Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria

Citation: Fritzsching, Keith J., Mei Hong, and Klaus Schmidt-Rohr. "Conformationally Selective Multidimensional Chemical Shift Ranges in Proteins from a PACSY Database Purged Using Intrinsic Quality Criteria." J Biomol NMR 64, no. 2 (January 19, 2016): 115–130.

As Published: http://dx.doi.org/10.1007/s10858-016-0013-5

Publisher: Springer Netherlands

Persistent URL: http://hdl.handle.net/1721.1/105515

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

J Biomol NMR (2016) 64:115–130
DOI 10.1007/s10858-016-0013-5

ARTICLE

Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria

Keith J. Fritzsching[1] · Mei Hong[2] · Klaus Schmidt-Rohr[1]

Received: 23 October 2015 / Accepted: 8 January 2016 / Published online: 19 January 2016

Abstract We have determined refined multidimensional chemical shift ranges for intra-residue correlations (^{13}C–^{13}C, ^{15}N–^{13}C, etc.) in proteins, which can be used to gain type-assignment and/or secondary-structure information from experimental NMR spectra. The chemical-shift ranges are the result of a statistical analysis of the PACSY database of >3000 proteins with 3D structures (1,200,207 ^{13}C chemical shifts and >3 million chemical shifts in total); these data were originally derived from the Biological Magnetic Resonance Data Bank. Using relatively simple non-parametric statistics to find peak maxima in the distributions of helix, sheet, coil and turn chemical shifts, and without the use of limited "hand-picked" data sets, we show that ~94 % of the ^{13}C NMR data and almost all ^{15}N data are quite accurately referenced and assigned, with smaller standard deviations (0.2 and 0.8 ppm, respectively) than recognized previously. On the other hand, approximately 6 % of the ^{13}C chemical shift data in the PACSY database are shown to be clearly misreferenced, mostly by ca. −2.4 ppm. The removal of the misreferenced data and other outliers by this purging by intrinsic quality criteria (PIQC) allows for reliable identification of secondary maxima in the two-dimensional chemical-shift distributions already pre-separated by secondary structure. We demonstrate that some of these correspond to specific regions in the Ramachandran plot, including left-handed helix dihedral angles, reflect unusual hydrogen bonding, or are due to the influence of a following proline residue. With appropriate smoothing, significantly more tightly defined chemical shift ranges are obtained for each amino acid type in the different secondary structures. These chemical shift ranges, which may be defined at any statistical threshold, can be used for amino-acid type assignment and secondary-structure analysis of chemical shifts from intra-residue cross peaks by inspection or by using a provided command-line Python script (PLUQin), which should be useful in protein structure determination. The refined chemical shift distributions are utilized in a simple quality test (SQAT) that should be applied to new protein NMR data before deposition in a databank, and they could benefit many other chemical-shift based tools.

Keywords Protein chemical shift · Databases · Protein secondary structure · Data mining · PIQC · PACSY · PLUQin · SQAT

Introduction

The measurement and assignment of isotropic chemical shifts is a ubiquitous step in studies of protein structure and dynamics by NMR. As a result, one by-product of the community's efforts is an impressive collection of chemical shift assignments. A crucial resource in this effort has been the Biological Magnetic Resonance Bank (BMRB) (Ulrich et al. 2008), which has archived more than 3 million protein chemical shift assignments. The correlation of

Electronic supplementary material The online version of this article (doi:10.1007/s10858-016-0013-5) contains supplementary material, which is available to authorized users.

✉ Keith J. Fritzsching
kfritzsc@brandeis.edu

✉ Klaus Schmidt-Rohr
srohr@brandeis.edu

[1] Department of Chemistry, Brandeis University, Waltham, MA 02453, USA

[2] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

these chemical shift data with torsion angles has proven especially successful (Han et al. 2011; Shen and Bax 2013; Spera and Bax 1991). In comparison, using the chemical shift data in the database for assignment purposes is less developed. It has mainly been used to generate simple tables based on averages (Hazan et al. 2008; Wang and Jardetzky 2002b); in addition, some methods using 1D probability distributions have been developed to aid assignment (Moseley et al. 2004). We have previously shown that the multidimensional chemical shift distributions have interesting features that can be used to gain additional assignment and structural information (Fritzsching et al. 2013).

While well-structured proteins give narrow lines that are straightforward to assign, chemical shift assignment for disordered or large proteins is still tedious or even intractable. The problem is especially apparent in solid-state NMR, where lines are often relatively broad, which can lead to ambiguities in both type and sequential assignments (Tycko 2015). As a result, there will be many sets of assignments equally compatible with the experimental data, and the goal of the assignment procedure should be to identify all of these sets. Tycko and coworkers have introduced a Monte Carlo/simulated-annealing algorithm (Hu et al. 2011; Tycko and Hu 2010) that attempts to provide all possible assignment sets based on input of grouped resonance lists and the possible type assignments. A related approach inspired by this method uses a genetic algorithm for the optimization (Yang et al. 2013). The input into the algorithm consists of lists of correlated chemical shifts with possible amino-acid type assignments and definitions that link the lists. Without accurate knowledge of the chemical shift ranges of the 20 canonical amino acids, even amino-acid types with only a marginal probability of resonating at the observed chemical shifts need to be included in input (Fritzsching et al. 2013; Tycko 2015). Identifying chemical shift ranges so that all inputs into the assignment algorithms can be validated at a chosen well-defined statistical threshold was the original motivation for the current work.

To this end, we decided to improve on a previously introduced simple program called PLUQ (Fritzsching et al. 2013) that takes the input of an intra-residue chemical shift list to query the PACSY database and returns possible assignments. When only two chemical shifts are inputted, the output assignments are ranked according to the number of times (within a chosen radius in ppm) that the pair of chemical shifts is found in the database. When more than two chemical shifts are entered, the database is queried with the chemical shifts in a pairwise fashion; the assignments are then ranked by the product of the number of hits found in each residue grouping. This has shown to be quite successful in correctly predicting the assignment (and also the secondary structure) from a limited number of peaks. However, it is not ideal for generating a list of all possible assignments, because for each query one has to decide when to stop including results. Also, there is no normalization for the occurrence frequency; for example, an assignment to Ala will almost always be ranked higher than a trp assignment simply because there are approximately six times more Ala than trp data in the database.

Ideally, while making assignment hypotheses a user would be able to define a certain statistical cut-off value and more or less blindly include all assignment possibilities returned on that basis. This would allow for the elimination of many of the judgment calls in amino-acid type assignment. It is then up to the optimization algorithm to remove assignments that do not satisfy the rest of the spectroscopic constraints. To achieve this, chemical shift ranges need to be defined with a cut-off based on the distribution of chemical shifts. These distributions should be relatively free of errors and defined by as much experimental data as possible. Furthermore, it should be possible to correlate ^{1}H, ^{13}C and ^{15}N peaks not only from backbone but also from side-chain atoms.

Unfortunately, as previously recognized (Iwadate et al. 1999; Zhang et al. 2003) despite IUPAC and BMRB recommendations there are still problems with chemical shift referencing within protein chemical shift databases. Chemical shift referencing for carbon-13 is especially problematic due to the continued use of the traditional standard TMS as a primary reference instead of the more recently recommended DSS. The difference of approximately −2.4 ppm between neat TMS and a 1 % DSS solution (Saito et al. 2010) is so similar to the variations in chemical shift due to secondary structure that a referencing error is not always obvious.

Due to these problems with chemical shift referencing as well as typographical and assignment errors, many methods have been proposed to identify and correct errors before submission to chemical shift archives or to prepare data for computational tools.(Ginzinger et al. 2007, 2009; Moseley et al. 2004; Neal et al. 2003; Shen et al. 2009; Wang et al. 2005, 2010; Wang and Markley 2009; Wang and Wishart 2005; Zhang et al. 2003) These methods usually rely on previously prepared and heavily curated chemical shift databases of only a few hundred proteins (Han et al. 2011; Shen and Bax 2013). Some of these programs disregard the referencing problems. For example, the commonly used assignment validation suite (AVS) is able to flag typographical errors and misassignments by comparing chemical shifts to chemical shift averages and excluding grossly (>8σ) incorrect chemical shifts, but it cannot detect a small referencing error.

Other methods (Ginzinger et al. 2007, 2009; Wang et al. 2007; Wang and Wishart 2005; Wang and Markley 2009; Wang and Jardetzky 2002b; Zhang et al. 2003) identify offsets by determining the secondary structure of a given residue (or of the whole protein on average) and calculating the difference between its measured chemical shift and the 'expected' chemical shift in that secondary structure and amino-acid type (or the expected average value). The average difference is used as the offset. The different approaches for identifying the secondary structure of the residue under consideration are summarized below. For the typical chemical shifts, most of these methods (Moseley et al. 2004; Wang et al. 2007; Wang and Markley 2009; Wang and Jardetzky 2002b) at some point rely on averages of chemical shifts generated from carefully chosen but nonetheless older and relatively small databases. Additionally, they also assume (sometimes implicitly) that distributions of the chemical shifts for each secondary structure are Gaussian.

SHIFTCOR, the algorithm behind RefDB (Zhang et al. 2003), obtains the dihedral angles of the residue in question from the known 3D structure of the protein and relies on SHIFTX (Neal et al. 2003) for calculating the expected chemical shift, using empirical chemical-shift hyper-surfaces (Spera and Bax 1991) produced from a limited database of <200 proteins re-referenced "by hand" (Zhang et al. 2003). PSSI (Wang and Wishart 2005) uses conformation-dependent ^{1}H chemical shifts to predict secondary structure for C′, Cα, and Cβ. For determining the "expected" ^{13}C chemical shifts, secondary-structure classifications of 6100 amino acids in Wang and Jardetzky's (Wang and Jardetzky 2002a) database were used. PANAV (Wang et al. 2010) appears to use a similar approach as PSSI to find referencing errors, and identifies misassignments and typographical errors by calculating the product of the probability densities (assuming Gaussian chemical shift distributions) for the backbone chemical shifts within a residue; the joint probability density found for surrounding residues gives a score of error likelihood. LACS (Wang et al. 2005; Wang and Markley 2009) exploits the empirical correlation between δCα and δCβ and secondary structure, based on data from RefDB, which appears to derive from fewer than 400 proteins. Additionally, accurate random coil chemical shift values had to be determined from a database of 651 proteins (Wang et al. 2007).

CheckShift (Ginzinger et al. 2007, 2009) compares experimental and predicted C′, Cα, Cβ and N chemical shifts using secondary-structure dependent density functions. The protein's H:C:E ratio is predicted from the amino acid sequence using PROFphd (Rost and Sander 1994) or PSIPRED (McGuffin et al. 2000). The expected chemical shift distributions were constructed using <250 proteins from the database for TALOS (Cornilescu et al. 1999) and classified by secondary structure using STRIDE. The average offset that gives the best match was used to re-reference the experimental data.

These methods do seem to work for their intended purpose (although tests with intentionally misreferenced data indicate residual errors of 0.2–1.4 ppm, (Ginzinger et al. 2007) and persistent ^{13}C offset errors in the BMRB, see below, indicate that these re-referencing methods are not always used); PANAV (Wang et al. 2010) in particular is able to detect referencing as well as assignment errors, without the need for the solved structure. RefDB (Zhang et al. 2003) contains well-referenced chemical shifts and the data are periodically updated. Still, the assumption of Gaussian character, reliance on limited numbers of hand-picked "ideal" chemical shifts values, and the possibility of different referencing within one protein limits their use when trying to discover novel features within the chemical-shift distributions.

Additionally, most consistently referenced databases do not incorporate chemical shift information for the side-chains, even though these reflect referencing errors more clearly than the backbone shifts with their large conformation-induced variability. To our knowledge, the only exception is the database for SHIFTX2 (Han et al. 2011), which was, however, limited to <200 proteins. The PACSY (Lee et al. 2012) database, although not consistently referenced, contains the majority of chemical shift data in the BMRB from proteins with determined structures. PACSY is easily interrogated thanks to its relational design, but has not been purged of misreferenced data.

In this paper, we first continue the development of curation methods for chemical shift data. Our approach, purging by intrinsic quality criteria (PIQC), uses an unreferenced chemical-shift database for proteins with known 3D structures. It has a higher resolution than previous methods, due to the larger number of data used, and indicates that most protein chemical shift referencing is actually more accurate than previously appreciated. At the same time, it exposes, again more clearly than previous studies, that a significant fraction ($\sim$6 %) of ^{13}C spectra are misreferenced by more than -1.2 ppm. We supply generated chemical-shift statistics for all proteins in the PACSY database (i.e. for almost all BMRB proteins with solved structures). Second, we use the refined database to construct multidimensional chemical shift ranges. Some of the previously seen "lobes" in the 2D chemical shift distributions are exposed as artifacts of referencing errors, while others are shown to be real and indicative of specific torsion angles or neighboring residues. Finally, applications of the chemical shift probability density functions are demonstrated, for instance to validate type assignments. To perform this validation programmatically, a command-line Python script (PLUQin) is introduced.

Methods

A CSV-formatted version (updated June 28, 2015) of the PACSY database was downloaded and built into a MySQL database using an appropriate Python script with the MySQLdb module. It contained 1,200,207 ^{13}C, 333,133 ^{15}N, and >1400,000 ^{1}H chemical shifts from >3000 proteins with 3D structures. Chemical-shift referencing checks were only performed when more than 15 chemical shifts assigned to a given isotope (^{1}H, ^{13}C, or ^{15}N) were available for the protein. All analyses of the PACSY database, delineations of chemical shift ranges, and subsequent analyses of experimental data were performed using Python. Elements of non-standard Python that were used include: MySQLdb, NumPy (van der Walt et al. 2011), scikit-learn (Pedregosa et al. 2011), Shapely, Matplotlib (Hunter 2007), BMRB Star Parser and NMRglue (Helmus and Jaroniec 2013); these libraries are available on-line. Our library of Python code contains functionality for generating PACSY database queries, processing the returned data, and analyzing chemical shift and torsion angle data.

The constructed chemical shift regions are manipulated and stored as many-sided polygons. The regions can be efficiently and conveniently stored and interrogated using open source geographic information system tools (Open Source Geospatial Foundation 2003), originally developed for mapping applications. Experimental chemical shifts are evaluated against the regions using an efficient point-in-polygon algorithm provided by Shapely. The derived probability density functions are stored as arrays in the HDF5 file format. The PLUQin script for identifying possible type (and secondary-structure) assignments can be found at ksrlab.org/pluqin-sqat.

Results and discussion

Refining chemical shift data

For the purpose of defining chemical shift ranges, small errors in chemical shift referencing or even occasional incorrect assignments have little effect. Even occasional large errors do not significantly affect the final chemical shift ranges if appropriate smoothing parameters are chosen. However, systematic offsets on the order of the width of the chemical shift distribution can be detrimental. The 2.4 ppm difference because of misreferencing to TMS instead of DSS is in this category. To eliminate chemical shifts with this type of referencing error from the chemical shift database analysis, it is desirable to determine a relative offset for the resonances in each protein in the BMRB database.

The first step of our PIQC method relies on finding the most likely chemical shift value for every atom type. Many attempts have been made to determine these idealized chemical-shift values for amino acids in proteins. Most of the previous methods have relied on generating a "high quality" data set, for which the arithmetic mean is found and used as the representative value (Zhang et al. 2003), or almost equivalently a Gaussian distribution is assumed (Wang et al. 2007). A priori selection of high quality data sets is difficult and limits the number of proteins used, which will result in poorer statistics. To circumvent having to rely on a hand-picked data set or use only a small subset of the chemical shift data, three assumptions were made in our work: (1) the largest influence of chemical shift for each atom type is the residue type and the residue's α-helix (H), coil (C), turn (T), or β-sheet (E) secondary-structure classification. (2) Within each classification, the ideal chemical shift is the chemical shift that is most frequently observed (i.e. the mode of the distribution). As long as misreferenced data account for $\lesssim$30 % of the total, they do not affect this ideal chemical shift significantly. (3) The protein structures in the PDB are correct regardless of potential chemical shift misreferencing. It should be noted that in our final referenced database the misreferenced proteins are removed so that their potentially distorted structures do not affect the final results. In this framework, all other factors influencing chemical shift are considered as perturbations resulting in broadening of the distributions without significantly shifting the maximum. The narrow offset distributions ($\sigma < 0.25$ ppm) obtained for ^{13}C and ^{1}H data justify these assumptions.

Secondary structure classification

Conveniently, the PACSY database contains secondary-structure classifications for each residue that were generated using STRIDE (Heinig and Frishman 2004), based on the three-dimensional structure of the protein. To determine secondary structure, the lowest-numbered structure was used, since by convention it has the lowest energy. The chemical shifts of the C′, Cα, Cβ, H, Hα, NH and N sites were analyzed independently in the four secondary structure classifications helix (H), coil (C), sheet (E), and turn (T), while the remaining side chain chemical shifts were analyzed without secondary-structure classification. The original STRIDE classifications also include π-helix (I), 3_{10}-helix (G), and isolated β-bridge (B or b). These were incorporated into the other structural categories as follows: helix includes H, G, I; and sheet includes E, B, b. In the presentations of the multidimensional chemical-shift regions, coil and turn were grouped together for simplicity.

Determination of distribution modes

Instead of the arithmetic mean of each distribution, PIQC uses the mode (position of the maximum) as the representative value. Finding the mode of discrete data is equivalent to locating the value of x that maximizes the probability distribution function (PDF) $p_x(x)$ (Romano 1988). Kernel density estimation, a non-parametric method, with a Gaussian kernel was used to estimate PDFs from the discrete data. The bandwidth (sometimes referred to as a smoothing parameter) of the Gaussian kernel has a large effect on the quality of the final PDF. For robustness, the bandwidth was chosen using a grid-search and tenfold cross-validation routine (Hastie et al. 2009). Figure 1 shows three examples of how the ideal chemical shift values were determined from distributions of chemical shifts of a specific amino-acid type in a specific conformation. The chemical shift mode $\delta_{m,A}$ in each distribution p_δ of chemical shift δ with the atom type condition A (which includes the amino-acid residue type, atom type [i.e. Cα, HN, etc.] and secondary structure) is given by

$$\delta_{m,A} = argmax[p_\delta(\delta|A)]$$

Due to the large number of chemical shift values in each distribution, obtained from the data for tens of thousands of amino-acid residues in PACSY, the generated PDFs are accurate representations of the underlying distributions, see Fig. 1. The p_δ functions show 'shoulders' at lower ppm values, which we will show to be at least partially due to misreferenced data of certain proteins. Due to the asymmetry in the experimental distribution, the mean is not equal to the mode. While the arithmetic or any other mean would be affected by these compromised data, this is not the case for the maximum position or mode used in our analysis, as long as any secondary components are small.

For each of the 20 canonical amino acids, the four different secondary-structure types were used to separate C′, Cα, and Cβ chemical shift distributions, while each of the other side-chain atoms was a single category. In total, 282 of such distributions were analyzed for ^{13}C. To aid comparison of the values obtained, the chemical shift modes for carbon sites in the 20 common amino-acid types in helix, sheet, coil, and turn are listed in Table S1. Chemical shift ranges were obtained using a quantile-based algorithm for bagplots (Rousseeuw et al. 1999). Briefly, to find the range of x at a given confidence fraction C (e.g. 0.95), n equally spaced points x were used to sample the probability density function p_x. The $ceil((1-C)n)$ values with the lowest probability densities $(\text{i.e.} p_x(\{x_i, x_{i+1}, \ldots, x_n)\}|A))$ were removed. The minimum and maximum remaining x positions were defined as the limits of the range.

The expected chemical shifts determined from modes correlate well with the results found using the RefDB approach for the available Cα and Cβ chemical shifts, as seen in the correlation plot shown in Fig. S1. The average difference between the modes and the RefDB values is 0.05 ± 0.4 ppm and the distribution looks Gaussian, see Fig. S2. It is an advantage of the new method that it also works for side-chain carbons and is able to provide accurate chemical shift ranges with few assumptions. Furthermore, the mode-based analysis of PIQC has no reliance on previously determined values, and the PACSY data did not need to be culled before PIQC analysis. In the future, as more data is added to the database, the precision of the expected values and reliability of the ranges determined by PIQC will continue to increase.

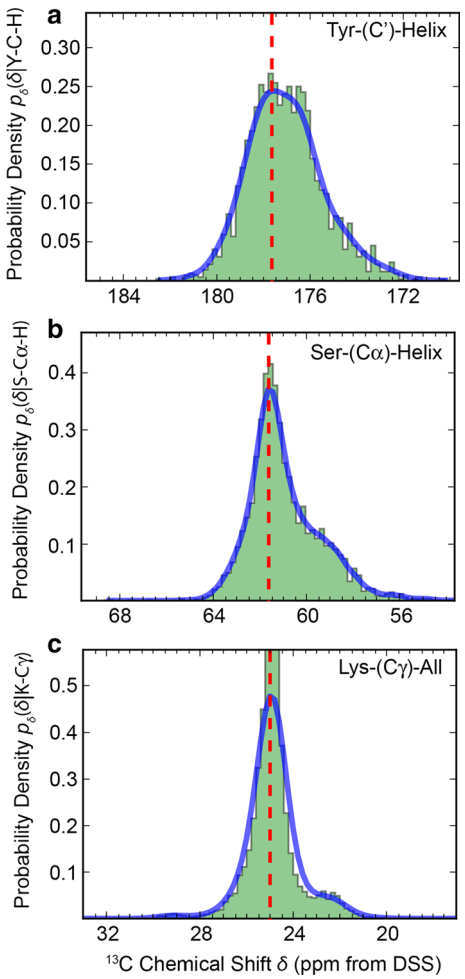

Fig. 1 Typical ^{13}C chemical shift histograms for a specific site of a given amino acid type in one of the three secondary-structure types (in *green*). The generated probability distribution function (PDF) is shown in *blue* (y-scale in ppm^{-1}). The *dashed red line* indicates the position of the maximum of the PDF, which is the most probable chemical shift for a given classification. **a** Tyr-C′-Helix ($N = 1936$). **b** Ser-Cα-Helix ($N = 4531$). **c** Lys-Cγ-All ($N = 18{,}203$). The distributions are clearly not monomodal. For each selection, $p_\delta(\delta|A)$ is shown from $\delta_{m,A} - 8$ ppm to $\delta_{m,A} + 8$ ppm. This analysis was performed for 288 ^{13}C chemical shift distributions

The precision in the determination of the ideal chemical shifts should improve with the amount of data utilized: if a distribution as in Fig. 1a is less noisy, its center can be determined more precisely. Quantitatively, the uncertainty σ_M of the position of the center (mean) of a normal distribution is the standard deviation (SD) σ of the distribution, divided by the square root of the number N of data points in the distribution, $\sigma_M = \sigma/\sqrt{N}$. With an average of 80 ^{13}C amino-acid residues in a protein (only considering proteins with ^{13}C data) in the BMRB, each of the 60 types of Cα carbons will occur about 80/60 = 1.3 times. The data in Fig. 1 suggest a typical SD of $\sigma = 3$ ppm, due, for instance, to chemical-shift effects of neighboring residues. Using a typical curated database of 80 proteins, $N = 1.3 \times 80 = 104$ and the uncertainty in the ideal Cα chemical shift is $\sigma_M = \sigma/\sqrt{N} = 0.3$ ppm, while 2000 proteins in our analysis give a much better $\sigma_M = 0.06$ - ppm. Since PIQC uses data from several thousands of proteins, it is superior to previous analyses based on a smaller number of proteins. A large fraction of the $\pm$0.4 ppm deviation between our ideal Cα chemical shifts and those of RefDB, see Fig. S2, should be attributed to this factor, and this analysis suggests that our values are more accurate by 0.25 ppm, due to the larger number of proteins evaluated.

Determining protein-level chemical shift offsets

At this point, it would be convenient to simply flag all of the chemical shifts that deviate from the mode by more than a given offset. However, the second component is too poorly resolved (see Fig. 1) to allow for a clean separation here. In addition, there are several different ways to explain the multicomponent distribution other than a simple referencing error offset. For instance, one could attribute it to imperfect secondary-structure characterization.

A ^{13}C chemical shift referencing error will, however, equally affect every ^{13}C chemical shift of the protein under consideration. It therefore contributes to every deviation $\Delta\delta_A$ of the chemical shift δ_A from the corresponding ideal value $\delta_{m,A,}$ of that atom type A:

$$\Delta\delta_A = \delta_A - \delta_{m,A}.$$

Wang and Wishart (2005) denoted this quantity by $-\Delta\delta^{offset}$ since they used $\Delta\delta$ for conformational shift (i.e., deviation from random coil); note that our definition of $\Delta\delta_A$ has the opposite sign. Calculation of $\Delta\delta_A$ for each atom of a given protein (labeled by its BMRB ID#) allows for the creation of a distribution $p_{\Delta\delta}$ of the differences between reported and ideal values for that protein. [Note that here, as is typical in the generation of probability distributions, the index A distinguishing different incidences of $\Delta\delta$ disappears when $\Delta\delta$ becomes the argument of the probability distribution, just as for $j = 1,\ldots, N$ position measurements x_j used to produce a probability distribution $p_x(x)$].

The distributions for two proteins are shown in Fig. 2. While the distribution in Fig. 2a is centered close to zero, as should be expected, the example in Fig. 2b has a mode close to the known offset of about -2.4 ppm between neat TMS and DSS. These distributions clearly have a more unimodal character than those in Fig. 1, which indicates that the chemical-shift deviations for a given protein are produced by apparently random effects, while the asymmetries in Fig. 1 are due to systematic referencing errors. The mode (position of the maximum) of each $p_{\Delta\delta}(\Delta\delta|\mathrm{ID\#})$ distribution is considered as the chemical shift offset $\Delta\delta_{m,ID\#}$ for the protein ID#,

$$\Delta\delta_{m,ID\#} = argmax[p_{\Delta\delta}(\Delta\delta|\mathrm{ID\#})]$$

This method was used to analyze the >1.2 million ^{13}C chemical shifts of >3000 proteins (all with solved 3D structures) currently in the PACSY database. A table with

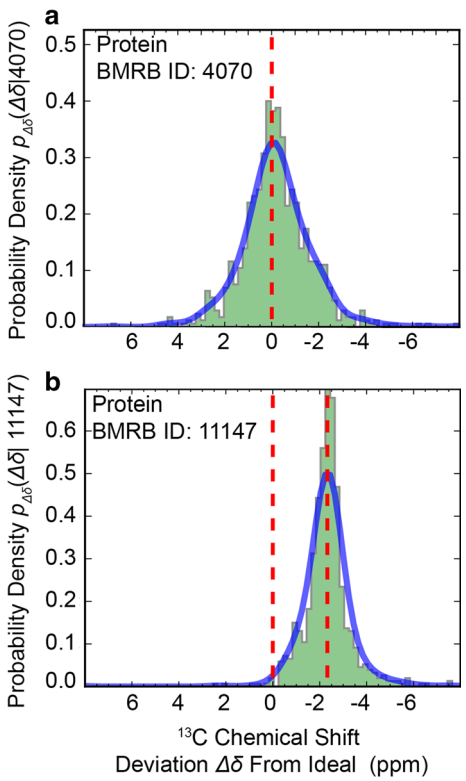

Fig. 2 Representative ^{13}C chemical shift distributions $p_{\Delta\delta}$ of the chemical-shift deviation $\Delta\delta_A = \delta_A - \delta_{m,A}$ for each classification A, for all resonances in a single protein. The histogram is depicted in *green* and the probability distribution function (PDF) in *blue*. All distributions are expected to be centered on 0 ppm, marked with a *dashed red line*. The maximum of the PDF represents the most likely offset $\Delta\delta_{m,ID\#}$. **a** BMRB ID: 4070 and **b** BMRB ID: 11147. This analysis was performed for each of the 3060 proteins with ^{13}C data in PACSY

the offset of each protein will be updated periodically and made available at ksrlab.org/pluqin-sqat.

Limited precision of the ideal chemical shifts could affect the width of the distributions $p_{\Delta\delta}$ for each protein as shown in Fig. 2, but given typical standard deviations of 1 ppm in these distributions, it appears that the <0.2 ppm effect is relatively small.

Identification of problematic chemical shift data

The chemical-shift offsets $\Delta\delta_{m,ID\#}$ of all proteins were used to create a distribution $p_{\Delta\delta m}(\Delta\delta_m)$ (where again the incidence index, presently ID#, has disappeared). The resulting $p_{\Delta\delta m}$ distribution is shown in Fig. 3. It is at least bimodal. The peak centered on 0 ppm and of SD $\sigma \approx 0.2$ ppm results from proteins where the ^{13}C chemical shifts were correctly referenced to DSS. The SD of the main peak can be estimated even in the presence of outliers by calculating the median absolute deviation from the mode (MAD) and multiplying by 1.483; (Hampel 1974) with MAD = 0.112 ppm for our peak, this yields $\sigma \approx 1.483 \times 0.112$ ppm = 0.17 ppm. This SD, which is an upper limit to the root-mean-square referencing error in these proteins, is remarkably small. The second peak near −2.4 ppm very likely results from misreferencing to TMS. Approximately 7 % of the proteins (6 % of chemical shifts) in PACSY fall into this category. For the distribution as a whole, the $avg(\Delta\delta_{m,ID\#}) = -0.2 \pm 0.7$ ppm when only

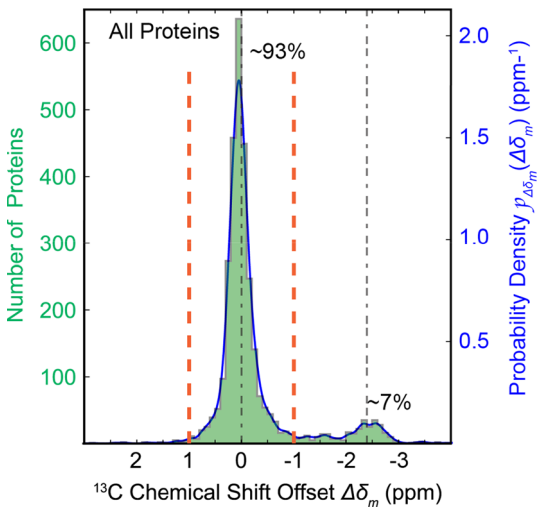

Fig. 3 The distribution of the offsets $\Delta\delta_m$ for all 3060 proteins with ^{13}C chemical shifts in the PACSY database, based on 1,200,207 chemical-shift values. There are two main local maxima: the expected maximum at 0 ppm, with a SD of 0.2 ppm, and a maximum at −2.4 ppm that probably results from erroneous TMS referencing; the chemical-shift positions are marked with *dashed-dotted vertical black lines*. *Orange dashed vertical lines* at ±1 ppm enclose the part of the distribution that was retained. Approximately 7 % of the proteins in the database appear to be incorrectly referenced, corresponding to 6 % of the chemical-shift data

$\Delta\delta_{m,ID\#}$ within 4 ppm of the mode are considered. For $-4 \text{ ppm} \le \Delta\delta_{m,ID\#} \le -1.1$ ppm and $1.1 \text{ ppm} \le \Delta\delta_{m,ID\#} \le 4$ ppm the $avg(\Delta\delta_{m,ID\#}) = -2.2 \pm 1.0$ ppm.

A PSSI analysis by Wishart et al. (Wang and Wishart 2005) showed the same type of bimodal distribution as in Fig. 3, but with significantly lower resolution, probably due to the ~20-times smaller amount of data used, as a result of the limited number of carbon sites and proteins evaluated. The additional component near −2.4 ppm in ref. Wang and Wishart (2005) appeared more like a foot than a resolved second peak. Since the center of their distribution is shifted off the main maximum by ~0.5 ppm, most of the outlying intensity was shown within −2 ppm from the center, which obscured the connection to the −2.4 ppm referencing error. The left, main peak was so broad that the authors of ref. (Wang and Wishart 2005) decided to discard data with deviations by more than −0.5 ppm from the overall average, while our analysis shows these data to be perfectly normal.

In reference Wang and Wishart (2005), the distributions were produced separately for Cα, Cβ and C′, using data from only 450 proteins and with 0.5 ppm wide histogram bars. Correspondingly, Fig. S5a shows the distribution obtained only from Cα data, which is broader than in Fig. 3 by $\sim\sqrt{5}$ due to the ~5-times smaller number of data points used. Next Fig. S5b displays this distribution for only 450 proteins; it has the same width as in part a but exhibits more noise. Finally, Fig. S5c shows the distribution of part b with 0.5 ppm wide histogram bars.

The resolution in Fig. 3 is much improved compared to that of a single protein as in Fig. 2: The distribution of N ^{13}C chemical shifts of a protein, with a SD σ, as before gives a SD of the mean of $\sigma_{mean} = \sigma/\sqrt{N}$. With $\sigma \sim 2$ ppm and $N \sim 400$ in Fig. 2, we predict $\sigma_{mean} = 0.1$ ppm, quite comparable with the SD of 0.2 ppm of the main peak in Fig. 3. Since it combines data from 282 atom types, the distribution in Fig. 3 is also ca. $\sqrt{282} = 17$ times narrower than the distributions in Fig. 1 for a single atom type, where the misreferenced data produced only an unresolved shoulder.

Selection of high-quality chemical shift data

Figure 4 shows a plot correlating the chemical-shift offset $\Delta\delta_{m,ID\#}$ with the SD from the mean of the $p_{\Delta\delta}$ distribution (as in Fig. 2) for each protein. Most proteins that were misreferenced are seen to have a similar SD as the correctly DSS-referenced proteins, which indicates that the different average chemical shifts cannot be attributed to a fraction of sites with unusual conformations and chemical shifts. The large standard deviations >5 ppm for some proteins in Fig. 4 may be due to paramagnetic shifts.

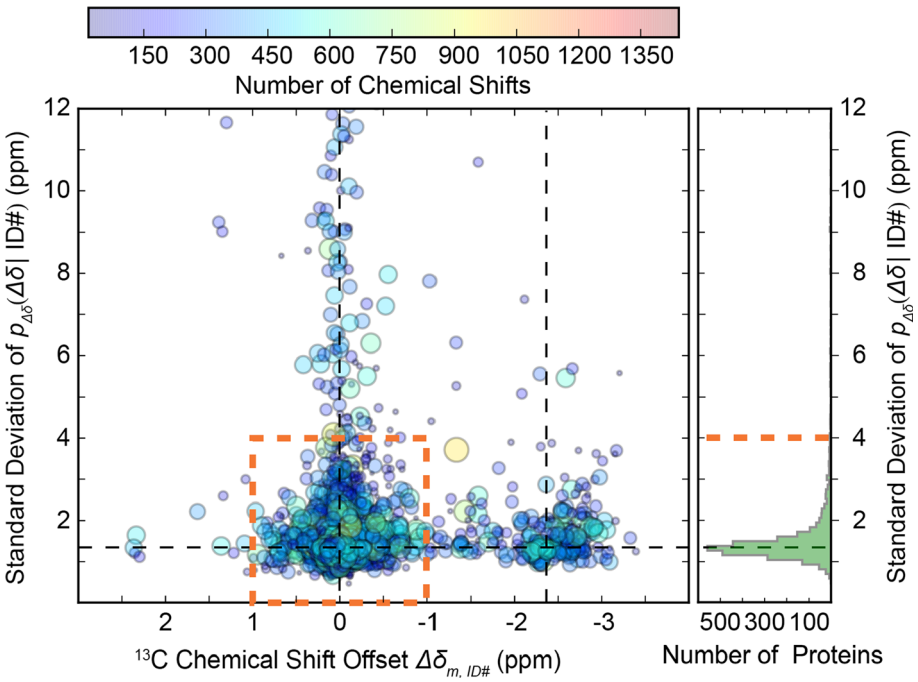

Fig. 4 Chemical shift offset, $\Delta\delta_m$, versus the SD of $p_{\Delta\delta}$, for 3060 proteins in the PACSY database. Each *circle* represents all the ^{13}C chemical shifts of a single protein. The position on the x-axis gives the same information as in Fig. 3, the chemical shift offset $\Delta\delta_m$ for each protein *ID#*. The y-axis represents the SD of the $p_{\Delta\delta}$ distribution as shown in Fig. 2. The *color* and *size* of the circles represents the number of chemical shifts (see *color scale* at the *top* of the figure). Again two local maxima are observed in the distribution. The standard deviations of the proteins with −2.4 ppm offset are comparable to those at the expected 0 ppm offset. A histogram of the standard deviations is shown on the *right*. All proteins within ±1.0 ppm from the 0 ppm offset and with a SD < 4 ppm were used as the purged PACSY dataset. They account for 86 % of all proteins in the database. The same type of plots for hydrogen and nitrogen, where no secondary maxima are observed, are provided in the supporting information (Fig. S1)

Offset or referencing errors cannot be corrected to better than the 0.2 ppm SD of the main peak in Fig. 3. Therefore, it is best to remove improperly referenced data; our approach can tolerate this slight reduction in the amount of data since it takes advantage of data from the large number (>3000) of proteins in the PACSY database.

A new PACSY-like database was generated by excluding all information from proteins with a deviation $\Delta\delta_m > 1.0$ - ppm or with $\sigma > 4$ ppm, (i.e. those outside the orange dashed lines in Fig. 3 and the dashed box in Fig. 4). The ±1 ppm $\Delta\delta_m$ criterion is near the 6σ value of the correctly referenced, dominant sub-distribution centered on 0 ppm in Fig. 3. The −1 ppm cut-off value ensures that most datasets from the sub-distribution offset by −2.4 are removed. Overall, the fraction of chemical shift data removed was 14 %. The chemical shift statistics for each protein were tabulated and the data incorporated into additional PACSY-like SQL tables. The data tables will be shared, along with the Python analysis code that generated it.

Monomodal ^{1}H and ^{15}N NMR data

PIQC was also applied to ^{1}H and ^{15}N (except side-chain ^{15}N) data. The number of p_δ distributions analyzed was 88 for ^{15}N and 213 for ^{1}H. The plots of $p_{\Delta\delta m}$ for ^{1}H and ^{15}N chemical shifts show monomodal, nearly Gaussian distributions with a width of $\sigma = 0.08$ and 0.8 ppm, respectively (see Fig. S3). Only 0.8 % of the 3139 ^{15}N data deviate by >4 ppm (5σ). These fairly narrow distributions appear to be in disagreement with claims (Ginzinger et al. 2007; Wang and Markley 2009; Zhang et al. 2003) of significant ^{1}H and ^{15}N NMR referencing errors. Based on LACS analysis, it was reported that 35 % of ^{15}N chemical shifts are misreferenced by >0.7 ppm, but only 25 % of ^{13}C by >0.5 ppm (Wang et al. 2005). We do find that 25 % of ^{15}N chemical shifts deviate by >0.7 ppm from the expected values but our analysis also shows that this is not a significant offset; it is only approximately 1σ, and in any random distribution, a significant fraction of values of >1σ will be found. Given that the number of chemical shift data is about 3.2 times smaller for ^{15}N than for the main peak in the ^{13}C distribution, the ^{15}N SD should be divided by $\sqrt{3.2}$ before comparison with that of ^{13}C. On that basis, the two values (0.5 vs. 0.2 ppm) are more similar. Overall, our analysis suggests that traditional approaches (Wang et al. 2005; Wang and Wishart 2005; Zhang et al. 2003) will over-correct some data that actually have accurate offsets.

Multidimensional chemical shift regions

After the removal of the questionable data, we re-analyzed (Fritzsching et al. 2013) correlated chemical shifts in 2D spectra. The problem of finding 2D chemical shift range(s) for a given confidence level is equivalent to identifying the smallest-area polygon(s) that will incorporate a given fraction (e.g. 95 %) of all peaks or discrete data points. Multidimensional PDFs were constructed for each distribution, using a Gaussian kernel, i.e. replacing each measured data point with a Gaussian. The bandwidth of the Gaussian kernel was optimized using a grid-search and threefold cross-validation procedure. To find the intensity of the PDF at which to contour, i.e. to define the polygon, a density quantile/percentile algorithm (Rousseeuw et al. 1999) was used. Examples of defined regions of Ala Cβ–Cα correlations for each secondary-structure classification are shown in Fig. 5 at a 95 % confidence level.

Removal of spurious data from 2D correlations

Figure 5 demonstrates how the elimination of the 14 % of misreferenced or large $\sigma_{\Delta\delta}$ data removes spurious secondary maxima, or "lobes", from the chemical shift data and makes real lobes in the distributions recognizable. The distributions before cleanup are shown on the left, those after removal of misreferenced data and outliers in the right column. For the helix data, top row, the removal of a lobe shifted by ~ -2.4 ppm in both dimensions is especially clear, since the true maximum in the distribution is particularly pronounced. For the other two secondary structures, the corresponding artifact towards the upper right is more diffuse, but in all cases, the chemical-shift distributions are significantly tighter after cleanup.

Non-Gaussian chemical shift distributions

The removal of misreferenced chemical shifts by PIQC allows for close examination of the distributions. In the high-quality data in the right column of Fig. 5, distinct lobes can now be reliably discerned for each of the secondary structure categories. For helix and sheet, lobes near (52.5, 19.3 ppm) appear to coincide with the maximum of the coil distribution. This could prompt a reassessment of the secondary structure associated with these outlying peaks. Conversely, in the coil distribution, in addition to the pronounced maximum lower intensity extends to the maxima of helix and sheet distributions. This should not be unexpected since dihedral angles matching those in helix and sheet can be compatible with a random coil conformation.

Most interestingly, the cleaned-up coil distribution, Fig. 5d, shows a clear secondary maximum near (50.5, 18.1 ppm), which does not coincide with maxima for other secondary structures. Similar multimodal distributions (with additional maxima or 'lobes') are found in several (Cα, Cβ)-coil distributions, see Figs. 6 and 7, as well as some (Cα, N)-sheet chemical shifts, see Fig. 8. Further investigations of the structural features that set some of these lobes apart are described in the following. In our previous study (Fritzsching et al. 2013) we had abandoned such selective analyses after the analysis of the spurious secondary peak in Fig. 5a showed no difference in the torsion-angle distribution.

Special structural features of some secondary maxima

Ramachandran diagrams of the distributions of torsion angles associated with various distinctive spectral regions marked by rectangles in the Ala coil Cα–Cβ distribution of Fig. 6a are displayed in Fig. 6b–e. Compared with the conformations associated with the main maximum for Ala in coil conformations, some of the secondary spectral maxima are associated with clearly different torsion angle distributions. For instance, the region near the α-helix maximum shows predominantly torsion angles corresponding to α-helical conformations. We attribute these to isolated residues with such torsion angles in a non-helical coil environment.

The secondary maximum near (50.5, 18.1 ppm) shows a slight preference for polyproline conformations, see Fig. 6c. Nevertheless, its most distinctive feature is revealed by an analysis of the neighboring amino-acid types. Figure 6f shows that 75 % of these alanines are followed by proline, and no alanine followed by proline resonates near the coil maximum. A similar result is obtained for the corresponding peak in the distribution for Lys in coil conformations, see Fig. 7. Here, 2/3 of the following residues are Pro.

The Lys data in Fig. 7a show an additional distinctive lobe near (57.4, 29 ppm). Our analysis reveals, see Fig. 7e, that it corresponds nearly exclusively to torsion angles in the left-handed helix region. Again, it seems likely that some coil residues take these sterically allowed positive torsion angles without actually being part of a helix. Note also in Fig. 7d the nearly exclusively helix-like dihedral angles of Lys coil residues near the maximum of the helix distribution.

Two distinct local maxima are observed in the N–Cα sheet correlations for Asp, see Fig. 8a. All of the signals arise from residues near the ends of β-sheets, which is a preferred location for Asp and Asn. They appear to be distinguished by their torsion angles and hydrogen bonding as indicated in Fig. 8b, c. While the ^{15}N, ^{13}C

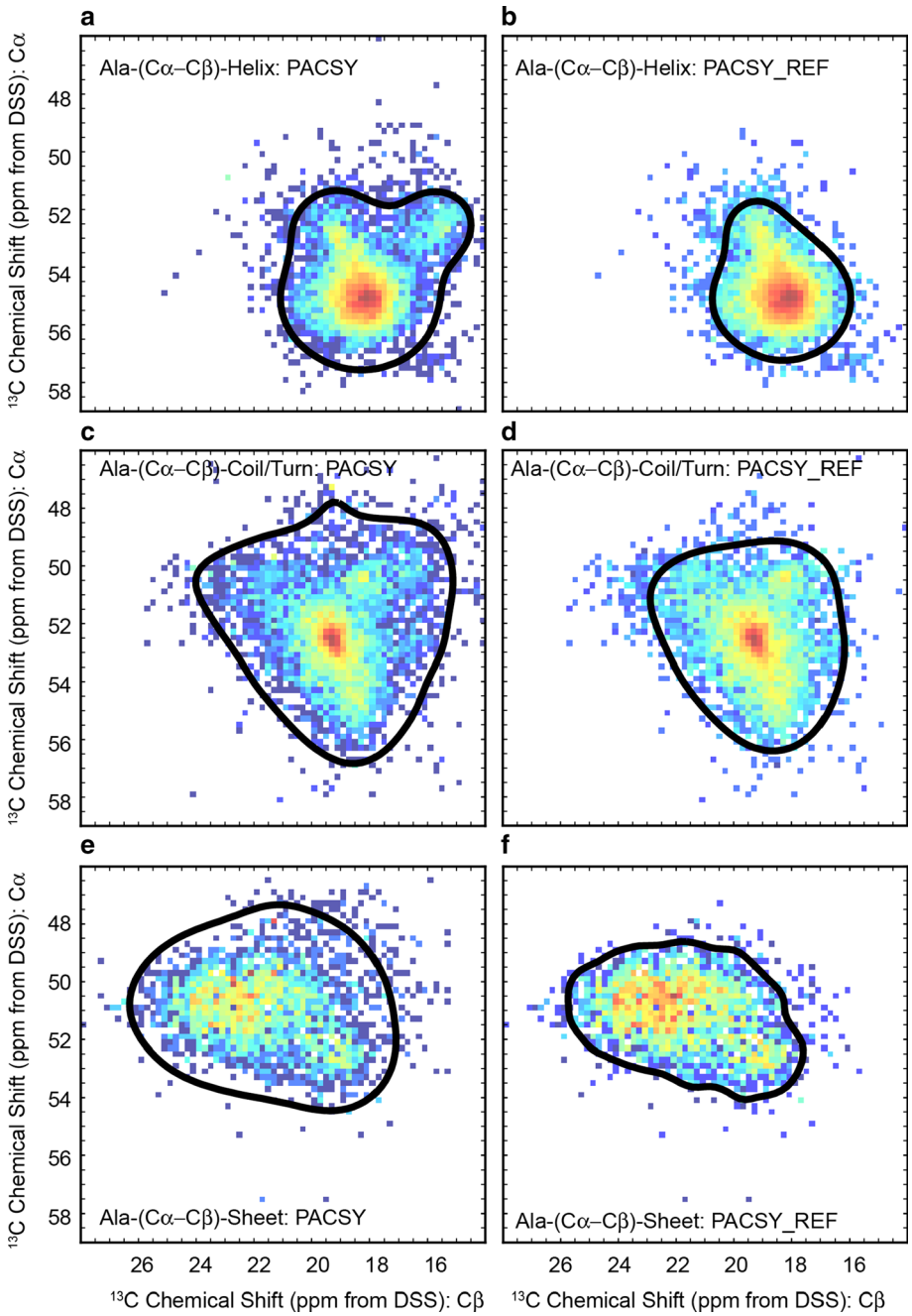

Fig. 5 Comparison of two-dimensional chemical shift distributions and associated 95 % containment regions of Ala Cβ–Cα correlations for different secondary-structure classifications before (*left column*) and after (*right column*) removal of misreferenced data and outliers from the PACSY database. **a**, **b** Helix; **c**, **d** coil and turn; and **e**, **f** sheet. The *color scale* is logarithmic

chemical shifts at (125, 54) ppm correspond mostly to the canonical hydrogen-bonding geometry, the chemical shifts near (121, 55.5) ppm are from structures that either lack backbone-N hydrogen bonding or where the bonding partner residue is not classified as sheet. The average of the corresponding dihedral angle distributions (ϕ, ψ) calculated using circular (or directional) statistics (Berens 2009) are ($-101° \pm 28°$, $126° \pm 31°$) and ($-99° \pm 39°$, $136° \pm 34°$) [circular average $\pm$ circular SD] for the (127, 53) and the (122, 55.5) ppm distribution, respectively.

Chemical shift maps

The 2D chemical shift ranges for different correlations can be combined to generate maps that are useful for direct comparisons with spectra. As an example, the map of one- and two-bond ^{13}C–^{13}C correlations in Leu is presented in Fig. 9. The corresponding maps for all 20 common amino acids are shown in the SI. They generally show helix best resolved, in particular in the C′–Cα correlation. Sheet conformations have the least overlap from coil in Ala, Thr, and Ser Cα- Cβ, and in Gly C′-Cα correlations.

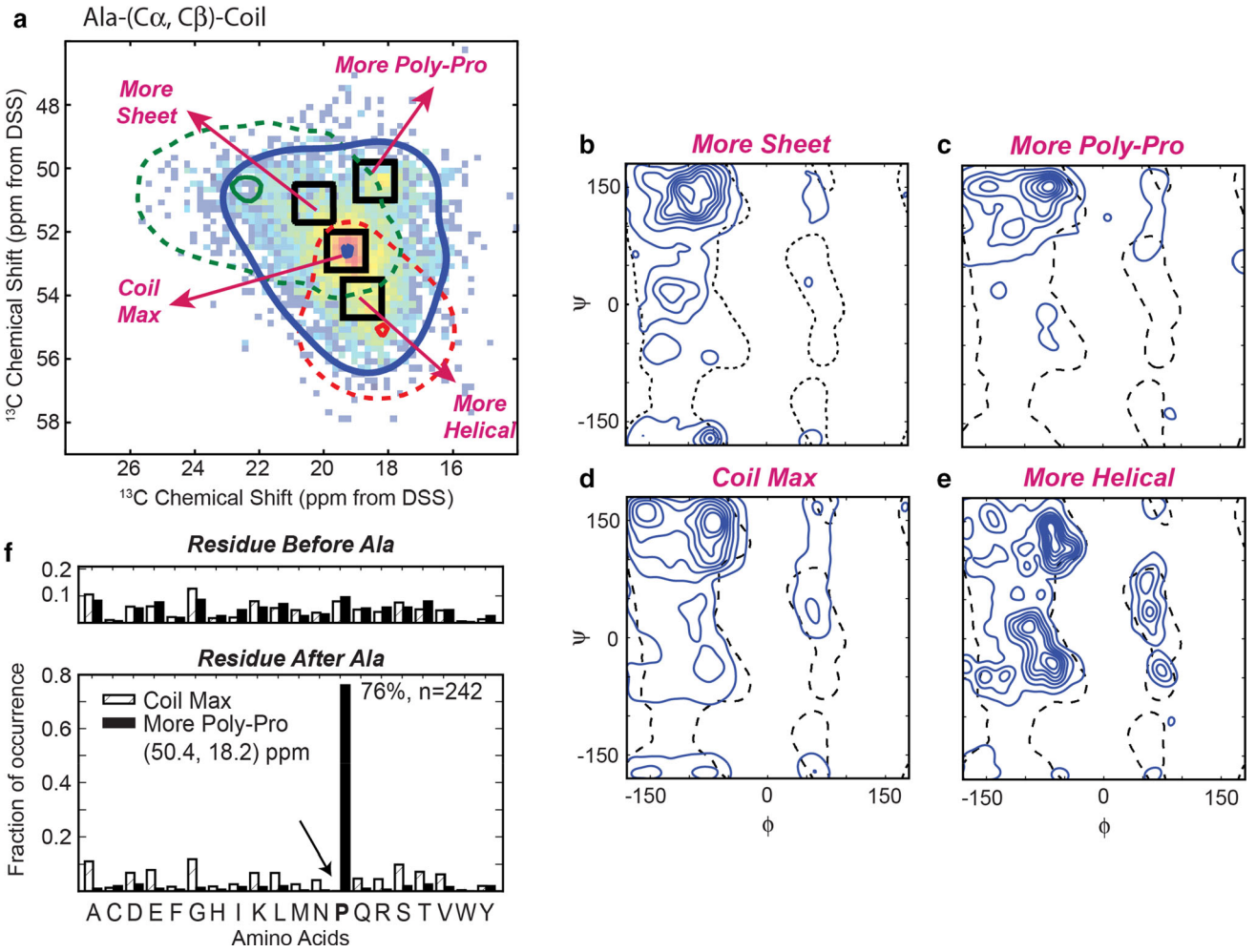

Fig. 6 **a** Ala-(Cα, Cβ)-coil chemical shift distribution. The *color scale* is logarithmic. **b**–**e** The distributions of *dihedral angles* in Ramachandran plots for various chemical shift regions indicated in (**a**): **b** Region around (51.1, 20.2) ppm, toward the sheet region. **c** Region around (50.4, 18.2) ppm, toward the *upper right*. **d** Region around (52.6, 19.2) ppm, around the coil maximum. **e** Region around (54.2, 18.7) ppm, toward the *helical region*. The *dashed black lines* mark the 98 % confidence regions for 500 high-resolution protein structures (Lovell et al. 2003). **f** Fractions of amino acid types directly preceding (*top*) and following (*bottom*) Ala in the chemical shift region near (50.4, 18.2) ppm ("Poly-Pro"), shown as *filled bars*. The corresponding fractions of neighboring residues of alanines at the maximum of the coil distribution are shown as *striped bars* for reference. The *arrow* in (**f**) highlights the very low fraction of Ala with typical coil chemical shifts and a neighboring Pro

Impact on other chemical-shift based tools

Our highly precise ideal (or expected) chemical shift values characteristic of helix, coil, and sheet could improve the accuracy of many existing chemical-shift assignment and re-referencing programs, including PSSI and PANAV (Wang et al. 2010; Wang and Wishart 2005). For some of these tools, performance should increase by simply replacing the older expected values with the more precise values found in this analysis (see Tables S1–S3). Hopefully, this will allow for an improvement in misassignment identification and re-referencing results without the need for reinventing the existing tools. Programs such as AVS and PANAV (Moseley et al. 2004; Wang et al. 2010) that use distributions of chemical shifts to provide assignment probabilities would benefit from using the non-Gaussian distributions determined in our analysis. The large amount of data and the use of carefully chosen smoothing parameters yields accurate chemical shift ranges even at high confidence values, whereas the use of Gaussian-based statistics (Moseley et al. 2004; Wang et al. 2010) results in less accurate chemical shift ranges due to the truncation of shoulders or tails in the chemical shift distributions.

The relations between reliable chemical-shift data and dihedral angles from our large-scale analysis might also be useful for improving chemical-shift hypersurfaces as used in TALOS and SPARTA-like algorithms (Shen and Bax 2010, 2013). Some secondary structures with distinct

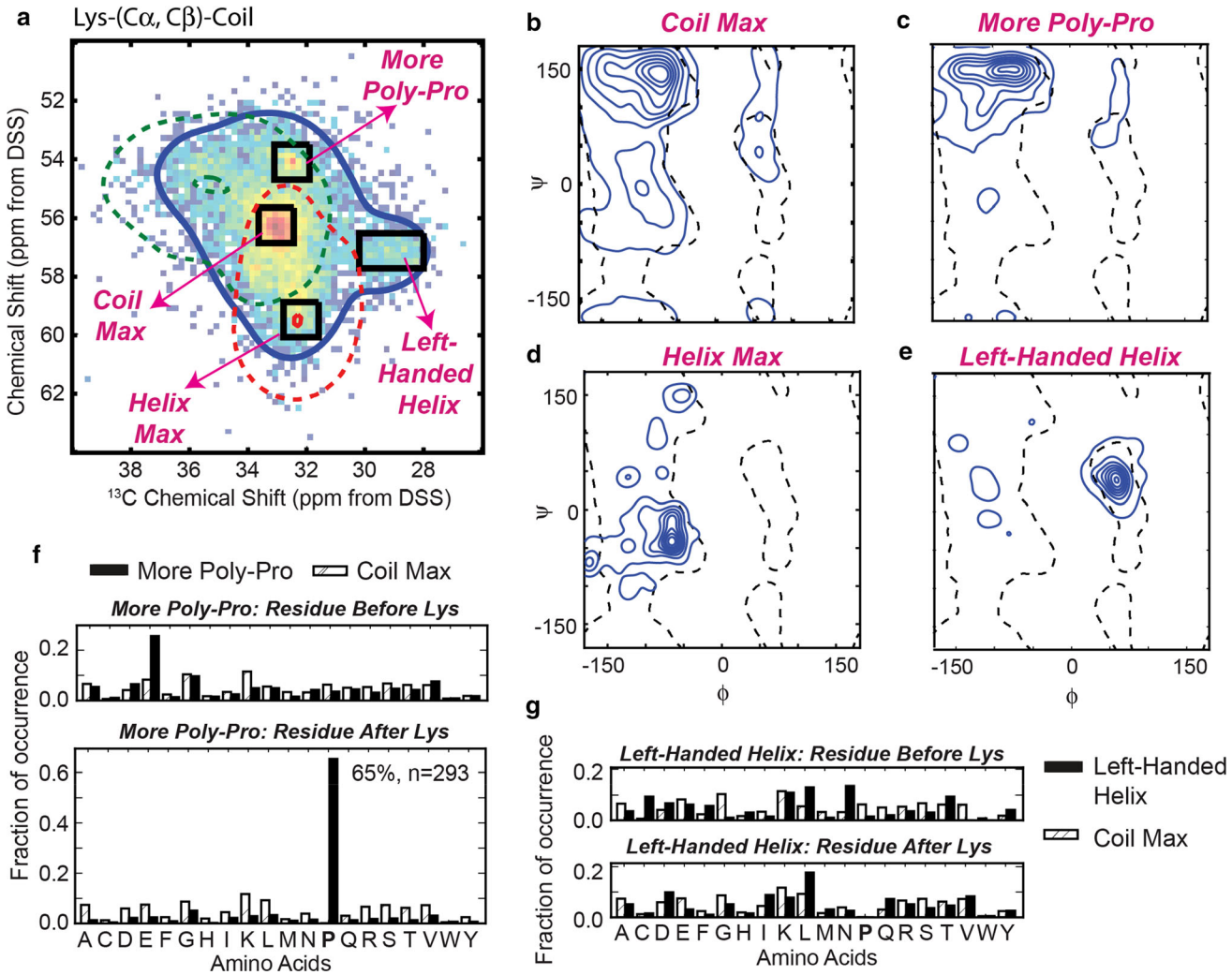

Fig. 7 **a** Lys-(Cα, Cβ)-coil chemical shift distribution. The *color scale* is logarithmic. The *closed blue line* contains 95 % of coil residues. Corresponding 95 % contours of helix and sheet are shown as *dashed lines*. **b–e** Distributions of *dihedral angles* in Ramachandran plots for various chemical shift regions indicated in (**a**). **b** Region around (56.3, 33.0) ppm, toward the coil maximum. **c** Region around (54.2, 32.5) ppm, toward the *upper right*. **d** Region around (54.2, 32.7) ppm, toward the helix maximum. **e** Region around (57.4, 29.1) ppm, toward the *center right*. **f** Fractions of amino acid types directly preceding (*top*) and following (*bottom*) Lys in the chemical shift region near (54.2, 32.5) ppm ("Poly-Pro"), shown as *filled bars*. The corresponding fractions of neighboring residues of Lys at the maximum of the coil distribution are shown as striped bars for reference. **g** Same as **f** for the left-handed helix chemical shift region

chemical-shift correlations are present only in very few proteins; for instance, there are <300 examples (out of >1 million chemical shifts) of the Lys "left-handed helix" correlation shown in Fig. 7. This suggests that machine-learning algorithms for predicting or interpreting chemical shifts that utilize a relatively small number of proteins may never be trained with these correlations.

Modes versus averages

The distributions in Fig. 1 are clearly asymmetric due to some data with referencing errors and therefore require an analysis in terms of their modes. On the other hand, the $p_{\Delta\delta}$ in Fig. 2 are mostly unimodal regardless of the referencing error, and therefore using the average $\Delta\delta_{a,ID\#}$ of $P_{\Delta\delta}(\Delta\delta|\mathrm{ID}\#)$ instead of the mode gives similar results for most proteins (for ^{13}C: $avg(\Delta\delta_{m,ID\#} - \Delta\delta_{a,ID\#}) = -0.02 \pm 0.17$ ppm for the proteins inside the dashed box in Fig. 4). However, when all proteins are included, the SD of the difference between mode and average for ^{13}C increases to ±1.4 ppm. While some deviations can arise from paramagnetic effects, cases where $\Delta\delta_{m,ID\#}$ and $\Delta\delta_{a,ID\#}$ are very different usually indicate typographical errors or incorrect assignments. When the errors are few (as is often the case) $\Delta\delta_{m,ID\#}$ is a better estimate of the referencing offset than is

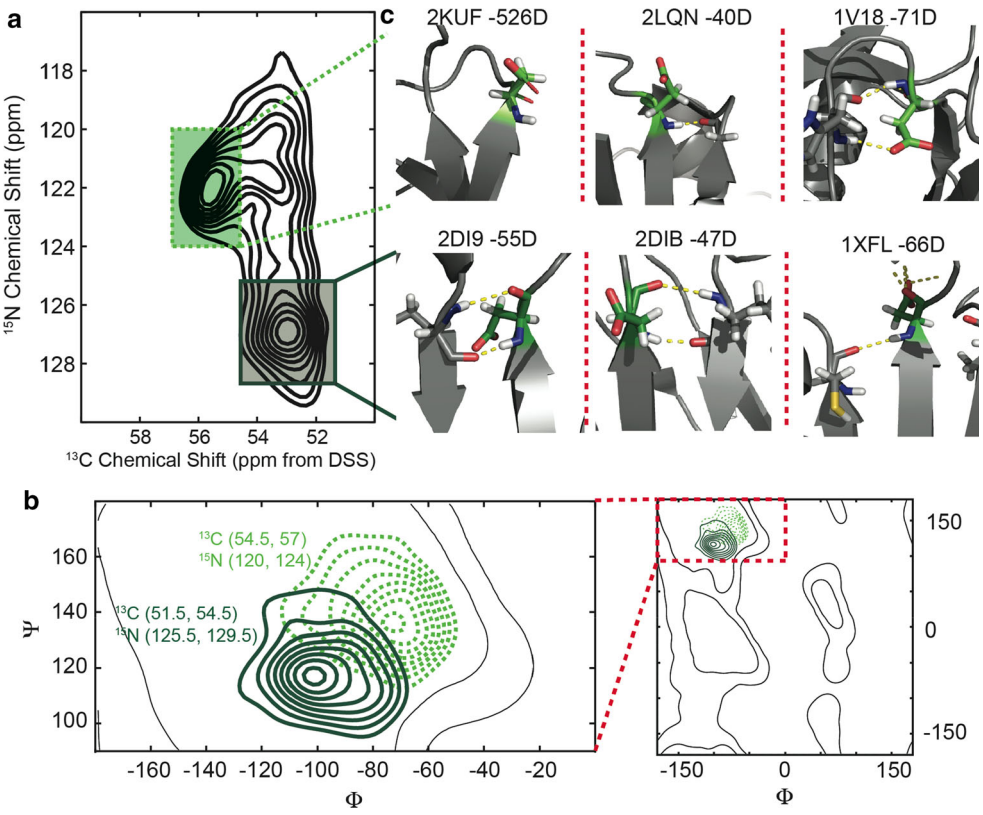

Fig. 8 **a** Asp(N, Cα)-sheet chemical shift distribution. Two distinct local maxima are marked by boxes. **b** The distributions of torsion angles for each of the local maxima in (**a**). The upper distribution (*dashed lines*) is close to the canonical β-turn torsion angles. **c** Examples of typical Asp residues from the two maxima in the chemical shift distribution. *Top row* residues without canonical H-bonding between neighboring β-strands. *Bottom row* residues with H-bonding between strands. All images are of the first model in the PDB, except for the 1XFL image, which is of the second model

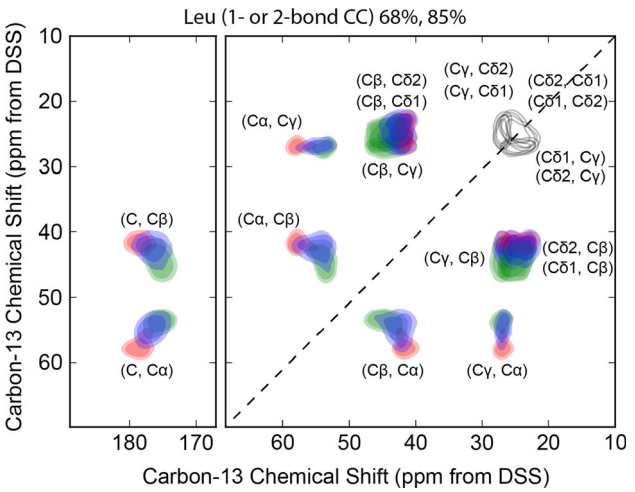

Fig. 9 Refined correlation patterns for ^{13}C–^{13}C chemical shifts of Leu for atoms within two bonds. Regions are shown at 68 and 85 % confidence levels. *Helix* is represented by *red*, coil and turn by *blue*, and sheet by *green*. Regions where neither atom contributes secondary-structure information are not filled in. Similar figures are provided in the supporting information for the other 19 common amino acids

$\Delta\delta_{a,ID\#}$. As an example, compare $\Delta\delta_{m,18910} = 0.2$ ppm with $\Delta\delta_{a,18910} = 46.8$ ppm for ^{13}C: The large average $\Delta\delta_a$ comes from the assignment of 16 Y Cδ atoms to chemical shifts >730 ppm; these outliers are far too few to shift the mode $\Delta\delta_m$. While this kind of error can be quite easily identified, there are others that are more difficult to classify. As an example, for the ^{13}C data of protein BMRB ID: 4150, the mode of $\Delta\delta_{m,4150} = 39.9$ ppm differs strongly from the average, $\Delta\delta_{a,4150} = 13.9$ ppm, and the values deviate so far from 0 that both must be incorrect. These unusual problems are rare enough that their effect on the distribution is not generally significant for our determination of chemical shift ranges. Further examination of $p_{\Delta\delta}(\Delta\delta|4150)$ shows four approximately normal distributions spaced by 40 ppm (see Fig. S4). This particular error was also identified using PANAV. PANAV allows each chemical shift to be re-referenced independently (even within one protein). Thus, these chemical shifts would have been re-referenced to the "ideal value". However, it is not clear to us how this maintains a distribution related to the experiment; indeed, tests of re-referencing programs with intentionally misreferenced data show residual errors of 0.2–1.4 ppm (see Table 2 of ref. Ginzinger et al. 2007).

Assignment selection by PLUQin

As outlined in the "Introduction", sequential-assignment algorithms (Hu et al. 2011; Tycko and Hu 2010; Yang et al. 2013) require as input all reasonably possible type assignments for a given set of chemical shifts. Based on the

chemical-shift ranges identified here, we provide a program, called PLUQin, that upon input of an individual chemical shift, or of two correlated chemical shifts, provides all possible type assignments at a chosen confidence level (e.g. 68 or 95 %). Compared to its precursor, PLUQ, PLUQin is normalized with respect to, and therefore insensitive to, the occurrence frequency of different amino-acid types in the database. Optionally, secondary structure information is also provided. Tested on the model protein GB1 (BMRB ID 15156), with a 95 % confidence level, 94 % out of 413 correct two-bond correlations were included in the assignments proposed by PLUQin. The average number of proposed amino-acid types for a given cross peak was 4; there were $\leq$3 possible assignments 69 % of the time (48 % for Cα and Cβ). It should be useful to run the sequential-assignment algorithm iteratively: First PLUQin should be run with tight (e.g. 68 %) confinement regions, corresponding to the highest-likelihood regions of the chemical shift distributions to assign the resonances with fairly typical chemical shifts. If a certain chemical-shift correlation is not assigned a residue in most of the solutions offered, its confinement level can be relaxed to include more unusual chemical shift assignment possibilities.

Nearly unique type assignments are possible along the margins of the ^{13}C–^{13}C chemical shift distributions. Figure 10 show these regions with a 90 % confidence level, i.e. the probability density for one amino acid type in such a region is at least ten times larger than for all others combined. The plot was obtained based on our smoother chemical shift distributions, and updates Fig. 6 from our previous paper (Fritzsching et al. 2013). Figure 10a shows the plot for a protein with the average amino-acid composition of many proteins, but this analysis can also take into account the amino-acid composition of the specific protein under study. As an example, in an extreme case the protein may not contain Cys, and therefore a peak near (53.5, 46) ppm cannot be assigned to Cys, which makes an assignment to Leu much more likely. Similarly, few Cys and large numbers of Leu in the sequence still make a Leu assignment more probable. This is the case for the VDAC protein (Raschle et al. 2009), whose nearly unique type assignment distributions are shown in Fig. 10b.

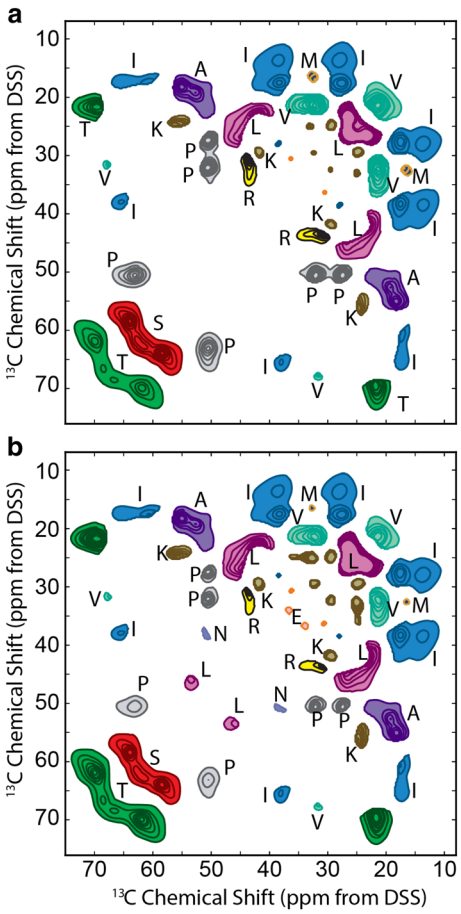

Fig. 10 All ^{13}C–^{13}C chemical shift regions where one residue type assignment is >10 times more likely than all other assignments combined, for **a** a hypothetical protein with the typical amino-acid fractions and for **b** the VDAC-1 (PDB ID = 2K4T) protein. These maps can be used to type assign a 2D chemical shift peak that falls into one of the colored regions with >90 % confidence

Simple offset and quality test (SQAT) for new protein data

The presence of ~6 % incorrectly referenced protein data indicated by PIQC suggests that successful chemical-shift re-referencing programs have not been sufficiently widely adopted. Our analysis suggests a simple quality test (SQAT) that should be applied to new protein NMR data before submission to the BMRB or for publication. At that stage, the secondary-structure information required by SQAT is usually available. The distribution $p_{\Delta\delta}$ should be calculated for the new protein structure, simply based on the tabulated modes from PIQC analysis (which does *not* have to be performed again) and the protein's measured chemical shifts with their amino-acid type and conformational assignments. This distribution easily reveals likely referencing offset and outliers. For instance, the incorrect referencing by about −2.4 ppm in Fig. 2b is apparent. Figure S4 shows additional examples where visual inspection immediately reveals problems with the data. The SQAT routine requires experimental chemical-shift data (BMRB Star or TALOS format) and the protein secondary structure (STRIDE or DSSP format) as input. The program is available at ksrlab.org/pluqin-sqat.

Conclusions

A simple self-referencing method, without handpicked data sets, for eliminating incorrectly referenced protein data has been introduced (purging by intrinsic quality criteria, or PIQC). Our analysis shows that >94 % of ^{13}C NMR spectra have been referenced with a SD of <0.2 ppm, which is better than previous analyses have suggested. Indeed, our distributions imply that traditional approaches will improperly "correct" >10 % of data that actually have accurate offsets; this also applies to ^{15}N chemical shifts. The increased resolution of our approach can be attributed to the >10 times larger amounts of data used. Our distribution of estimated offsets also demonstrates more clearly than previous studies that about 6 % of data are distinctly misreferenced by more than −1 ppm, mostly by −2.4 ppm, which strongly suggests use of a TMS rather than the preferred DSS standard. After removal of these outliers and of data from proteins with excessively wide chemical-shift distributions, more tightly defined spectral regions for 282 carbon types have been obtained. In particular, the helix, sheet, and coil regions for C′, Cα, and Cβ have been determined more accurately. The two-dimensional distributions of carbon–carbon and carbon–nitrogen correlations within many conformation types are distinctly non-Gaussian, revealing the commonly made Gaussian approximation as inadequate. Secondary maxima in the distributions have been revealed as being due to unusual conformations (e.g. "left-handed helix"), from ends of secondary-structure elements, or arising from following proline residues. Future analysis of secondary maxima or distinct lobes in other distributions promises to reveal more such features of interest for structural analysis. The more distinct conformationally selective distributions obtained here may be useful for other chemical-shift based tools and enable conformational analysis without sequential assignment. For instance, they enable amino-acid type assignment with improved confidence (PLUQin), as well as a simple quality test (SQAT) for detecting referencing errors and similar problems that should be applied to any new protein data before publication or deposition in a databank.

Acknowledgments K. S. R. gratefully acknowledges Brandeis University for support. This work was partly supported by NIH Grant GM066976 to M. H.

References

Berens P (2009) CircStat: a MATLAB toolbox for circular statistics. J Stat Softw 31:1–21. doi:10.18637/jss.v031.i10

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302. doi:10.1023/A:1008392405740

Fritzsching KJ, Yang Y, Schmidt-Rohr K, Hong M (2013) Practical use of chemical shift databases for protein solid-state NMR: 2D chemical shift maps and amino-acid assignment with secondary-structure information. J Biomol NMR 56:155–167. doi:10.1007/s10858-013-9732-z

Ginzinger SW, Gerick F, Coles M, Heun V (2007) CheckShift: automatic correction of inconsistent chemical shift referencing. J Biomol NMR 39:223–227. doi:10.1007/s10858-007-9191-5

Ginzinger SW, Skocibusic M, Heun V (2009) CheckShift improved: fast chemical shift reference correction with high accuracy. J Biomol NMR 44:207–211. doi:10.1007/s10858-009-9330-2

Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393. doi:10.1080/01621459.1974.10482962

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57. doi:10.1007/s10858-011-9478-4

Hastie T, Tibshirani R, Firedman J (2009) Model inference and averaging: the elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, Berlin

Hazan C et al (2008) Structural insights on the pamoic acid and the 8 kDa domain of DNA polymerase beta complex: towards the design of higher-affinity inhibitors. BMC Struct Biol 8:22. doi:10.1186/1472-6807-8-22

Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32:W500–W502. doi:10.1093/nar/gkh429

Helmus JJ, Jaroniec CP (2013) Nmrglue: an open source Python package for the analysis of multidimensional NMR data. J Biomol NMR 55:355–367. doi:10.1007/s10858-013-9718-x

Hu KN, Qiang W, Tycko R (2011) A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. J Biomol NMR 50:267–276. doi:10.1007/s10858-011-9517-1

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. doi:10.1109/Mcse.2007.55

Iwadate M, Asakura T, Williamson MP (1999) Cα and Cβ carbon-13 chemical shifts in proteins from an empirical database. J Biomol NMR. doi:10.1023/A:1008376710086

Lee W, Yu W, Kim S, Chang I, Lee W, Markley JL (2012) PACSY, a relational database management system for protein structure and chemical shift analysis. J Biomol NMR 54:169–179. doi:10.1007/s10858-012-9660-3

Lovell SC et al (2003) Structure validation by Cα geometry: Φ, Ψ and Cβ deviation. Proteins 50:437–450. doi:10.1002/prot.10286

McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404–405

Moseley HN, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR 28:341–355. doi:10.1023/B:JNMR.0000015420.44364.06

Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. J Biomol NMR 26:215–240. doi:10.1023/A:1023812930288

Open Source Geospatial Foundation (2003) GEOS—Geometry engine open source. http://trac.osgeo.org/geos/. Accessed Sept 2015

Pedregosa F et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Raschle T, Hiller S, Yu TY, Rice AJ, Walz T, Wagner G (2009) Structural and functional characterization of the integral membrane protein VDAC-1 in lipid bilayer nanodiscs. J Am Chem Soc 131:17777–17779

Romano JP (1988) On weak-convergence and optimality of kernel density estimates of the mode. Ann Stat 16:629–647. doi:10.1214/aos/1176350824

Rost B, Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72. doi:10.1002/prot.340190108

Rousseeuw PJ, Ruts I, Tukey JW (1999) The bagplot: a bivariate boxplot. Am Stat 53:382–387. doi:10.2307/2686061

Saito H, Ando I, Ramamoorthy A (2010) Chemical shift tensor—the heart of NMR: insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc 57:181–228. doi:10.1016/j.pnmrs.2010.04.005

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48:13–22. doi:10.1007/s10858-010-9433-9

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56:227–241. doi:10.1007/s10858-013-9741-y

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223. doi:10.1007/s10858-009-9333-z

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and C. alpha. and C. beta. ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492. doi:10.1021/ja00014a071

Tycko R (2015) On the problem of resonance assignments in solid state NMR of uniformly ^{15}N, ^{13}C-labeled proteins. J Magn Reson 253:166–172. doi:10.1016/j.jmr.2015.02.006

Tycko R, Hu KN (2010) A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. J Magn Reson 205:304–314. doi:10.1016/j.jmr.2010.05.013

Ulrich EL et al (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408. doi:10.1093/nar/gkm957

van der Walt Sf, Colbert SC, Varoquaux Gl (2011) The NumPy array: a structure for efficient numerical computation. Comput Sci Eng 13:22–30. doi:10.1109/mcse.2011.37

Wang Y, Jardetzky O (2002a) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084. doi:10.1021/ja026811f

Wang Y, Jardetzky O (2002b) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861. doi:10.1110/ps.3180102

Wang L, Markley JL (2009) Empirical correlation between protein backbone 15N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. J Biomol NMR 44:95–99. doi:10.1007/s10858-009-9324-0

Wang Y, Wishart DS (2005) A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. J Biomol NMR 31:143–148. doi:10.1007/s10858-004-7441-3

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 32:13–22. doi:10.1007/s10858-005-1717-0

Wang L, Eghbalnia HR, Markley JL (2007) Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. J Biomol NMR 39:247–257. doi:10.1007/s10858-007-9193-3

Wang B, Wang Y, Wishart DS (2010) A probabilistic approach for validating protein NMR chemical shift assignments. J Biomol NMR 47:85–99. doi:10.1007/s10858-010-9407-y

Yang Y, Fritzsching KJ, Hong M (2013) Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm. J Biomol NMR 57:281–296. doi:10.1007/s10858-013-9788-9

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195. doi:10.1023/A:1022836027055