J Biomol NMR (2007) 39:223–227
DOI 10.1007/s10858-007-9191-5

ARTICLE

CheckShift: automatic correction of inconsistent chemical shift referencing

Simon W. Ginzinger · Fabian Gerick ·
Murray Coles · Volker Heun

Received: 24 April 2007 / Accepted: 7 August 2007 / Published online: 26 September 2007

Abstract The construction of a consistent protein chemical shift database is an important step toward making more extensive use of this data in structural studies. Unfortunately, progress in this direction has been hampered by the quality of the available data, particularly with respect to chemical shift referencing, which is often either inaccurate or inconsistently annotated. Preprocessing of the data is therefore required to detect and correct referencing errors. We have developed a program for performing this task, based on the comparison of reported and expected chemical shift distributions. This program, named CheckShift, does not require additional data and is therefore applicable to data sets where structures are not available. Therefore CheckShift provides the possibility to re-reference chemical shifts prior to their use as structural constraints.

Keywords Chemical shifts · Re-referencing · NMR

S. W. Ginzinger (✉) · F. Gerick · V. Heun
Institut für Informatik, Ludwig-Maximilians-Universität
München, Amalienstraße 17, München 80333, Germany
e-mail: simon.ginzinger@bio.ifi.lmu.de

Present Address:
F. Gerick
European Bioinformatics Institute, Wellcome Trust Genome
Campus, Hinxton, Cambridge CB10 1SD, UK

M. Coles
Lehrstuhl für Organische Chemie II, Department für Chemie,
Technische Universität München, Lichtenbergstrasse 4,
Garching 85747, Germany

Introduction

The most common approach to extracting structural information from protein chemical shifts is to compare the shifts of the target protein to a database of reference shifts. This has been applied to direct refinement of protein structures (Schwieters et al. 2003), prediction of protein secondary structure (Wishart et al. 1992; Wang and Jardetzky 2002), inference of protein backbone angles (Cornilescu et al. 1999; Neal et al. 2006), structure validation (Oldfield 1995) and detection of structural similarities in proteins (Ginzinger and Fischer 2006; Ginzinger et al. 2007). In all of these methods, the quality of the database is crucial to the outcome, in terms of its size, the accuracy of the component structures, and consistent referencing of chemical shifts. The last factor is perhaps a larger obstacle than it may first appear, due to the number of different referencing compounds and methods in current use. Even with detailed information on the method, re-referencing of shifts to a single standard is difficult. In practice, incomplete or inconsistent annotation in the main repository, the Biological Magnetic Resonance Data Bank (BMRB) (Seavey et al. 1991), often makes this impossible, and cases where re-referencing is necessary can be difficult to detect. In many cases, the magnitude of referencing errors is of the same order as structure-dependent secondary shifts, and thus all data must be checked for accurate referencing before use (Zhang et al. 2003).

Several existing programs are capable of re-referencing chemical shifts, using expectation values calculated on a residue-by-residue basis either from high-resolution structures (Neal et al. 2003; Zhang et al. 2003) or secondary structure predictions based on correctly referenced $^{1}H_{\alpha}$ shifts (Wang and Wishart 2005). Here we present a method for automatically re-referencing chemical shift data, named

CheckShift, which takes the alternative approach of comparing the global chemical shift distribution of the target protein to a reference distribution. In addition to the chemical shift values, CheckShift requires only an estimate of the overall proportion of residues in β-sheet and α-helix secondary structures, a quantity that can be reliably predicted from primary sequence (Jones 1999). CheckShift minimizes the difference between the distributions' density functions. Due to this modus operandi, CheckShift is insensitive to outlying values. We show here that CheckShift is very accurate and compares well to other structure independent methods.

Methods

Calculation of re-referencing offsets

The following steps are performed to calculate the re-referencing offset for each atom type of a set of target chemical shifts. Each will be discussed in detail below.

1. Preparation of reference density functions: Secondary shift density functions from correctly referenced data sets are prepared as a reference. This step has to be performed only once.
2. Calculation of similarity: The reference density functions are compared to the density function of the secondary shifts in the target data set.
3. Re-referencing of data sets: The previous step is iterated while changing the re-referencing offset to find the best agreement of the target and the reference. The offset that minimizes the difference between the two density functions is suggested as re-referencing offset.

Preparation of reference density functions

We have used all ^{13}C', ^{13}C$_\alpha$, ^{13}C$_\beta$ and ^{15}N chemical shifts which are included in the TALOS (Cornilescu et al. 1999) reference database (78 proteins, referenced to DSS and liquid ammonia) to prepare the reference density function. Chemical shifts from cysteine residues are excluded as they strongly depend on the oxidation state of each residue, which is a structure dependent feature that can not be predicted using sequence information alone. Although structures are available for all entries from the TALOS reference database and, thus, cysteine oxidation states are known, this is not necessarily the case for the target chemical shifts.

Subsequently, the secondary shifts for all chemical shifts from the remaining 19 amino acids are derived by subtracting the amino acid- specific random coil shifts which are used by TALOS. The secondary structure associated with each chemical shift is calculated from the corresponding protein structure using STRIDE (Heinig and Frishman 2004). Therefore, the secondary shifts can be classified according to their secondary structure. This gives rise to three separate secondary shift density functions for each atom type (see Fig. 1). Please note that the number of shifts in each distribution of an atom type is different, leading to a prior probability $\rho = (\rho_H, \rho_E, \rho_C)$ for each of the three secondary structure states.

Calculation of similarity

When predicting the re-referencing offset for each atom type of a target, the three secondary structure dependent density functions serve as the reference: these are based on the empirical chemical shifts of proteins, which are referenced according to the IUPAC standard. Target chemical shifts which are given in the standardized way are expected to have a similar density function as the reference. On the other hand, if the density functions are found to be shifted this is an indicator of a referencing error.

For the comparison, secondary shifts are derived from the target's chemical shifts, except for cysteine. Subsequently, PSIPRED (Jones 1999) is used to predict the secondary structure of the target sequence. This is done due to the fact that structures are not always available for the target sequences, and thus neither a mapping nor a defined secondary structure can be derived. While PSIPRED in general gives good predictions of secondary structures, this prediction is not used to split the secondary shifts of the target according to the secondary structure, but only to calculate the ratio $\sigma = (\sigma_H, \sigma_E, \sigma_C)$ of the three secondary

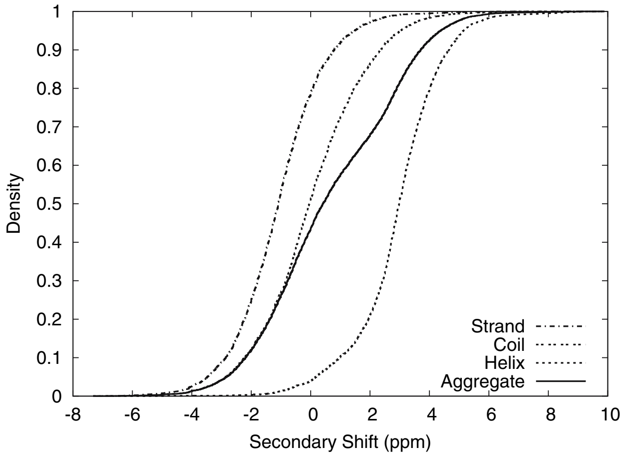

Fig. 1 Density function of ^{13}C$_\alpha$ secondary shifts from TALOS, used as expectation for secondary shifts of correctly TSP-referenced data sets. The density functions for each of the three secondary structures states (Strand, Coil, Helix) are shown together with the total density function (Aggregate)

structure states relative to each other. Later, for each of the three secondary structure states sec ∈ (H, E, C), the respective secondary structure dependent reference density function from TALOS with a prior ρ_{sec} is scaled by $\sigma_{\mathrm{sec}}/\rho_{\mathrm{sec}}$ to have the same ratio σ_{sec} as the target protein before combining and comparing them to the target's combined density function. Please note the difference between the two density functions in Fig. 2 for an illustration of this approach. This takes into account that proteins can have very different secondary structure content, having a related ratio σ that is not necessarily equal to the prior ρ from TALOS. Consequently, this leads to different expected secondary shift density functions. Additionally, this approach avoids a wrong assignment of secondary shifts to a specific secondary structure, which would occur by splitting secondary shifts based on the secondary structure prediction. Wrong prediction of secondary structure would result in inferior secondary shift density functions, consequently checking consistency to the reference distributions would be more difficult and error-prone. While PSIPRED makes correct predictions with a rate of about 83%, its strength is to correctly predict the overall architecture of whole secondary structure elements. On the other hand, the exact position of those elements is not always predicted correctly, but may vary by a few residues. Therefore, using only the information about the overall secondary structure architecture (i.e., secondary structure content), combining the three scaled density functions, and comparing the two density functions as described above, should be more accurate than using the information in a residue specific way.

Accounting for the secondary structure ratio mentioned above is done by multiplying the density functions for each secondary structure state, derived from the TALOS data set, by the ratio derived from the target protein. The final reference density function is then calculated as the sum of the three ratio-adjusted density functions.

To evaluate the quality of a certain re-referencing offset, we now calculate the averaged summed distance between the target and the reference density function. This value is inversely proportional to the quality of the proposed offset.

Re-referencing of data sets

The re-referencing is accomplished by searching for the optimal offset over a range defined by the reference distribution, using an increment of 0.1ppm. Subsequently, all chemical shifts of the data set can be adjusted by subtracting the determined offset, respectively, leading to a data set that is re-referenced to a consitent standard. For an illustration see Fig. 3. Alternatively, this offset can be used to estimate the actual referencing method for a data set, as re-referencing such data sets results in an offset that is consistent with one of the common referencing methods. The reference method that is most consistent with these offsets can be proposed and compared to the reference molecule declared in the data set.

Results & discussion

The database RefDB (Zhang et al. 2003) uses a structure dependent approach for re-referencing chemical shift data. This is done by comparing a data set to chemical shift data derived from the mapped structure using SHIFTX (Neal et al. 2003). While this approach is reported to work well and is the established standard, it is limited by the availability of structural data, which is not available for 61% of data sets from the BMRB. Furthermore, some

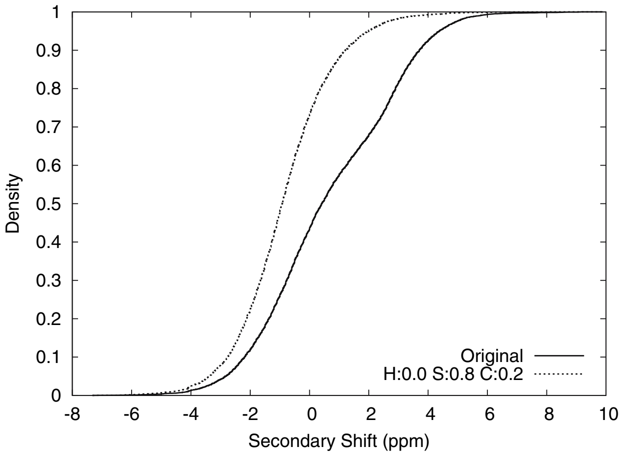

Fig. 2 The density function of the $^{13}C_{\alpha}$ secondary shifts from the TALOS data set together with the adjusted density function for a protein with 80% beta content, i.e., σ = (0.0%, 0.8%, 0.2%)

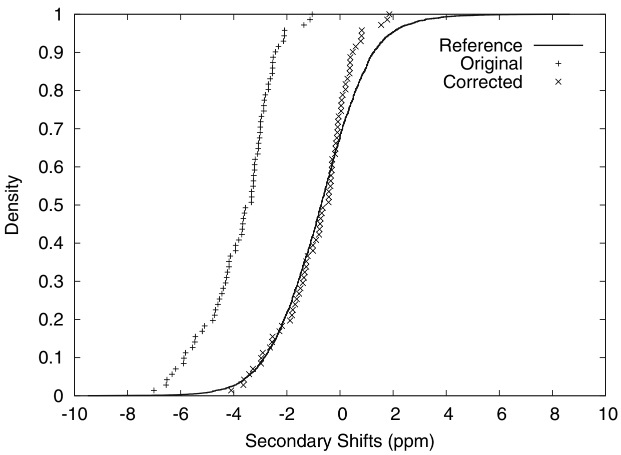

Fig. 3 Example of the density function of the target's C' shifts for a test protein, σ = (0%, 37%, 63%), (original and corrected) and the corresponding reference density function

entries in the RefDB still show inconsistencies after re-referencing due to insufficient handling of outliers; chemical shifts that differ from those predicted by SHIFTX by more than four times the expected SHIFTX RMS error (e.g., 5.0 ppm for ${}^{13}C_\alpha$) do not contribute to the average that is compared to the average of SHIFTX predictions. Therefore, data sets with many outliers are re-referenced by an offset that is too small.

Unlike the RefDB approach, Wang and Wishart (2005) introduced a protocol for adjusting inconsistently referenced chemical shifts that does not depend on structural data. ${}^1H_\alpha$ chemical shifts are used to determine the secondary structure of the protein. Sub- sequently, the re-referencing offset for each chemical shift is derived by comparison to a set of previously published averaged, secondary structure-dependent chemical shifts. These offsets are averaged for each nucleus over all residues to yield a consensus re-referencing offset for each nucleus. The re-referenced chemical shifts along with the original ${}^1H_\alpha$ chemical shifts are then used to derive the secondary structure and calculate the re-referencing offset as described before. This last step is iterated twice. CheckShift differs from Wang and Wishart (2005) in that overall shift distributions are compared, rather than individual shifts, and is therefore not sensitive to errors in secondary structure prediction for individual amino-acids.

Recently LACS (Wang et al. 2005) was developed, a method which calculates re-referencing offsets based on secondary chemical shift values alone. LACS uses linear equations to relate the difference between C_α and C_β shifts to the chemical shift value of C_α, C_β, C' and H_α. By solving these equations, the re-referencing offset for the respective atoms may be calculated. Two constraints have to be fulfilled for LACS to be applicable:

- Chemical shifts for C_α and C_β have to be available.
- C_α and C_β shifts have to be (mis-)referenced in the same way.

In comparison to LACS, CheckShift is not dependent on these constraints which proves valuable in cases where C_α or C_β shifts are missing or have been referenced differently. Additionally CheckShift calculates reference corrections for N, which is not possible using the LACS approach.

As it is often hard to check the reliability of chemical shift data, we used a set of 11 target structures (see Table 1 for details) which were provided by the group of Prof. Dr. Horst Kessler from the Technische Universität München for the performance evaluation of CheckShift.

To check the performance of CheckShift versus the method developed by Wang and Wishart and LACS, we introduced artificial referencing errors by adding an offset to the original chemical shift values. All multiples of 0.5 in the interval [–5, 5] are used as artificial re-referencing errors. This way we end up with 220 target chemical shift sets with an artificial error, plus the original 11 chemical shift sets. For each of these chemical shift sets we calculate the root mean square deviation (RMSD) between the error which was introduced and the negative re-referencing offset calculated by the respective re-referencing methods. The results of this evaluation are shown in Table 2. CheckShift strongly outperforms the re-referencing method by Wang and Wishart (2005) and performs equivalently to the LACS approach.

Table 1 Benchmark set

Name	Reference	Length	%Helix	%Sheet	%Coil
ß-ADT	Heller et al. (2004)	154	27	27	46
HAMP	Hulko et al. (2006)	54	69	0	31
KdpB	Haupt et al. (2006)	136	36	32	32
Mj0056	EMBO-J, in press	136	16	41	43
Ph1500N	Unpublished	83	13	41	46
PhS018	Coles et al. (2006)	92	22	52	26
VatN	Coles et al. (1999)	185	15	36	49
Josephin	Nicastro et al. (2005), Mao et al. (2005)	182	38	20	42

Additionally we use three unpublished chemical shift sets

CheckShift's calculations are based on a secondary structure prediction, which is of course not free of error. Therefore it is interesting to evaluate the dependence of CheckShift on a correct secondary structure assignment. We use 8 target structures from our test set, for which three-dimensional structural information is available (these are the ones listed in Table I). Now the secondary structure for these targets is calculated using STRIDE. Then a certain percentage of the secondary structure assignments is falsified randomly. This way we generate a set of targets with a secondary structure prediction correctness of 50%, 60%, 70%, 80%, 90%, and 100%. Then we evaluate CheckShift on all of these sets. The results (shown in Table 3) reveal that there is a slight dependence of the C_α and C_β corrections and essentially no influence on the C' and N corrections. This proves empirically that CheckShift is stable with respect to errors in secondary structure prediction of up to 50%.

Correct referencing of chemical shift data is vital for its further use. In the scope of this work, a re-referencing protocol was developed, which does not use structural

Table 2 RMSD of the re-referencing errors

Method	Cα	C_β	C'	N
CheckShift	0.25	0.24	0.55	0.71
Wang and Wishart (2005)	0.81	0.59	1.42	1.12
LACS	0.20	0.20	0.66	n/a

Table 3 RMSD of the re-referencing errors for different secondary structure prediction error rates

Correct (%)	C_α	C_β	C'	N
50	0.53	0.41	0.69	0.48
60	0.42	0.31	0.53	0.56
70	0.29	0.26	0.47	0.32
80	0.33	0.31	0.49	0.36
90	0.23	0.22	0.44	0.53
100	0.16	0.21	0.44	0.55

information, as opposed to established approaches. For this purpose, chemical shifts from a target protein are compared to chemical shift data from a set of correctly referenced proteins by comparing the two datasets' density functions. Subsequently, the target chemical shifts are re-referenced by applying an offset to the chemical shifts of the target. The offset that maximizes the similarity between the target- and reference chemical shift data is proposed as a re-referencing offset.

By assessing the performance of this approach, it was found the CheckShift performs very well in correcting referencing errors. CheckShift strongly outperforms another structure-independent re-referencing protocol by Wang and Wishart (2005). The comparison to LACS, a recently proposed re-referencing method, shows that CheckShift performs equivalently. Thereby CheckShift has the advantage of being able to re-reference the chemical shift for each atom independently and to give re-referencing offsets for nitrogen atoms.

Availability

CheckShift is available as a web application: http://shifts.bio.ifi.lmu.de/checkshift

On the website we also give a list of proposed reference corrections for each BMRB entry.

References

Coles M, Diercks T, Liermann J, Groger A, Rockel B, Baumeister W, Koretke KK, Lupas A, Peters J, Kessler H (1999) The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple $\beta\alpha\beta\beta$ element. Curr Biol 9(20):1158–1168

Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN (2006) Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure 14(10):1489–1498

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13(3):289–302

Ginzinger SW, Fischer J (2006) SimShift: identifying structural similarities from NMR chemical shifts. Bioinformatics 22(4):460–465

Ginzinger SW, Gräupl T, Heun V (2007) SimShiftDB: Chemical-Shift-based homology modeling'. Proceedings of the first International Conference on Bioinformatics Research and Development, BIRD 2007, Springer LNBI 4414

Haupt M, Bramkamp M, Heller M, Coles M, Deckers-Hebestreit G, Herkenhoff-Hesselmann B, Altendorf K, Kessler H (2006) The holo-form of the nucleotide binding domain of the KdpFABC complex from Escherichia coli reveals a new binding mode. J Biol Chem 281(14):9641–9649

Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32(Web Server Issue):W500–W502

Heller M, John M, Coles M, Bosch G, Baumeister W, Kessler H (2004) NMR studies on the substrate-binding domains of the thermosome: structural plasticity in the protrusion region. J Mol Biol 336(3):717–729

Hulko M, Berndt F, Gruber M, Linder JU, Truffault V, Schultz A, Martin J, Schultz JE, Lupas AN, Coles M (2006) The HAMP domain structure implies helix rotation in transmembrane signaling. Cell 126(5):929–940

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202

Mao Y, Senic-Matuglia F, Fiore PPD, Polo S, Hodsdon ME, Camilli PD (2005) Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain. Proc Natl Acad Sci U.S.A. 102(36):12700–12705

Neal S, Berjanskii M, Zhang H, Wishart DS (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. Magn Reson Chem 44:S158–S167

Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. J Biomol NMR 26(3):215–240

Nicastro G, Menon RP, Masino L, Knowles PP, McDonald NQ, Pastore A (2005) The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition. Proc Natl Acad Sci U.S.A. 102(30):10493–10498

Oldfield E (1995) Chemical shifts and three-dimensional protein structures. J Biomol NMR 5(3):217–225

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65–73

Seavey B, Farr E, Westler W, Markley J (1991) A Relational Database for Sequence-Specific Protein NMR Data. J Biomol NMR 1:217–236

Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32(1):13–22

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11(4):852–861

Wang Y, Wishart DS (2005) A simple method to adjust inconsistently ref- erenced 13C and 15N chemical shift assignments of proteins. J Biomol NMR 31(2):143–148

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(6):1647–1651

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25(3): 173–195