (ppm)	$^1\text{H}\alpha$ (ppm)
4184 (20)	0.17	0.16	0.22	0.41	0.11	0.08
4574 (15)	0.17	0.30	0.20	0.87	0.15	0.08
4593 (20)	0.32	0.26	N/A	0.73	0.12	0.08
4599 (20)	0.19	0.32	0.25	0.62	0.10	0.03
4619 (32)	0.24	0.22	N/A	0.65	0.10	0.10
Average	0.22	0.25	0.22	0.66	0.12	0.07

Accuracy and precision of SHIFTCOR reference offsets

As described earlier, SHIFTCOR uses the mean difference between SHIFTX-calculated shifts and observed chemical shifts to determine reference offsets. Based on a standard *t*-test and the estimated error in the SHIFTX calculations, it is possible to determine the reliability or confidence limits of each of these calculated reference offsets. These values are determined for each protein entry and for each nucleus and presented in every RefDB file as part of that entry's RefDB header. Typically these reference offset errors are on the order of ± 0.10 for ^1H , ± 0.20 for ^{13}C and ± 0.50 ^{15}N shifts, respectively.

However, it is also appropriate to ask whether these quoted error limits encompass other influences related to structural variations, ligand binding, or PDB selection strategies. For instance, how are the offsets affected by the resolution of the chosen structure? Do low-resolution structures have greater offsets than high-resolution structures of the same protein?

Table 4. Number and type of assignment-related errors (459 proteins)

Type of error	$^{13}\text{C}\alpha$	$^{13}\text{C}\beta$	^{13}CO	^{15}N	^1HN	$^1\text{H}\alpha$
Mis-assignment	70	206	19	54	27	N/A
Labelling/Typographical	2	3	2	4	1	1
Struct difference	8	8	N/A	N/A	N/A	318
Switches	13	5	N/A	N/A	N/A	N/A
Spectral folding	N/A	N/A	188	25	N/A	N/A
All categories	93	222	209	83	28	319
Total assignments	32104	25526	20036	34915	40414	35829

Do ensembles of NMR structures show considerable individual variance in their calculated offsets? Do ligands such as DNA, bound proteins or bound co-factors influence the calculated offsets significantly? How unreliable are the predictions if a 'bad' protein selection is made (i.e., the NMR assignments are for a ligand-free molecule yet the selected structure is of a ligand-bound molecule)?

To look at the influence of resolution and structure selection on reference offset calculations we applied SHIFTCOR to a number of proteins that had multiple (>40) PDB files, covering a wide range of resolutions (1.0 to 3.31 Å) and crystal forms, with a number of different ligands (with and without calcium, with and without bound antibodies, with and without trypsin, etc.). These proteins included myoglobin (bmr4061 – 131 structures), staphylococcal nuclease (bmr4052 – 47 structures), BPTI (bmr5359 – 59 structures) and human lysozyme (bmr5142 – 172 structures). The maximum difference or range between the maximum and minimum reference offsets calculated by SHIFTCOR was then determined for each nucleus. For myoglobin the range was 0.67 ppm for $^{13}\text{C}\alpha$, 0.12 ppm for $^{13}\text{C}\beta$, 0.59 ppm for ^{15}N and 0.31 ppm for ^1HN . For staphylococcal nuclease the range was 0.05 ppm for $^1\text{H}\alpha$, 0.24 ppm for $^{13}\text{C}\alpha$, 0.26 ppm for $^{13}\text{C}\beta$, 0.52 ppm for ^{13}CO , 0.72 ppm for ^{15}N and 0.20 ppm for ^1HN . For BPTI the range was 0.10 ppm for $^1\text{H}\alpha$, 0.58 ppm for $^{13}\text{C}\alpha$, 0.37 ppm for $^{13}\text{C}\beta$, 0.35 ppm for ^{13}CO , 1.01 ppm for ^{15}N and 0.20 ppm for ^1HN . For human lysozyme the range was 0.05 ppm for $^1\text{H}\alpha$, 0.25 ppm for $^{13}\text{C}\alpha$, 0.41 ppm for $^{13}\text{C}\beta$ ppm, 0.20 ppm for ^{13}CO , 0.45 ppm for ^{15}N and 0.18 ppm for ^1HN . Interestingly, these ranges correspond closely to the range of the 95% confidence limits determined by the

t-test. Hence, the choice of the structure file does not unduly influence the chemical shift reference offset calculations.

To assess the influence of different levels of resolution within the same structure, plots were generated between the calculated offsets and the reported X-ray resolution for each of myoglobiin, BPTI, nuclease and human lysozyme. No correlation ($r < 0.1$) could be detected between the calculated offsets and the resolution of any of the four proteins or individual nuclei (data not shown). On the other hand, the standard deviations increased slightly and the correlation coefficients generally dropped with decreasing resolution. Overall, it can be concluded that the calculated offsets are not significantly affected by the resolution of the chosen structure. Furthermore, among the PDB files containing large (protein) ligands, small (metal) ligands, or no ligands at all, no significant (>0.2 ppm) differences in the calculated offsets could be seen for ^1H , ^{13}C or ^{15}N nuclei. This is not to say that some proteins are not profoundly changed by the addition of ligands (esp. DNA binding proteins), however, proteins with substantially different structures than the structures used to generate the corresponding NMR assignments are easily identified (or discarded) through the correlation coefficient checks that RefDB and UPDATE currently use.

To assess the variations seen in reference offsets calculated for ensembles of NMR structures, we selected five proteins (bmr4184, bmr4574, bmr4593, bmr4599 and bmr4619), each of which had been solved by NMR methods and each of which had a minimum of 10 structures in its structural ensemble files (corresponding to 1F8H, 1D4B, 1D1D, 1CKR, 1E0E). The range between the maximum and minimum reference offsets calculated by SHIFTCOR was then determined for each nucleus. The results are shown in Table 3 along with the global average for each nucleus. As might be expected, these ranges are well within (by approximately 1/2) the range of the 95% confidence limits determined by the *t*-test. Hence, the choice of the individual structure within an NMR structural ensemble does not have a statistically significant influence the chemical shifts reference offset calculations.

As an additional check of the robustness or limitations of SHIFTCOR, we looked at the differences between the SHIFTCOR offsets calculated from a series of structurally similar (or identical) proteins assigned by the same laboratory. In other words we looked at the concordance of offset calculations be-

tween multiple BMRB files with a single PDB file. Ideally, if the laboratory used the same referencing protocols and the structures were essentially identical, the calculated offsets for ^{13}C and ^{15}N shifts should be within ~ 0.2 ppm of each other. Using NMR data for two troponin C variants (bmr4232, 4401, PDB IAVS), two calmodulin variants (bmr547, 4270, PDB 1MXE) and two maltose binding protein variants (bmr4354, 4987, PDB 1MPB), we determined the difference between each pair of reference offsets (calculated by SHIFTCOR) for each nucleus. The range, averaged over these three proteins, was 0.04 ppm for $^1\text{H}\alpha$, 0.08 ppm for $^{13}\text{C}\alpha$, 0.05 ppm for $^{13}\text{C}\beta$, 0.07 ppm for ^{13}CO , 0.20 ppm for ^{15}N and 0.02 ppm for ^1HN offsets, which confirms the reproducibility of the SHIFTCOR reference offset calculations.

These results clearly demonstrate that the 95% confidence intervals assigned to the reference offsets calculated by SHIFTCOR are sufficiently broad to cover file selection errors or unexpected chemical shift influences related to structural variations, ligand binding, or general PDB selection strategies used in the UPDATE program.

Classifying chemical shift referencing and assignment errors

Of particular interest for this study was a precise determination of the magnitude and extent of chemical shift errors or problems in the BioMagResBank. Based on previous experience, we identified six types of potentially classifiable chemical shift ‘errors’ including: (1) Referencing errors; (2) typographical errors; (3) assignment switches; (4) mis-assignments; (5) mis-assignments due to spectral folding; and (6) structural discrepancies. With the possible exception of referencing errors, the latter four types of chemical shift errors have to be inferred on the basis of manual inspection or ‘reconstruction’ of the assignment process. Some of these errors are easily identified, while others are far more subtle. For instance, the addition or deletion of digits or decimal points (84.0 vs. 8.40 for a ^1HN shift) is an obvious typographical error, whereas the exchange of two digits (8.34 vs. 8.43 for a ^1HN shift) is almost undetectable. As a general rule, if we couldn’t classify a chemical shift anomaly as either a typographical error, an assignment switch/exchange, a spectral folding problem or a mis-assignment, we would attribute it to a structural discrepancy (solution vs. solid state).

Referencing errors

Referencing errors or referencing adjustments are systematic errors arising from the improper referencing of ^1H , ^{13}C or ^{15}N chemical shifts. Most NMR spectroscopists are quite diligent in their chemical shift referencing protocols. However, even the most careful worker can make mistakes. These mistakes may arise from (1) incorrect instrument settings; (2) data processing errors; (3) sample preparation or decay; (4) failure to account for isotopic shifts; (5) failure to adhere to or failure to understand IUPAC/IUB referencing protocols; (6) use of obsolete referencing standards (TMS, NH_4Cl , H_2O); or (7) the use of certain shift-biasing NMR techniques (TROSY, for example). These kinds of systematic errors are of considerable concern in biomolecular NMR because they can affect nearly every chemical shift assignment. Furthermore, they can often be sufficiently large to make almost all secondary shifts undetectable or misleading (Wishart et al., 1995). What is most frustrating is that these types of chemical shift errors, particularly for ^{13}C and ^{15}N nuclei, have often been exceedingly difficult to identify.

For the purposes of this paper we will refer to two types of referencing problems: (1) Mis-referencing and (2) improper referencing. Mis-referencing refers to situations where the reported chemical shift assignments do not appear to be consistent with author-reported chemical shift referencing protocols. Improper referencing refers to situations where the authors have not adhered to IUPAC/IUB recommendations (Markley et al., 1998). In this particular study we investigated the occurrence of referencing errors for each type of nucleus (^1H , ^{13}C and ^{15}N) separately.

The first set of shifts we analyzed in RefDB was the ^1H shifts. Because almost all ^1H chemical shifts are determined using an internal primary reference (DSS, TSP) or a well characterized secondary reference (HDO) one would not expect to find any significant ^1H shift referencing errors. Indeed the data in RefDB bear this out as we found essentially no significant referencing errors among ~ 445 sets of ^1H assignments. The largest difference between any set of observed and predicted $^1\text{H}_\alpha$ shifts (i.e., the reference offset) was 0.29 ppm (bmr5102) with the vast majority (>90%) of $^1\text{H}_\alpha$ referencing offsets being less than 0.10 ppm. Similarly, the vast majority (>88%) of calculated ^1HN reference offsets were less than 0.20 ppm. The small proportion of ^1H shifts that appear to be mis-referenced can be adequately explained

by the presence of systematic temperature, isotope, solvent or pH effects (Wishart and Sykes, 1994a). On the other hand, because of the long-standing confusion over how to indirectly (or directly) reference ^{13}C or ^{15}N shifts, we found there were many more significant problems with these shifts. For instance, 93/345 (26.9%) of protein entries with ^{15}N assignments required reference adjustments (up or down) of more than 1 ppm. Furthermore, 64/291 (22.0%), 79/266 (29.7%) and 40/181 (22.1%) of protein entries in RefDB required reference adjustments of more than 0.5 ppm for their reported $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$ and ^{13}CO assignments, respectively.

Although there are many proteins that required chemical shift adjustments, most of these re-referencing changes fell into the range of 0.5–1.0 ppm for ^{13}C shifts and 1.0–1.5 ppm for ^{15}N shifts (Table 2). As might be expected, the relative frequency of these $^{13}\text{C}/^{15}\text{N}$ chemical shift referencing errors falls off exponentially relative to their magnitude. The largest referencing adjustment required for ^{13}C shifts was 2.95 ppm (bmr4431), whereas the largest ^{15}N chemical shift adjustment was 6.39 ppm (bmr4242). Among those proteins identified as requiring significant adjustments were twelve proteins which were fully deuterated (bmr4354, bmr4775, bmr4836, bmr4897, bmr4936, bmr4987, bmr5161, bmr5182, bmr5208, bmr5209, bmr5355, bmr5471). Since the ^{13}C and ^{15}N chemical shifts of deuterated proteins are shifted upfield (0.43 ppm for $^{13}\text{C}_\alpha$, 0.82 ppm for C_β , and 0.23 ppm for ^{15}N) relative to those expected for a fully protonated sample (Gardner et al., 1997; Bjorndahl et al., 2001), these chemical shift differences should not be classified as referencing errors. Indeed, their reported chemical shift displacement suggests that all seven samples were correctly referenced according to IUPAC conventions.

Figure 4 plots the frequency of chemical shift referencing errors for $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$, ^{13}CO and ^{15}N assignments versus the year of reporting/deposition. As can be seen from these graphs, heteronuclear chemical shift referencing problems were especially widespread prior to 1995. Almost all of these referencing errors could be classified as ‘improper’ referencing (i.e., not adhering to the IUPAC recommendations). After 1995 there appears to be a significant improvement, indicating a stricter adherence to IUPAC/IUB ^{13}C and ^{15}N chemical shift referencing recommendations (Wishart and Sykes, 1994a; Wishart et al., 1995; Markley et al., 1998). Interestingly, more than five years after the recommendations were first made, we still see that

Table 5. Averaged ^{13}C chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	52.84	(1.64)	54.83	(1.05)	51.53	(1.48)	53.44	(1.91)
Cys(ox)	55.60	(2.58)	58.00	(2.54)	54.99	(2.00)	56.02	(2.67)
Cys(red)	57.53	(3.05)	61.31	(3.50)	56.88	(2.02)	58.40	(3.32)
Asp	54.18	(1.60)	56.70	(1.61)	53.87	(1.64)	54.90	(2.01)
Glu	56.87	(1.82)	59.11	(1.16)	55.52	(1.67)	57.66	(2.09)
Phe	57.98	(2.02)	60.81	(1.90)	56.65	(1.59)	58.43	(2.57)
Gly	45.51	(1.05)	46.91	(1.10)	45.22	(1.17)	45.63	(1.18)
His	55.86	(1.96)	59.04	(1.74)	55.09	(1.78)	56.65	(2.44)
Ile	61.03	(1.90)	64.57	(1.74)	60.05	(1.57)	61.89	(2.65)
Lys	56.59	(1.78)	58.93	(1.44)	55.40	(1.34)	57.12	(2.11)
Leu	54.92	(1.70)	57.52	(1.23)	54.08	(1.31)	55.78	(2.06)
Met	55.67	(1.54)	58.09	(1.81)	54.58	(1.24)	56.58	(2.20)
Asn	53.23	(1.51)	55.45	(1.42)	52.74	(1.47)	53.69	(1.82)
Pro	63.47	(1.26)	65.49	(1.08)	62.64	(1.03)	63.61	(1.46)
Gln	56.12	(1.72)	58.47	(1.19)	54.83	(1.41)	56.77	(2.05)
Arg	56.42	(1.94)	58.93	(1.55)	55.14	(1.64)	57.11	(2.29)
Ser	58.38	(1.69)	60.88	(1.61)	57.54	(1.40)	58.74	(2.01)
Thr	61.64	(2.07)	65.61	(2.39)	61.06	(1.59)	62.31	(2.65)
Val	62.06	(2.16)	66.16	(1.55)	60.83	(1.64)	62.82	(2.90)
Trp	57.78	(1.71)	60.01	(1.77)	56.41	(1.87)	58.05	(2.34)
Tyr	57.97	(2.17)	60.98	(1.76)	56.83	(1.71)	58.21	(2.52)
Total number of chemical shifts	8466		8003		7055		24900	

approximately 20% of newly deposited protein entries are incorrectly referenced. Closer investigation of these entries (esp. from 1999–2002) indicates that the vast majority (>80%) of these problem entries can be ascribed to ‘mis-referencing’. That is, the IUPAC referencing scheme that the authors claim to be using is not being properly implemented. This suggests that chemical shift referencing is still problematic for a significant number of individuals in the biomolecular NMR community.

Outside of improved education, improved lab practices and stricter rules about adherence to IUPAC recommendations, it may be that the best approach to dealing with this chemical shift referencing problem is to make greater use of computer programs such as SHIFTCOR or TALOS as an integral part of the data checking/validation process prior to submitting or accepting data at the BMRB. Currently, the BMRB uses data checking routines to flag shifts that are more than five standard deviations away from the average value for a given atom or nucleus. Depositors have 7 days

to submit additions or corrections after which the entry is released. Similar data checking and validation procedures for NCBI sequence submission and PDB coordinate submission/evaluation are also starting to appear (Berman et al., 2000; Vaguine et al., 1999; Laskowski et al., 1993). Indeed, given the spreading variability in quality and the exponential growth in quantity, it appears that the development of data validation and data checking programs have become a major thrust for just about every major biological or bioinformatic database (Vaguine et al., 1999).

Mis-assignments, typos and other errors

While our principle concern was to develop software tools and methods to identify and fix chemical shift referencing errors, we found that other chemical shift errors could also be detected. Indeed, as Williamson et al. (1995) has already pointed out, accurate, structure-based chemical shift calculations can be used quite effectively to identify ^1H assignment errors.

Table 6. Averaged $^{13}\text{C}\beta$ chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	19.06	(1.26)	18.26	(0.88)	21.14	(2.05)	19.22	(1.78)
Cys(ox)	40.97	(3.19)	39.44	(2.86)	43.90	(4.18)	41.62	(3.95)
Cys(red)	29.35	(2.52)	27.75	(2.07)	30.16	(1.97)	29.14	(2.33)
Asp	40.85	(1.32)	40.51	(1.33)	42.30	(1.62)	41.03	(1.50)
Glu	30.20	(1.55)	29.37	(0.99)	32.01	(1.98)	30.19	(1.74)
Phe	39.45	(1.98)	38.78	(1.31)	41.54	(1.74)	40.08	(2.09)
Gly	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)
His	29.97	(2.42)	29.54	(1.46)	31.85	(2.22)	30.29	(2.23)
Ile	38.65	(1.69)	37.60	(1.15)	39.86	(1.98)	38.81	(1.93)
Lys	32.79	(1.67)	32.27	(0.88)	34.63	(1.78)	33.09	(1.72)
Leu	42.38	(1.64)	41.65	(1.05)	43.79	(2.00)	42.52	(1.80)
Met	33.36	(2.26)	32.27	(1.66)	35.05	(2.29)	33.25	(2.28)
Asn	38.55	(1.41)	38.61	(1.31)	40.12	(2.07)	38.93	(1.66)
Pro	31.94	(0.95)	31.46	(0.95)	32.27	(1.20)	31.94	(1.02)
Gln	29.14	(1.69)	28.51	(0.92)	31.28	(1.93)	29.39	(1.80)
Arg	30.66	(1.67)	30.14	(1.14)	32.19	(1.80)	30.83	(1.68)
Ser	64.03	(1.27)	63.08	(1.12)	65.16	(1.51)	64.15	(1.50)
Thr	70.12	(1.33)	68.88	(1.17)	70.75	(1.51)	70.07	(1.54)
Val	32.71	(1.37)	31.49	(0.72)	33.91	(1.61)	32.87	(1.68)
Trp	29.67	(1.74)	29.30	(1.40)	31.50	(1.70)	30.23	(1.88)
Tyr	38.95	(1.84)	38.25	(1.11)	40.97	(1.85)	39.71	(2.02)
Total number of chemical shifts		6115		6313		5870		19309

Unlike systematic referencing errors, these ‘random’ errors are not easily classified (or identified) without manual inspection or some prior knowledge about the nuances of the NMR assignment process. Similarly, the correction of these errors also requires manual intervention.

Table 4 provides a summary of the number and type of assignment-related errors that were manually identified. It must be emphasized that these are ‘probable’ errors as we cannot confirm their origin or cause without access to the raw experimental data. In all likelihood this is an underestimate of the true number of errors in the data set. As seen in this table, we identified 12 apparent typographical errors in 7 different proteins (bmr915, bmr1673, bmr4115, bmr4452, bmr4894, bmr4909, and bmr4984.). We also found 17 instances of $^{13}\text{C}\alpha/^{13}\text{C}\beta$ switches (bmr4068, bmr4232, bmr4599, bmr4726, bmr4740, bmr4979 and bmr5077) for threonine and serine. These assignment switches are quite understandable in light of the unusual downfield $^{13}\text{C}\beta$ shifts for serine and threonine and their

proximity to $^{13}\text{C}\alpha$ values. For those ^{13}C and ^{15}N shifts that differed by more than 5–6 standard deviations from the predicted values, but which fell within the allowed range of ^{13}C or ^{15}N shifts (regardless of amino acid type), we classified as ‘mis-assigned’. Clearly, some of these resonances may be correctly assigned and that their substantive differences arose from structural effects or our imperfect understanding of chemical shift principles. Based on the data collected for Table 4, we estimate that nearly 40% (181/459) of all protein entries in the BMRB have at least one mis-assignment. While this may seem high, if one calculates the actual fraction of mis-assignment relative to all reported assignments, their level of abundance ($\sim 0.3\%$) is not at all unreasonable. Indeed, it is well within the expected or acceptable frequency of expected mis-assignments. Interestingly, our analysis shows that the frequency of mis-assignments over the past 15 years appears to be remarkably constant.

While it was relatively easy to identify ^{13}C and ^{15}N mis-assignments, it was essentially impossible to

Table 7. Averaged ^{13}C O chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	177.67	(1.57)	179.40	(1.32)	176.09	(1.51)	178.16	(1.99)
Cys	174.93	(1.89)	176.16	(1.64)	173.57	(1.64)	174.76	(2.01)
Asp	176.31	(1.34)	178.02	(1.33)	175.54	(1.57)	176.69	(1.66)
Glu	176.43	(1.36)	178.61	(1.21)	175.35	(1.40)	177.25	(1.87)
Phe	175.59	(1.60)	177.13	(1.38)	174.25	(1.63)	175.65	(1.99)
Gly	173.89	(1.42)	175.51	(1.23)	172.55	(1.58)	173.97	(1.63)
His	174.83	(1.72)	176.98	(1.29)	174.17	(1.54)	175.34	(1.94)
Ile	175.57	(1.67)	177.72	(1.29)	174.86	(1.39)	176.05	(1.90)
Lys	176.34	(1.43)	178.40	(1.46)	175.31	(1.29)	176.85	(1.89)
Leu	176.89	(1.71)	178.53	(1.30)	175.67	(1.47)	177.26	(1.91)
Met	175.35	(1.89)	177.95	(1.12)	174.83	(1.40)	176.67	(2.00)
Asn	175.08	(1.51)	176.91	(1.55)	174.64	(1.65)	175.47	(1.78)
Pro	176.89	(1.34)	178.34	(1.45)	176.18	(1.40)	177.01	(1.53)
Gln	175.90	(1.52)	177.97	(1.29)	174.88	(1.38)	176.58	(1.87)
Arg	176.02	(1.69)	178.26	(1.43)	175.14	(1.36)	176.79	(1.98)
Ser	174.49	(1.31)	175.94	(1.39)	173.55	(1.50)	174.65	(1.66)
Thr	174.70	(1.47)	175.92	(1.15)	173.66	(1.50)	174.62	(1.65)
Val	175.66	(1.47)	177.65	(1.38)	174.80	(1.39)	175.91	(1.87)
Trp	176.15	(1.14)	178.05	(1.57)	175.41	(1.66)	176.60	(1.87)
Tyr	175.39	(1.67)	177.36	(1.40)	174.54	(1.45)	175.54	(1.89)
Total number of chemical shifts		5258		5445		4560		16216

determine whether substantial ^1H shift discrepancies arose from mis-assignments or from structural differences. Consequently, we chose to err on the side of caution and ascribed these extreme outliers to probable structural differences (solid vs. liquid state, imperfect refinement, N or C terminal changes, etc.) as opposed to mistaken assignments. An example of the kind of issues one might find when analyzing ^1H shifts is found in bmr4766 (Figure 1). As can be seen in this example, there are two near-contiguous regions exhibiting larger-than-expected deviations in their $^1\text{H}\alpha$ chemical shifts. The chemical shift differences seen for residues 76–78, likely arises from structural differences between the solution and crystal state. Specifically, the ring of Phe 76 is probably much closer to the backbone in the crystal structure than in solution, thereby leading to an upfield ring-current induced shift for nearby $^1\text{H}\alpha$ nuclei. On the other hand, the chemical shift differences seen for residues 102–109, most assuredly arises from the fact that the protein structure solved by X-ray crystallography was shorter than the protein assigned by NMR. This C-terminal truncation in the

X-ray structure likely lead to a real structural change (i.e., the loss of a helix) that is manifested in the substantially different $^1\text{H}\alpha$ chemical shifts for this region. Given the different conditions and different samples used by X-ray crystallographers relative to NMR spectroscopists, these kind of small discrepancies were not uncommon, nor were they unexpected.

Perhaps the most dramatic example of an assignment error in RefDB was found for the ^{13}C O resonances in bmr4775. While displaying good overall correlations for $^1\text{H}\alpha$ (0.719), ^{15}N (0.747) and $^{13}\text{C}\alpha$ (0.942) shifts, we found the ^{13}C O shifts were strongly negatively correlated (-0.783)! As no other protein analyzed by SHIFTCOR had shown such a negative correlation for any set of chemical shifts, we decided to investigate this situation further. On closer inspection it became obvious that the ^{13}C O spectrum for this protein must have been folded prior to its assignment (perhaps due to the use of incorrect offset pulses, a far too narrow sweep width or inappropriate data processing). Given the intrinsically narrow range of ^{13}C O shifts and the lack of any kind of characteristic

Table 8. Averaged ^{15}N amide chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	123.61	(3.77)	121.44	(2.37)	124.47	(4.39)	122.83	(4.04)
Cys	117.96	(3.88)	117.68	(3.33)	121.04	(4.53)	119.16	(4.33)
Asp	119.95	(4.41)	119.22	(2.69)	122.17	(4.40)	120.22	(4.07)
Glu	120.43	(4.05)	119.04	(2.82)	122.09	(3.95)	120.23	(3.67)
Phe	119.67	(4.65)	119.16	(3.33)	121.08	(4.45)	120.12	(4.23)
Gly	109.13	(3.91)	107.51	(2.69)	109.32	(3.94)	108.98	(3.73)
His	118.72	(4.59)	117.95	(2.63)	120.49	(4.51)	119.09	(4.09)
Ile	120.87	(5.31)	119.71	(2.88)	122.85	(4.63)	121.37	(4.46)
Lys	120.45	(4.08)	119.20	(2.64)	122.21	(4.32)	120.52	(3.85)
Leu	121.48	(4.27)	119.61	(2.73)	124.05	(4.26)	121.55	(4.13)
Met	119.66	(3.95)	118.18	(2.75)	121.66	(4.02)	119.48	(3.66)
Asn	118.22	(4.71)	117.30	(2.85)	121.58	(4.35)	118.83	(4.46)
Pro	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)
Gln	119.49	(4.11)	118.45	(2.84)	121.08	(4.13)	119.54	(3.73)
Arg	120.42	(4.33)	118.90	(2.83)	122.31	(4.25)	120.35	(3.96)
Ser	115.55	(3.95)	114.87	(2.99)	116.89	(4.02)	115.89	(3.81)
Thr	113.36	(4.98)	114.60	(3.99)	116.46	(5.00)	114.94	(4.94)
Val	119.77	(5.45)	119.19	(3.59)	121.90	(5.05)	120.62	(4.85)
Trp	120.16	(5.30)	119.84	(3.11)	122.09	(5.15)	120.90	(4.65)
Tyr	119.52	(5.10)	119.17	(2.91)	121.43	(4.78)	120.36	(4.55)
Total number of chemical shifts	8920		8498		8089		26909	

'marker' shifts (such as those seen with glycine for ^{15}N and $^{13}\text{C}\alpha$) it is not difficult to understand how this kind of error could be made nor how it could go undetected. In addition to this particularly interesting example of spectral folding, another 7 protein entries (bmr4082, bmr4153, bmr5161, bmr5123, bmr5124, bmr5125, bmr5474) also appear to suffer from some minor spectral folding problems. These likely arose due to the choice of an ^{15}N sweepwidth that was too narrow.

In its current form, the RefDB database still includes these anomalous proteins and their errors or mis-assignments. Hence, it is important for users of RefDB to exercise caution in their selection of files and to include or exclude BMRB entries based on a careful evaluation of all of the evidence presented in the RefDB header and the published description.

Re-evaluating secondary chemical shifts

While the primary purpose of this study was to identify, enumerate and correct chemical shift referencing and chemical shift assignment errors, we also wanted

to demonstrate how this 'corrected' data could be used in a more practical sense. One obvious application would be to use this data to improve upon the accuracy of chemical shift calculation routines (Iwadate et al., 1999; Osapay and Case, 1994; Wishart and Nip, 1998; Neal et al., submitted). A second application might be to improve upon secondary structure identification (Metzler et al., 1993; Wishart et al., 1992, 1995) or in dihedral angle calculation (Cornilescu et al., 1999). A third application might be in developing more accurate or consistent methods for alignment-based chemical shift prediction (Wishart et al., 1997; Potts and Chazin et al., 1998).

Rather than attempt to address all three areas here, we decided to focus on re-evaluating the so-called secondary chemical shifts or secondary-structure induced shifts associated with ^1H , ^{13}C and ^{15}N nuclei. To generate this data set, corrected chemical shifts from RefDB were assembled for each residue type along with the experimentally observed secondary structure. A specially 'cleansed' version of RefDB was prepared to consisting of reference-corrected entries that had

Table 9. Averaged $^1\text{H}\alpha$ chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	4.26	(0.33)	4.03	(0.33)	4.77	(0.55)	4.29	(0.49)
Cys	4.65	(0.39)	4.15	(0.67)	5.15	(0.51)	4.79	(0.64)
Asp	4.60	(0.28)	4.43	(0.22)	4.94	(0.40)	4.62	(0.34)
Glu	4.28	(0.33)	4.01	(0.24)	4.78	(0.49)	4.28	(0.45)
Phe	4.54	(0.47)	4.16	(0.46)	5.09	(0.46)	4.67	(0.61)
Gly	3.96	(0.35)	3.81	(0.38)	4.20	(0.60)	3.98	(0.41)
His	4.53	(0.50)	4.33	(0.34)	5.06	(0.48)	4.64	(0.52)
Ile	4.15	(0.38)	3.67	(0.33)	4.68	(0.48)	4.23	(0.60)
Lys	4.26	(0.41)	3.99	(0.30)	4.69	(0.51)	4.29	(0.49)
Leu	4.36	(0.37)	4.00	(0.34)	4.82	(0.46)	4.37	(0.52)
Met	4.38	(0.41)	4.07	(0.34)	4.96	(0.47)	4.40	(0.53)
Asn	4.66	(0.36)	4.48	(0.22)	5.06	(0.49)	4.71	(0.42)
Pro	4.37	(0.35)	4.22	(0.29)	4.60	(0.50)	4.40	(0.39)
Gln	4.26	(0.34)	3.99	(0.28)	4.80	(0.49)	4.30	(0.48)
Arg	4.24	(0.43)	3.99	(0.32)	4.74	(0.50)	4.29	(0.50)
Ser	4.47	(0.35)	4.25	(0.25)	4.91	(0.48)	4.56	(0.45)
Thr	4.45	(0.36)	4.00	(0.34)	4.86	(0.46)	4.53	(0.51)
Val	4.12	(0.41)	3.58	(0.36)	4.60	(0.48)	4.20	(0.61)
Trp	4.55	(0.48)	4.38	(0.37)	5.19	(0.50)	4.79	(0.58)
Tyr	4.52	(0.44)	4.09	(0.39)	5.10	(0.54)	4.70	(0.63)
Total number of chemical shifts	9637		7312		7956		26294	

correlation coefficients greater than 0.60 of $^1\text{H}\alpha$, 0.90 for $^{13}\text{C}\alpha$, 0.98 for $^{13}\text{C}\beta$, 0.60 for ^{13}CO , 0.60 for ^{15}N and 0.20 for ^1HN . This was done to avoid the inclusion of serious typos or assignment errors that might affect the overall calculations. The final file consisted of 309 entries and 144,373 corrected assignments. Secondary structures were calculated directly from PDB files using VADAR (Wishart et al., 1994). Because it is based on objective measures of peptide geometry, VADAR provides a far more consistent assignment of secondary structure location than those made by individual crystallographers or NMR spectroscopists. The results of these calculations are shown in Tables 5–10 where we have calculated the average characteristic shifts and standard deviations for residues in helices, beta-strands and ‘coil’ regions for $^1\text{H}\alpha$, ^1HN , $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO and ^{15}N nuclei. With 144,373 ‘corrected’ assignments, this collection represents the largest and most complete set of shifts for which this kind of calculation has been done (Wishart et al., 1991; Wishart and Sykes, 1994a; Wishart and Nip, 1998). Given that the smallest number of assignments for any one

category was at least 50 (^{13}CO assignments for tryptophans in helices), we can be quite confident about the statistics (mean, median, standard deviation, range, etc.) for these numbers.

Overall, these chemical shifts show a very good level of agreement relative to previously published sets (Wishart et al., 1991; Wishart and Sykes, 1994a; Wishart and Nip, 1998), with the possible exception of some of the less-abundant residues and/or nuclei (esp. tryptophan, methionine and histidine). Interestingly, with a much larger data set and better chemical shift referencing, the upfield/downfield trends for helices and beta sheets are now much more obvious for ^{15}N , ^1HN and ^{13}CO resonances. These trends had likely been obscured in previous studies because of the ‘noise’ arising from incorrectly referenced chemical shift assignments. Looking at the results obtained for each of the 20 different residues for any given nucleus, it is also obvious that there are certain residue-specific trends concerning the extent of the upfield/downfield shifts. These may reflect intrinsic structural limitations (restricted phi/psi or side chain chi angles) or a statis-

Table 10. Averaged ^1H N chemical shift values (in ppm) categorized according to secondary structural assignment (the standard deviation is given in parentheses)

Residue type	Coil		Helix		Beta strand		Average	
Ala	8.15	(0.72)	8.08	(0.52)	8.44	(0.76)	8.20	(0.66)
Cys	8.25	(0.71)	8.20	(0.69)	8.80	(0.64)	8.49	(0.73)
Asp	8.36	(0.62)	8.18	(0.56)	8.51	(0.61)	8.33	(0.60)
Glu	8.37	(0.68)	8.22	(0.62)	8.53	(0.62)	8.34	(0.64)
Phe	8.17	(0.83)	8.18	(0.62)	8.75	(0.72)	8.42	(0.77)
Gly	8.33	(0.78)	8.29	(0.67)	8.34	(0.86)	8.33	(0.76)
His	8.21	(0.79)	8.10	(0.56)	8.62	(0.74)	8.30	(0.73)
Ile	7.98	(0.84)	8.02	(0.52)	8.68	(0.70)	8.30	(0.75)
Lys	8.23	(0.72)	7.99	(0.56)	8.48	(0.68)	8.21	(0.68)
Leu	8.08	(0.76)	8.05	(0.54)	8.60	(0.71)	8.24	(0.70)
Met	8.18	(0.65)	8.09	(0.58)	8.64	(0.67)	8.27	(0.65)
Asn	8.40	(0.78)	8.22	(0.58)	8.60	(0.64)	8.40	(0.71)
Pro	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)	N/A	(N/A)
Gln	8.23	(0.65)	8.04	(0.55)	8.48	(0.66)	8.22	(0.62)
Arg	8.25	(0.67)	8.07	(0.55)	8.56	(0.64)	8.26	(0.64)
Ser	8.23	(0.65)	8.14	(0.56)	8.50	(0.67)	8.29	(0.65)
Thr	8.16	(0.69)	8.04	(0.51)	8.51	(0.61)	8.28	(0.65)
Val	8.04	(0.65)	8.02	(0.65)	8.62	(0.69)	8.31	(0.72)
Trp	7.92	(0.89)	8.12	(0.74)	8.59	(0.83)	8.28	(0.86)
Tyr	8.06	(0.77)	8.07	(0.62)	8.68	(0.76)	8.37	(0.79)
Total number of chemical shifts	10615		9261		9331		30745	

tical proclivity to be located in less mobile (or more mobile) regions of a polypeptide.

As indicated earlier, these tables may be of some utility in predicting chemical shifts (Wishart and Nip, 1998), in assessing preliminary chemical shift assignments, in automating chemical shift assignments (Moseley and Montelione, 1999), in identifying secondary structure (Wishart et al., 1992; Metzler et al., 1993) or evaluating nearest neighbor effects (Schwarzinger et al., 2001).

RefDB as a new model for bioinformatic databases

With the increasing movement of towards storing vast quantities of biological data on electronic databases, it is clear that data handling and data storage will become increasingly important for just about everyone in the life sciences. Given the difficulty associated with handling and assimilating so much data from so many sources, we believe that it will be important to develop new approaches for automatically handling and analyzing biological data. In our view, RefDB

may serve as a useful model for a new generation of self-updating, self-correcting bioinformatic databases. Specifically RefDB makes use of the fact that all of the data it needs can be retrieved from the web through automated data mining tools (web-bots or web-spiders), automatically checked and modified (through resident data validation/checking software) and automatically displayed or accessed (via a self-updating web interface and CGI scripts). In other words, unlike most current biological databases, RefDB was designed to function autonomously, without the need for frequent human intervention or human data entry. While the removal of the 'human factor' from the database side does have its occasional down-side (run-away processes, mix-ups due to unannounced data format changes, mis-assignment or misnaming of structures). These can be overcome by occasional checks both by the users and RefDB archivists. An example of one type of mis-assignment error that occasionally happens in RefDB occurred for two recent entries. Two sets of assignments (5358 and 5359 corresponding to trypsin-bound and free forms of BPTI, respectively)

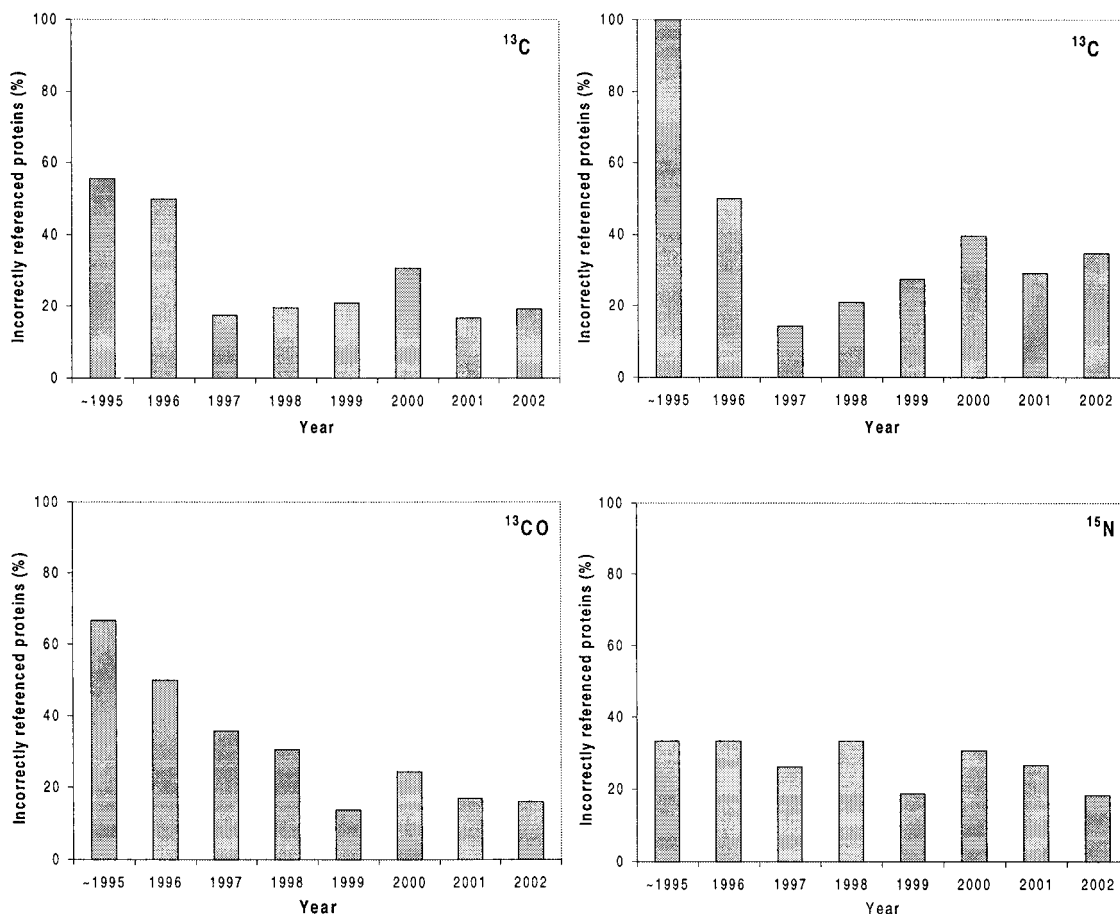


Figure 4. The percentage of referencing errors for $^{13}\text{C}\alpha$, $^{13}\text{C}\beta$, ^{13}CO and ^{15}N shifts versus year of deposition/submission.

were recently deposited into the BMRB. UPDATE selected one PDB file (1F5R) corresponding to the trypsin-bound form of BPTI and calculated the offsets for both. Technically this is an error, but had UPDATE chosen the correct forms (PDB 1F5R for 5358 and PDB 5PTI for 5359) the nucleus specific reference offsets would have only differed by 0.04 ppm for $^{13}\text{C}\alpha$, 0.10 ppm for ^{13}CO and 0.37 ppm for ^{15}N (which is statistically insignificant). Note, however, that because one entry (bmr5358) does not contain $^{13}\text{C}\beta$ chemical shifts and the other entry (bmr5359) does, SHIFT-COR averages the $^{13}\text{C}\alpha$ and $^{13}\text{C}\beta$ offsets differently, leading to an erroneous ^{13}C offset of -0.46 ppm (versus -0.15 ppm) for entry bmr5358. While this is an error that could have been prevented through careful manual curation, the net difference is relatively small (~ 0.3 ppm) and structurally insignificant. Nevertheless, this example underlines the importance for RefDB users to carefully analyze the data provided in

the RefDB header in advance of using the corrected shifts or in drawing any quick conclusions.

Despite these caveats, we have been operating and updating RefDB continuously for the better part of two years, without the need for any part-time student annotators or full-time dedicated staff members. Furthermore, the data in RefDB is never more than 1 week out of date and not particularly subject to data entry slow-downs due to staff turnover or holidays. The concept of self-updating databases is relatively new in the field of bioinformatics and yet given the abundance of web-based tools, it is something that can be relatively easily implemented. Given the increasing web accessibility of biological data, it appears that the ideas behind RefDB could be generalized to a much wider variety of biological or chemical databases.

Conclusion

There can be little doubt that chemical shifts are playing an increasingly important role in biomolecular NMR. Not only are they the 'mileposts' which map atomic structure to NMR detectable parameters, but they also provide a means for NMR spectroscopists to share and exchange raw experimental data. With the observation that chemical shifts contain a considerable amount of useful structural information, the importance of chemical shifts in biomolecular NMR has grown even further (Wishart and Sykes, 1994a; Szilagy, 1995; Case, 2000; Williamson and Asakura, 1997). However, much of the utility of chemical shifts, both for assignment and for structural purposes depends on their accuracy and reliability. Recently, this reliability has been called into question (Wishart and Case, 2001).

In this study we have demonstrated that a significant portion of ^{13}C and ^{15}N chemical shift assignments made prior to 1995 need to be re-referenced – in some cases by as much as 6 ppm. We have also demonstrated that, while NMR spectroscopists are increasingly adhering to IUPAC recommendations, ~20% of newly deposited protein chemical shifts are still mis-referenced. Furthermore, it appears that approximately 0.5% of all reported assignments may also be mis-assigned. In an effort to help sort out these persistent chemical shift referencing problems and to assist with the identification of potential mis-assignments, we have developed a self-updating database (RefDB) and a set of computational tools (SHIFTX, SHIFTCOR, UPDATE). As shown here, these tools should help correct these problems and facilitate both chemical shift analysis and chemical shift referencing.

Specifically, we believe RefDB and its associated programs could serve as: (1) A suite of programs and a set of criteria with which to assess, annotate and correct new (or old) BMRB entries; (2) a suite of programs and set of criteria with which individuals can assess and correct their own assignments and structures (during refinement, or prior to submission); (3) a resource to help test, refine and develop chemical shift prediction programs; (4) a resource with which to test, refine and develop methods to predict protein structural features (helix caps, beta turns) from chemical shift data; and (5) a resource from which accurate chemical shift dependent patterns (secondary shifts, periodicity in shifts) may be derived and useful chemical shift ranges may be calculated.

No doubt more sophisticated approaches for both chemical calculation and chemical shift validation will eventually be developed (as they need to be), however, it is our hope that RefDB and its associated software will at least initiate a concerted movement towards improving the quality of data that NMR spectroscopists deposit in the field of biomolecular NMR.

Availability

RefDB, along with web-server versions of SHIFTX and SHIFTCOR are freely available at <http://redpoll.pharmacy.ualberta.ca>.

Acknowledgements

Financial support by the Natural Sciences and Engineering Research Council (NSERC) and by the Protein Engineering Network of Centres of Excellence (PENEC Inc.) is gratefully acknowledged.

References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucl. Acids Res.*, **25**, 3389–3402.
- Archer, S.J., Vinson, V.K., Pollard, T.D. and Torchia, D.A. (1994) *FEBS Lett.*, **37**, 145–151.
- Beger, R.D. and Bolton, P.H. (1997) *J. Biomol. NMR*, **10**, 129–142.
- Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gillicand, G., Weissig, H. and Westbrook, J. (2000) *Nat. Struct. Biol.*, **7**(Suppl.), 957–959.
- Bewley, C.A., Gustafson, K.R., Boyd, R.M., Covell, D.G., Bax, A., Clore G.M. and Gronenborn, A.M. (1998) *Nat. Struct. Biol.*, **5**, 571–578.
- Bjorndhal, T.C., Watson, M.S., Slupsky, C.M., Spyropoulos, L., Sykes, B.D. and Wishart, D.S. (2001) *J. Biomol. NMR*, **19**, 187–188.
- Campbell-Burk, S., Domaille, P.J., Starovas, M.A., Boucher, W. and Laue, E.D. (1992) *J. Biomol. NMR*, **2**, 639–646.
- Case, D.A. (2000) *Curr. Opin. Struct. Biol.*, **10**, 197–203.
- Chandrasekhar, K., Campbell, A.P., Jeng, M.F., Holmgren, A. and Dyson, H.J. (1994) *J. Biomol. NMR*, **4**, 411–432.
- Copie, V., Battles, J.A., Schwab, J.M. and Torchia, D.A. (1996) *J. Biomol. NMR*, **7**, 335–340.
- Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.
- Davis, J.H., Agard, D.A., Handel, T.M. and Basus, V.J. (1997) *J. Biomol. NMR*, **10**, 21–27.
- Fogh, R.H., Schipper, D., Boelens, R. and Kaptein, R. (1995) *J. Biomol. NMR*, **5**, 259–270.
- Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997) *Biochemistry*, **36**, 1389–1401.
- Gronenborn, A.M. and Clore, G.M. (1994) *J. Biomol. NMR*, **4**, 455–458.

- Haigh, C.W. and Mallion, R.B. (1979) *Prog. Nucl. Magn. Reson. Spectrosc.*, **13**, 303.
- Hodsdon, M.E., Toner J.J., Cistola D.P. (1995) *J. Biomol. NMR*, **6**, 198–210
- Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199–211.
- Lam, Y.F. and Kotowycz G. (1977) *FEBS Lett.* **78**, 181–183.
- Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Cryst.*, **26**, 283–291.
- Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341–348.
- Manber, U. and Bigot, P. (1997) In *USENIX Symposium on Internet Technologies and Systems (NSITS'97)*, Monterey, California, pp. 231–239.
- Manber, U. and Wu, S. (1994) In *Usenix Winter 1994 Technical Conference* (best paper award), San Francisco (January 1994), pp. 23–32.
- Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 1–23.
- Maurer, T. and Kalbitzer, H.R. (1996) *J. Magn. Reson.*, **B113**, 177–178.
- Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) *Biochemistry*, **32**, 13818–13829.
- Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
- Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436–9444.
- Osapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215–320.
- Ottinger, M., Zerbe, O., Guntert, P. and Wüthrich, K. (1997) *J. Mol. Biol.*, **272**, 64–81.
- Pelton, J.G. Torchia, D.A., Meadow, N.D., Wong, C. and Roseman, S. (1991) *Biochemistry*, **30**, 10043–10057.
- Potts, B.C. and Chazin, W.J. (1998) *J. Biomol. NMR*, **11**, 45–57.
- Qin, J., Clore, G.C. and Gronenborn, A.M. (1996) *Biochemistry*, **35**, 7–13.
- Schwarzinger, S., Kroon, G.J., Foss, T.R., Chung, J., Wright, P.E. and Dyson, H.J. (2001) *J. Am. Chem. Soc.*, **123**, 2970–2978.
- Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
- Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
- Szilagy, L. (1995) *Prog. Nucl. Magn. Reson. Spectrosc.*, **27**, 325.
- Vaguine, A.A., Richelle, J. and Wodak, S.J. (1999) *Acta Cryst.*, **D55**, 191–205.
- Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52–56.
- Wagner, G., Pardi, A. and Wüthrich, K. (1983) *J. Am. Chem. Soc.*, **105**, 5948.
- Wang, A.C., Grzesiek, S., Tschudin, R., Lodi, P.J. and Bax, A. (1995) *J. Biomol. NMR*, **5**, 376–382.
- Weisemann, R., Rüterjans, H. and Bermel, W. (1993) *J. Biomol. NMR*, **3**, 113–120
- Williamson, M. P. and Asakura, T. (1997) *Meth. Mol. Biol.*, **60**, 53–69.
- Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541–546.
- Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.
- Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**, 153–163.
- Wishart, D.S. and Sykes, B.D. (1994a) *Meth. Enzymol.*, **239**, 363–392.
- Wishart, D.S. and Sykes, B.D. (1994b) *J. Biomol. NMR*, **4**, 171–180.
- Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, **6**, 135–140.
- Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333.
- Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.
- Wishart, D.S., Watson, M.S., Boyko, R.F. and Sykes, B.D. (1997) *J. Biomol. NMR*, **10**, 329–336.
- Wishart D.S., Willard L., Richards F.M. and Sykes B.D. (1994) VADAR: A comprehensive program for protein structure evaluation. Version 1.2. Edmonton, Alberta, Canada.
- Xu, X.P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321–333.
- Yamazaki, T., Hinck, A.P., Wang, Y.-X., Nicholson, L.K., Trchia, D.A., Wingfield, P.T., Stahl, S.J., Kaufman J.D., Chan, C.-H., Domaile, P.J. and Lam, P.Y.S. (1996) *Protein Sci.*, **5**, 495–506.
- Yamazaki, T., Yoshida, M., Kanaya, S., Nakamura, H. and Nagayama, K. (1991) *Biochemistry*, **30**, 6036–6047.