J Biomol NMR (2012) 52:179–190
DOI 10.1007/s10858-011-9600-7

ARTICLE

A procedure to validate and correct the ^{13}C chemical shift calibration of RNA datasets

Thomas Aeschbacher · Mario Schubert · Frédéric H.-T. Allain

Received: 13 October 2011 / Accepted: 13 December 2011 / Published online: 18 January 2012

Abstract Chemical shifts reflect the structural environment of a certain nucleus and can be used to extract structural and dynamic information. Proper calibration is indispensable to extract such information from chemical shifts. Whereas a variety of procedures exist to verify the chemical shift calibration for proteins, no such procedure is available for RNAs to date. We present here a procedure to analyze and correct the calibration of ^{13}C NMR data of RNAs. Our procedure uses five ^{13}C chemical shifts as a reference, each of them found in a narrow shift range in most datasets deposited in the Biological Magnetic Resonance Bank. In 49 datasets we could evaluate the ^{13}C calibration and detect errors or inconsistencies in RNA ^{13}C chemical shifts based on these chemical shift reference values. More than half of the datasets (27 out of those 49) were found to be improperly referenced or contained inconsistencies. This large inconsistency rate possibly explains that no clear structure–^{13}C chemical shift relationship has emerged for RNA so far. We were able to recalibrate or correct 17 datasets resulting in 39 usable ^{13}C datasets. 6 new datasets from our lab were used to verify our method increasing the database to 45 usable datasets. We can now search for structure–chemical shift relationships with this improved list of ^{13}C chemical shift data. This is demonstrated by a clear relationship between ribose ^{13}C shifts and the sugar pucker, which can be used to predict a C2′- or C3′-endo conformation of the ribose with high accuracy. The improved quality of the chemical shift data allows statistical analysis with the potential to facilitate assignment procedures, and the extraction of restraints for structure calculations of RNA.

Keywords RNA · NMR spectroscopy · Chemical shift · ^{13}C referencing · A-form RNA · C2′-endo · Sugar pucker

Introduction

NMR chemical shifts of biomolecules are a rich source of local structural and dynamic information (Mulder and Filatov 2010; Wishart and Case 2001). Their extensive use for protein structure determination is well documented ranging from facilitating resonance assignment (Grzesiek and Bax 1993), detecting cis-peptide bonds (Schubert et al. 2002), predicting secondary structure (Wishart et al. 1992), deriving angle restraints (Cornilescu et al. 1999) to the generation of 3D structures (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). Their application for RNA structure determination is still limited (Lam and Chi 2010). Especially the information content of ^{13}C chemical shifts of RNA has not been systematically exploited although recent studies showed a strong potential in providing structural information for RNA (Fares et al. 2007; Ohlenschlager et al. 2008).

Despite the fact that frequencies can be measured very accurately with modern NMR spectrometers, the chemical shift is a relative measure that depends strongly on correct calibration to a standard. Inaccurate or incorrect chemical shift referencing can blur or distort the information contained in the chemical shift data. The standard procedure

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9600-7) contains supplementary material, which is available to authorized users.

T. Aeschbacher · M. Schubert (✉) · F. H.-T. Allain (✉)
Institute for Molecular Biology and Biophysics, ETH Zürich, 8093 Zürich, Switzerland
e-mail: schubert@mol.biol.ethz.ch

F. H.-T. Allain
e-mail: allain@mol.biol.ethz.ch

for calibrating chemical shifts of biomolecules is well documented (Wishart et al. 1995) and should be applied prior to any chemical shift assignment. A reliable chemical shift database is indispensable for comparing chemical shifts of different structures, and to reveal structure–chemical shift relationships. Unfortunately, a significant percentage of deposited chemical shifts in the Biological Magnetic Resonance Data Bank (BMRB) (Seavey et al. 1991) is still incorrectly calibrated. A study from 2003 revealed that 25% of all protein entries contained incorrectly referenced ^{13}C chemical shifts, and 40% of all protein entries appeared to have assignment errors (Zhang et al. 2003). In the meantime, a variety of protocols and programs exist to detect and eventually correct calibration errors in deposited protein chemical shifts (Ginzinger et al. 2007; Wang and Wishart 2005; Zhang et al. 2003).

To date, such a procedure is not available for RNA chemical shift depositions. Recent studies of structure–^{13}C chemical shift relationships of RNAs (Fares et al. 2007; Ohlenschlager et al. 2008) noted that inconsistent calibration is a serious problem for RNA chemical shift data. Therefore, a procedure to check ^{13}C calibration in RNAs would be highly desirable. We therefore decided to establish such a procedure. Our analysis of over sixty ^{13}C chemical shift datasets deposited in the BMRB database identified various sources of inconsistencies in ^{13}C chemical shifts allowing us to correct several datasets, and therefore to increase the number of usable chemical shifts datasets. From this improved quality of the datasets, we can start to build reliable statistics that should help us deciphering clear relationships between RNA structure and ^{13}C chemical shifts.

Materials and methods

Data mining

We collected all available ^{13}C chemical shifts of RNAs without binding partners from the BMRB (Seavey et al. 1991) (Table 1). Chemical shifts of six additional RNAs reported only in publications (Butcher et al. 1997; Jucker and Pardi 1995; SantaLucia and Turner 1993; Sich et al. 1997; Smith and Nikonowicz 1998; Szewczak and Moore 1995) and correctly referenced chemical shift data of six RNA stem-loops from our laboratory (unpublished) were added to the final database (Table 1). The secondary and tertiary structure of all datasets was extracted from the associated pdb coordinates, publications and the BMRB star file. The local structure of the terminal nucleotides of all RNAs was determined manually by analyzing the 3D structure using the pdb files, or from the secondary structure if the coordinates were not available. Subsequently, a script written in C++ was used to extract all available chemical shift values for each previously characterized nucleotide from the corresponding star files in the BMRB. These data were then converted into Microsoft Excel format. RNA chemical shift data from publications were entered manually.

Chemical shift correlations

Microcal Origin (Microcal Software Inc. MA) was used to create 2D scatter plots of chemical shift correlations. The expected chemical shift ranges for the five internal reference values (green boxes in Fig. 4) were defined as 138.7–139.7 ppm for C8 of 5′G, 136.4–137.6 ppm for C8 of 5′GG, 97.4–98.8 ppm for C5 of 3′C, 92.5–93.4 for C1′ of 3′C and 69.4–70.4 ppm for C3′ of a 3′C.

NMR measurements

NMR experiments were performed on AVANCE III (600 or 700 MHz) and AVANCE (900 MHz) Bruker spectrometers equipped with cryogenic probes. Unless indicated otherwise, spectra were recorded at 303 K. Six RNA stem loops with concentrations of 1.5–2.5 mM were used (their secondary structures are depicted in Supplementary Fig. 1 and their preparation is described in the Supplementary Text). With all RNA samples 2D ^{1}H-^{1}H TOCSY, 2D ^{1}H-^{13}C natural abundance HSQC and 2D NOESY spectra were recorded in D_2O and a 2D NOESY spectrum in H_2O. Typical parameters for the 2D NOESY experiments in D_2O were 48 scans, t_{1max} = 55 ms, 2,048 × 1,100 recorded data points, a mixing time of 250 ms and a relaxation delay of 1 s. Typical parameters for the 2D NOESY experiments in H_2O were 96 scans, t_{1max} = 33 ms, 2,048 × 1,000 recorded data points, a mixing time of 300 ms and a relaxation delay of 1 s. Typical parameters for the 2D ^{1}H-^{1}H TOCSY experiments were 4 scans, t_{1max} = 25 ms, 2,048 × 512 recorded data points, a mixing time of 50 ms and a relaxation delay of 1 s. The 2D ^{1}H-^{13}C natural abundance HSQC experiment was typically recorded with 220 scans, t_{1max} = 7.5 ms, 2,048 × 300 data points, and a relaxation delay of 1 s. For testing the influence of temperature on the chemical shifts, ^{1}H-^{13}C natural abundance HSQC spectra of stem-loop TASL2 were recorded at 283, 293, 303 and 313 K. Temperatures were calibrated using methanol-d_4 (>98.8% D, Armar AG, Switzerland) according to Findeisen et al. (Findeisen et al. 2007). The NMR spectra were processed with the software Topspin 2.1 (Bruker), and analyzed using the software SPARKY (Goddard and Kneller 1999). Spectra were referenced by an external sucrose/DSS sample which is described in detail in the Supplementary Material. The assignment of the six RNA stem-loops will be reported elsewhere.

Table 1 Datasets used for our analysis of chemical shift inconsistencies

I Correct datasets		II Partly correct or mis-calibrated datasets		III Datasets with unclear inconsistencies	
Correct datasets		a Unsystematic error		5007[d]	
4226		4780		5632	
4346		6062		5773	
5256		b Part of the data usable		6239	
5259		5170[d]		6320	
5371		5919		6756	
5655		5932		7090	
5705		(6485[e])		15538[d]	
5834[a]		15656[d]		15856	
5852		15786[f]		15859	
5962		1AFX		IV Datasets lacking internal reference values	
6076		c Shifted by 2.7 ppm			
6077		7403		4120	
6485[b] corrected		7404		4253	
6543		7405		4816	
7098		15869		5278	
15080		1SCL		5553	
15417		d Shifted by another value		5559	
15571		5703		6094	
15572		6633		6477	
15780		1YFV		6509	
15781				6562	
17RA				6652	
1UUU				15745	
New data				15081	
FZL2[c]				15858	
FZL4[c]				1RNG	
RP1[c]					
TASL1[c]					
TASL2[c]					
TASL3[c]					

[a] Deeper evaluation showed a systematic offset of all base ^{13}C chemical shifts of Ade and Ura nucleotides

[b] Corrected with the help of the original raw data

[c] Unpublished data from our laboratory

[d] Deeper evaluation showed a systematic offset of part of the data as illustrated for entry 15656 in Fig. 5

[e] Uncorrected file, was corrected and appears afterwards in category I

[f] BMRB entry 15786 contains two shift lists, list 2 falls in category II. Since no reference value for base carbons was assigned and the few assigned base chemical shifts are far away from typical values, base assignments were omitted in our final combined dataset. List 1 would fall in category III and is excluded

BMRB entries are listed by their entry number, datasets derived from publications are listed by their associated PDB code and datasets from our laboratory are listed by name. The five colored circles represent the 5 reference values in the order 5′G $^{13}C_{C8}$, 5′GG $^{13}C_{C8}$, 3′C $^{13}C_{C5}$, 3′C $^{13}C_{C1'}$, 3′C $^{13}C_{C3'}$. We applied a color code system to indicate if each individual reference value is in the expected range (*green*), systematically shifted (shifted by 2.7 ppm or shifted the same amount as another reference value) (*yellow*), not assigned (*black*), absent in the RNA sequence (*black*), outside the expected ranges and with no systematic errors that could be detected by the reference values (*red*). More details are found in Supplementary Table 2

BMRB accession codes

Chemical shifts of six newly assigned stem-loops were deposited in the BMRB under the accession numbers 17326, 17559, 17560, 17566, 17567 and 17568.

Determination of the sugar pucker

The backbone torsion angles δ were extracted from pdb files using the program AMIGOS (Duarte and Pyle 1998). δ angles between 130° and 190° were classified as C2′-endo (S-type) (Varani et al. 1996). δ angles between 50° and 110° were classified C3′-endo (N-type). These ranges were derived from high-resolution crystal structures, and are used in our laboratory (Oberstrass et al. 2006; Schubert et al. 2007). The δ angle range for C2′-endo is identical, and the range for C3′-endo is very similar to the angles described by Varani et al. (1996) (55°–115°). If the average of the δ angles of the structural ensemble lay in none of these regions, then the pucker was classified as unclear. Cases where the δ angles were found in the C3′-endo region that stand in contrast to experimental data indicating C2′-endo characteristics (e.g. H1′–H2′ couplings or a H1′–H2′ cross peak in the 2D ^{1}H–^{1}H COSY or 2D ^{1}H–^{1}H TOCSY spectrum) were also classified as ambiguous. Covariance ellipses were derived assuming an underlying bivariate normal distribution (Meyer 1975).

Results

Data mining and initial chemical shift analysis

Our initial aim was to perform a statistical analysis of ^{13}C RNA chemical shifts. We used all available BMRB entries containing ^{13}C data of RNA. To eliminate the influence of binding partners in our analysis, we excluded the chemical shift depositions of RNA complexes. This resulted in a database of 58 BMRB ^{13}C datasets. For our subsequent analysis, we added six datasets extracted from publications, and six unpublished datasets of RNA stem-loops, which were prepared for this work. All 70 entries are listed in Table 1. A simple two-dimensional plot of the ^{13}C versus ^{1}H chemical shifts of aromatic C6–H6 and C8–H8 pairs shows an interesting pattern (Fig. 1a). Guanine C8–H8, Adenine C8–H8 and pyrimidine C6–H6 are found in distinct regions. More surprisingly, it appears that within this grouping the peaks split into two clusters, which are separated by 2.5–3 ppm in the ^{13}C dimension (Fig. 1a). One explanation for these two clusters is that ^{13}C chemical shifts were calibrated using at least two different standards. RNA chemical shift data should be referenced like other biomolecules in aqueous solution to 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS). However, referencing to other standards like tetramethylsilane (TMS)—that is the general standard for substances in organic solvents (Fig. 2a)—was observed. In order to systematically analyze the datasets, we looked for chemical shifts that could serve as internal ^{13}C reference values in RNA.

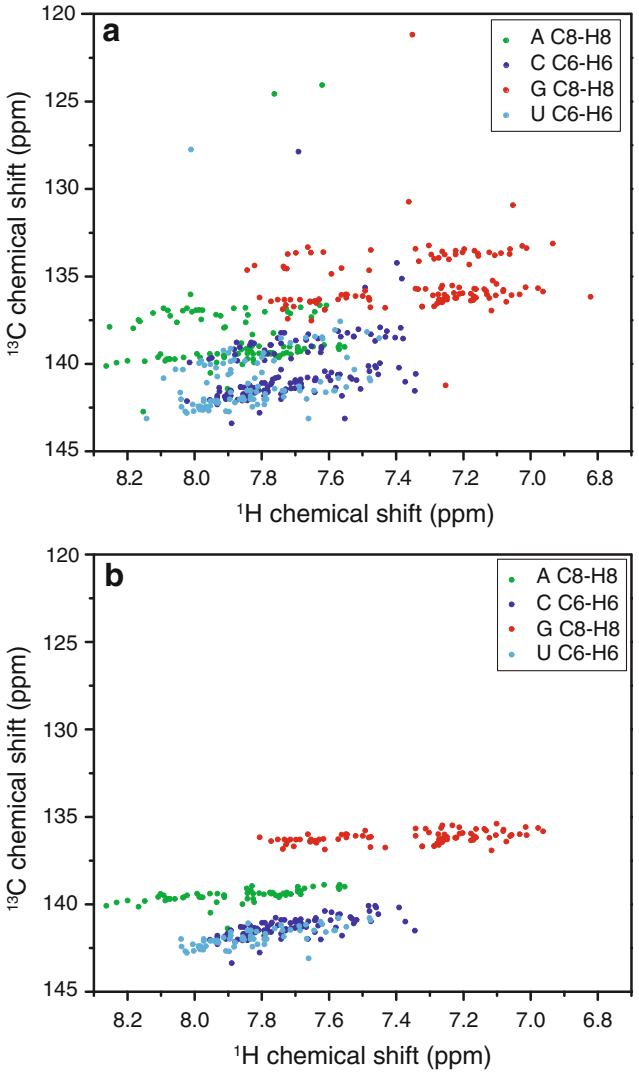

Fig. 1 RNA ^{1}H-^{13}C chemical shift correlations of bases in an A-form helix environment of the initial chemical shift data (**a**) and after validation and recalibration (**b**). Correlations of guanines are colored in *red*, of adenines in *green*, of cytosines in *blue* and of uracils in *cyan*. A/U chemical shifts of entry 5834 were excluded in the final chemical shifts (see footnote of Supplementary Table 2 for more details)

Selecting internal ^{13}C reference values for the chemical shift calibration

^{13}C chemical shifts of each nucleotide are highly dependent on the RNA sequence. Nevertheless we could find a set of five chemical shifts that are present in most RNA

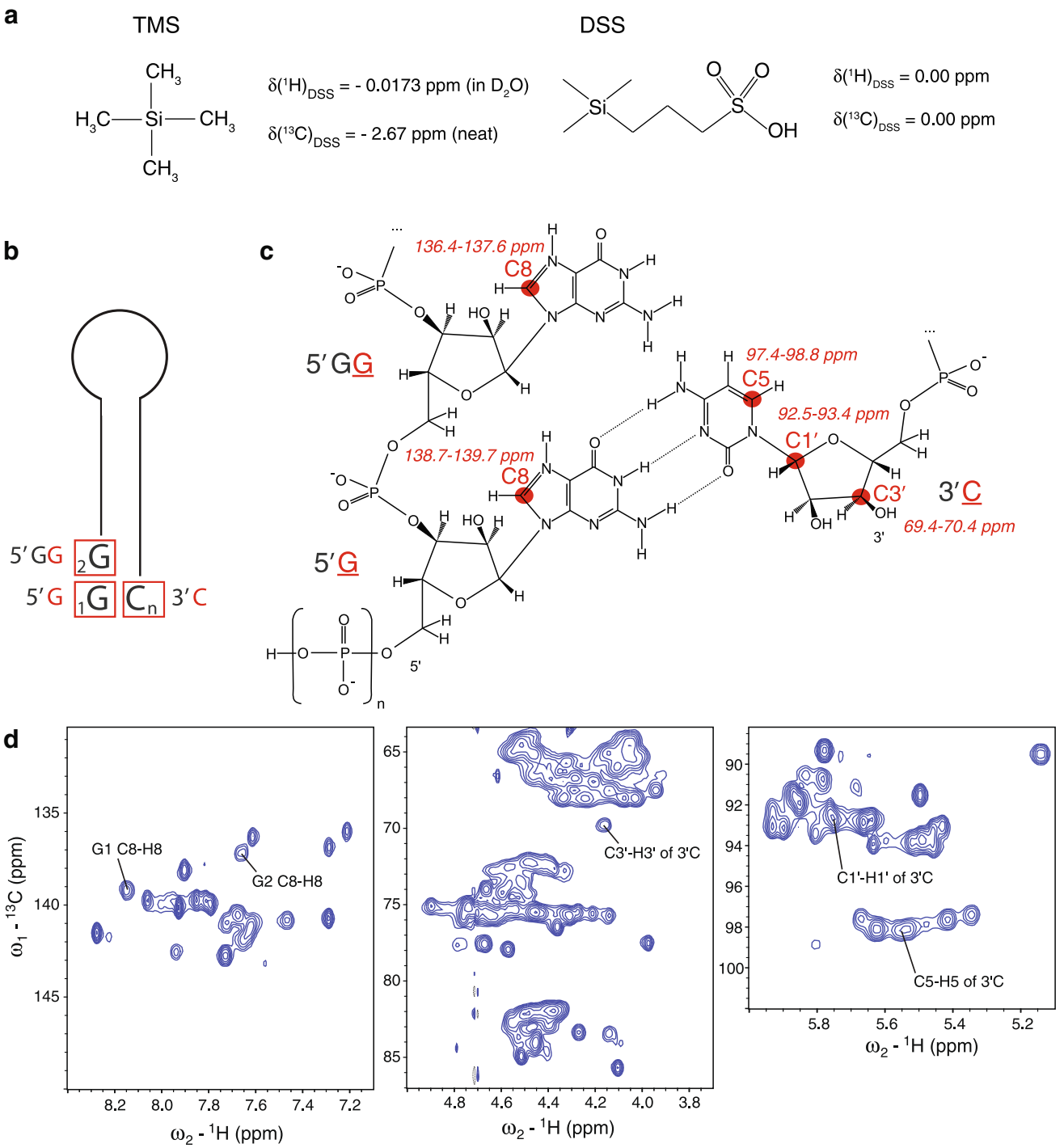

Fig. 2 Commonly occurring nucleotides that were used to extract chemical shift reference values. **a** Structures of the chemical shift reference standards tetramethylsilane (TMS) and 2,2-Dimethyl-2-silapentane-5-sulfonic acid (DSS). The chemical shifts of TMS in respect to DSS were reported previously (Markley et al. 1998; Morcombe and Zilm 2003). **b** Schematic structure displaying the involved nucleotides in an RNA stem-loop with two G–C closing base pairs. Chemical shifts of nucleotides in and adjacent to mismatches were not used as reference points. **c** Schematic atomic structure of the 5′- and 3′-end indicating the chemical shift reference values in *red*. **d** Regions of a ^{13}C-HSQC spectrum of the stem-loop FZL2 highlighting the C8 chemical shifts of two consecutive guanosines at the 5′-end, and the C3′, C5 and C1′ of a Watson–Crick base-paired cytosine at the 3′-end

datasets, and whose values are found in narrow shift ranges in the majority of the datasets. Therefore they are ideally suited as internal references to check the chemical shift calibration. The first two of these 'reference' ^{13}C chemical shifts are the C8 resonances of G1 and G2 found at the 5′-end of most RNAs prepared by in vitro transcription, and

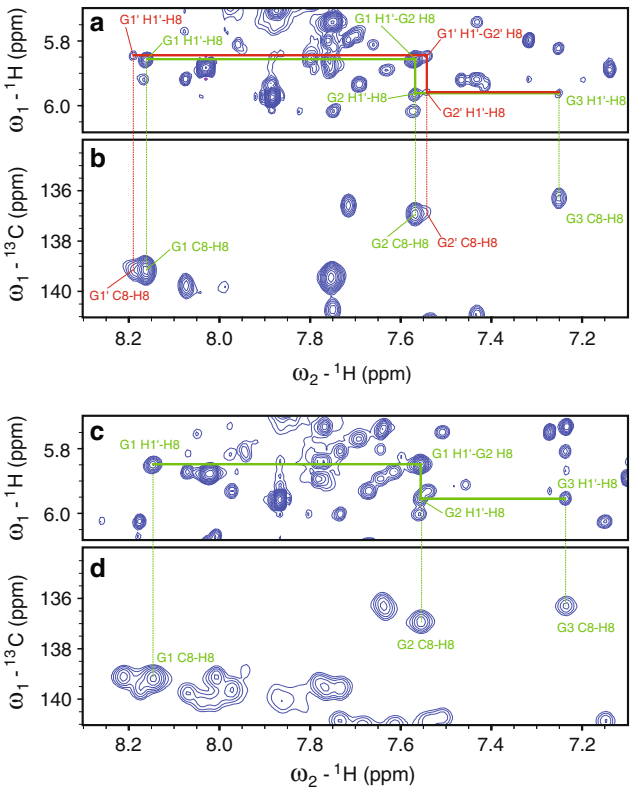

Fig. 3 Illustration of the effect of tri- versus monophosphate at the 5′-end on ^{1}H and ^{13}C chemical shifts. **a** and **b** 2D NOESY and ^{13}C-HSQC spectra of the RNA stem-loop TASL1 transcribed in the presence of GMP. The H6/H8-H1′ walk is indicated by *lines*. There are two sets of signals visible for the 5′-end corresponding to a terminal monophosphate (*red*) and triphosphate (*green*). The ^{13}C chemical shifts of both sets are virtually identical. **c** and **d** 2D NOESY and ^{13}C-HSQC spectra of the RNA stem-loop TASL3 transcribed in the absence of GMP. Only one set of signals is visible for the 5′-end corresponding to a terminal triphosphate (*green*)

denoted here as 5′G or 5′GG, respectively (Fig. 2b, c). Characteristic C8–H8 cross peaks occur at ~139.1/~8.15 ppm and ~137.0/~7.65 ppm in a ^{13}C-HSQC spectrum for 5′G and 5′GG, respectively (Fig. 2d). The terminal 5′G lacks a 5′ stacking neighboring base, thus resulting in a very distinct shift for its C8–H8 making it easily accessible. A mono- or a triphosphate at the 5′-end does not appear to modify the C8 chemical shift (within 0.1 ppm, see Fig. 3). Even a complete lack of phosphate, as found in chemically synthesized RNAs, does not significantly influence the $^{13}C_{C8}$ chemical shifts of the 5′G; the value is for example 138.8 ppm in entry 15571. Since GG is a frequently used starting sequence for RNA made by in vitro transcription, the ^{13}C C8 resonance of G2 is a good second reference value (5′GG) for most RNAs (44 out of 70). The third reference value is the C3′ ^{13}C chemical shift of the last 3′-nucleotide (Fig. 2b, c), which also occurs in a distinct position of a ^{13}C-HSQC spectrum (~69.9/~4.19 ppm, see Fig. 2d), because this nucleotide is lacking a phosphate at the 3′-end. This value is apparently independent of the 5′-neighbour. The fourth and fifth reference values are the C1′ and C5 chemical shifts of the 3′ terminal cytosine (3′C) involved in a Watson–Crick base pair with 5′G1 displaying ^{13}C values of ~92.9 ppm and ~98.1 ppm, respectively. In contrast to the other reference chemical shifts, these two resonances are not found in a very distinct region of the ^{13}C-HSQC spectrum (Fig. 2d) and a slight dependence of the 5′ neighbor might be possible. Nevertheless, these values are usually correctly assigned and can provide information to help detecting systematic errors in chemical shift datasets.

Correlations of internal reference values reveal correct calibration

In order to evaluate the ^{13}C calibration, we analyzed the chemical shift distributions of these five reference values in all collected ^{13}C RNA datasets using 2D correlation plots. Figure 4 shows four 2D correlations among the five references: between the two C8 of 5′G and 5′GG (Fig. 4a), between the C8 of 5′G and C3′ of 3′C (Fig. 4b), between the C1′ and C5 of the 3′C (Fig. 4c) and between the C1′ and C3′ of the 3′C (Fig. 4d). In all correlation plots the majority of the datasets cluster within ranges of about 1 ppm, indicating correct referencing (green boxes in Fig. 4).

However, several datasets present equally shifted carbon chemical shift values for both resonances, and therefore appear shifted along a line with a slope of 1 drawn in each figure. Along this line, a second cluster appears shifted by ~2.7 ppm in all four 2D plots (Blue box). This 2.66 ppm offset is likely to coincide with the ^{13}C chemical shift difference between 2,2-dimethylsilapentane-5-sulfonic acid (DSS) and tetramethylsilane (TMS). TMS is the default ^{13}C standard on Bruker spectrometers. However, biomolecules should be referenced via the absolute ^{1}H frequency of DSS multiplied by the ratio 0.251449530, yielding the absolute ^{13}C frequency of DSS which is then set to 0 ppm (Markley et al. 1998). Since ^{1}H chemical shifts of proteins are almost always calibrated correctly in contrast to heteronuclear data (Wang and Wishart 2005), we assume this holds true for RNA chemical shifts. The origin of this 2.66 ppm offset is described in more detail in the Supplementary Material. Although indirect chemical shift referencing was introduced as the standard for biomolecular NMR (Wishart et al. 1995), it is still not generally followed. However, this offset of 2.66 ppm can be easily corrected by a simple addition. When 2.66 ppm is added, all datasets lying in the blue box are found in the correct green box.

The origin for other calibration inconsistencies as depicted in Fig. 4 is not always clear. Since the C8 shifts of

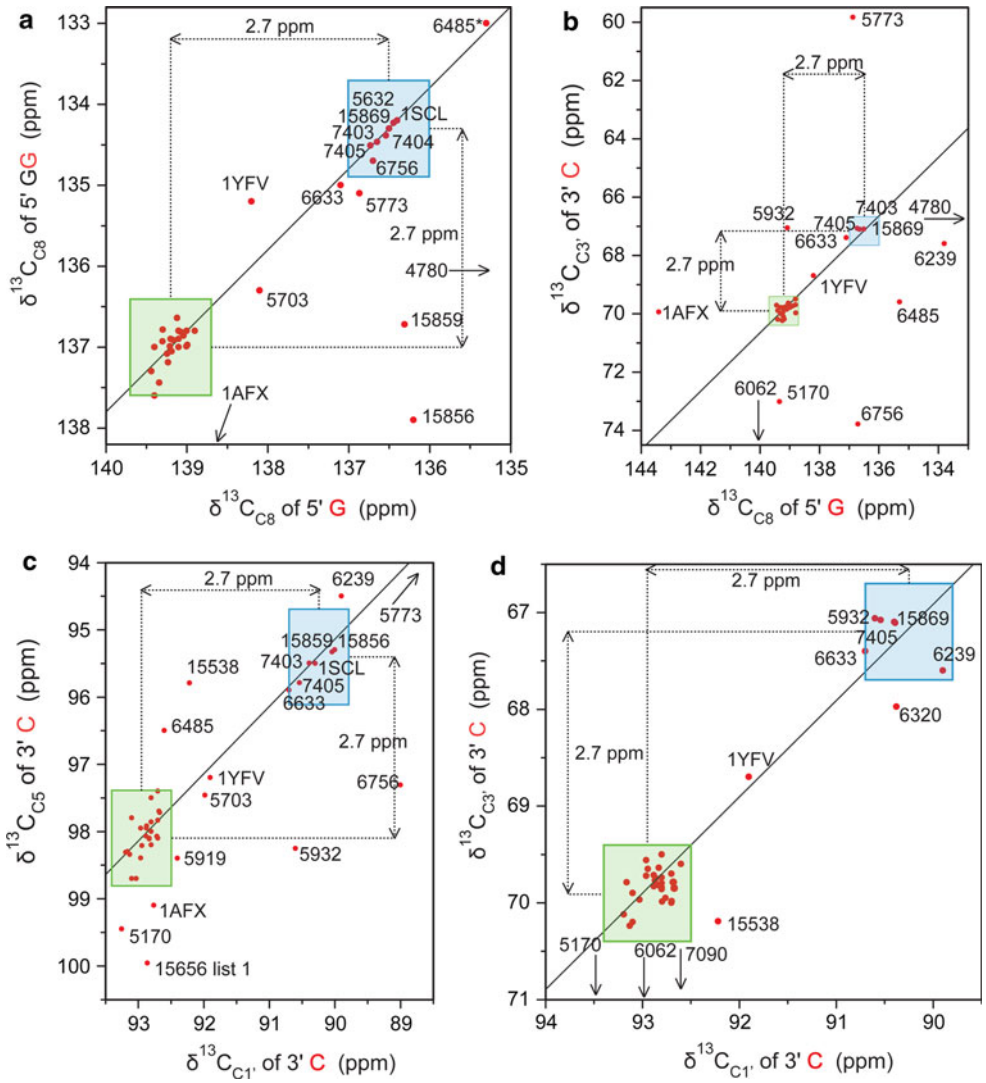

Fig. 4 Carbon-carbon chemical shift correlations of RNA chemical shift reference values of all ^{13}C RNA datasets (excluding the six unpublished datasets from our laboratory). **a** Correlations between C8 chemical shifts of Watson–Crick base paired guanosines at the 5′-end ($\delta^{13}C_{C8}$ 5′G) and C8 chemical shifts of Watson–Crick base paired guanosines following a guanosine at the 5′-end ($\delta^{13}C_{C8}$ 5′GG). **b** Correlations between C8 shifts of guanosines at the 5′-end ($\delta^{13}C_{C8}$ 5′G) and C3′ shifts of cytidines at the 3′-end ($\delta^{13}C_{C3'}$ 3′C). **c** Correlations between C1′ chemical shifts of guanosines at the 3′-end ($\delta^{13}C_{C1'}$ 3′C) and C5 chemical shifts of cytidines at the 3′-end ($\delta^{13}C_{C5}$ 3′C) and **d** Correlations between C1′ chemical shifts of cytidines at the 3′-end ($\delta^{13}C_{C1'}$ 3′C) and C3′ chemical shifts of cytidines at the 3′-end ($\delta^{13}C_{C3'}$ 3′C). The *green boxes* indicate the expected ranges for correctly calibrated reference values. The *blue boxes* are shifted by 2.7 ppm compared to the *green boxes*. The *black lines* have a slope of 1. For each off-diagonal data point either the BMRB entry number or the PDB code is indicated. A *black arrow* indicates data points lying outside the range of the figure

5′G and 5′GG are usually recorded in the same spectra, an off-diagonal correlation cannot originate from mis-calibration, and must therefore result from a mis-assignment (Fig. 4a). The same considerations are true for the sugar shifts of the 3′C (C1′ and C3′, Fig. 4c). Chemical shifts that could originate from two different spectra could potentially differ in calibration. In this case, a correlation away from the diagonal could be the result of two differently calibrated spectra, or from a mis-assignment. Such cases appear in Fig. 4b for 5′G C8—3′C C3′ correlations.

Experimental chemical shift of internal ^{13}C reference values

We transcribed six RNAs ranging from 20 to 30 nts (Supplementary Fig. 1) and assigned them by NMR spectroscopy. All internal reference values of those RNAs cluster in even narrower ranges within the green boxes. To verify that the chemical shifts of the internal referencing values stay within the defined tolerances (green boxes) under a variety of solution conditions, we measured spectra

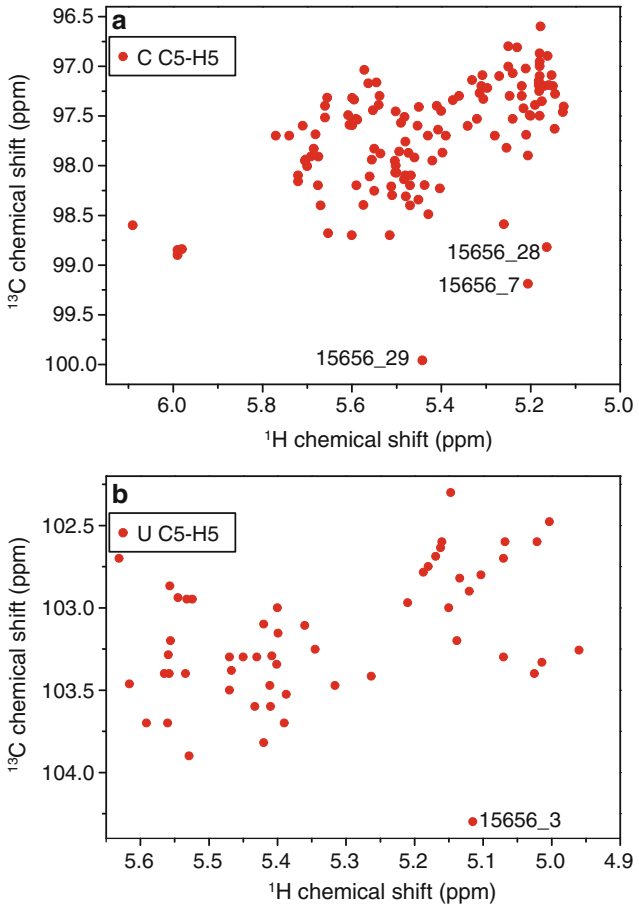

Fig. 5 H5-C5 correlations of cytosines (**a**) and uracils (**b**) in an A-helix environment of the corrected database containing categories I and IIa-c. The C5 reference value of the dataset 15,656 was outside the expected region. Here it is apparent that all C5 chemical shifts are systematically downfield shifted by ~2 ppm. This deeper analysis can help to distinguish systematic from nonsystematic errors

of the 26 nt stem-loop TASL2 at several temperatures ranging from 10 to 40°C, at several pH conditions ranging from 5 to 8, and at different NaCl concentrations ranging from 0 to 200 mM, with or without KH_2PO_4/K_2HPO_4 buffer. The five chemical shift reference values vary only within a small range (≤0.1 ppm compared to conditions at 30°C pH 6.0), and are therefore independent of temperature, pH and salt concentration (Supplementary Table 1). One exception is the small deviation observed for the C3′ ^{13}C of the 3′C which varies for low and high temperature by −0.2 ppm at 10°C and +0.2 ppm at 40°C. In addition, the C8 ^{13}C chemical shifts of the 5′G increases by +0.2 ppm at 200 mM NaCl. The following ranges were measured, namely 139.1–139.2 ppm for C8 of 5′G, 136.8–136.9 ppm for C8 of 5′GG, 97.9–98.2 ppm for C5 of 3′C, 92.8–92.9 for C1′ of 3′C and 69.8–69.9 ppm for C3′ of a 3′C. The ^{13}C chemical shifts were indirectly referenced to DSS (2,2-dimethyl-2-silapentane-5-sulfonic acid) according to the recommendations for biomolecules (Markley et al. 1998).

Correction of the chemical shift data

Forty-nine of the 64 RNA ^{13}C chemical shift datasets (without our 6 RNAs) contain at least two of the internal ^{13}C reference chemical shifts that allowed us to evaluate the calibration of these datasets (Table 1). We used a color code to indicate if each individual reference value is correct (green), either shifted by 2.66 ppm or diagonally shifted (yellow), is not assigned (black), absent in the RNA sequence (blank) or outside the expected ranges without detectable systematic error (red). For 23 datasets all assigned internal reference frequencies are lying within the expected chemical shift range, and are therefore counted correctly referenced (Table 1, category I). In addition we added six correctly referenced datasets from our laboratory which extend category I to 29 datasets. 17 datasets (category II) contained inconsistent shift values, but could be recovered by either detecting correct parts in the datasets or

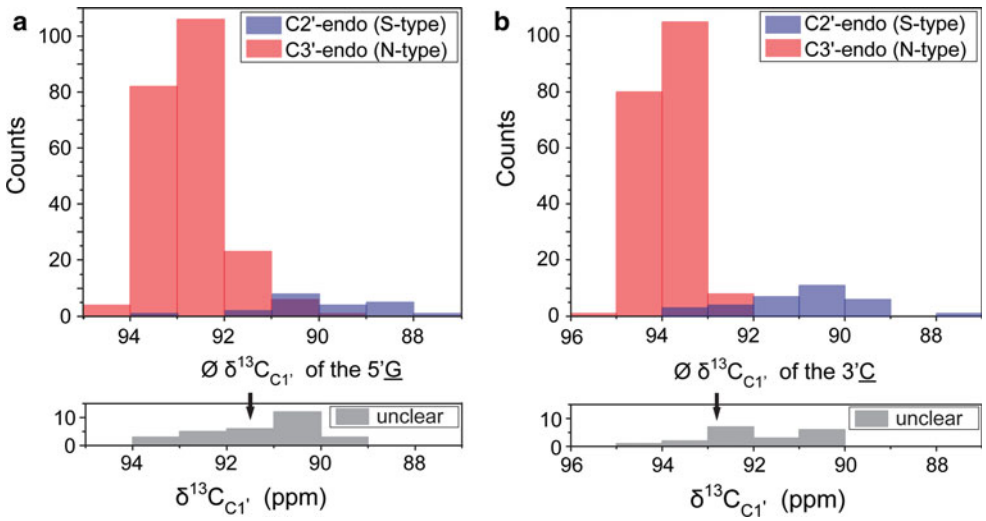

Fig. 6 Dependence of ^{13}C chemical shifts on the sugar pucker. *Histograms* showing the correlation between RNA C1′ ^{13}C chemical shifts of (**a**) purines and (**b**) pyrimidines with the sugar pucker conformation. Riboses adopting C3′-endo conformation are colored *red*; riboses adopting C2′-endo conformation are colored *blue*. Riboses with unclear sugar pucker conformation are colored *grey*. Average values found for the $\delta^{13}C_{C1'}$ of the 5′G and the $\delta^{13}C_{C1'}$ of the 3′C are indicated by *black arrows*

by recalibrating the datasets. There are two cases (category IIa) with a single outlier of more than 30 ppm indicating that the outlier is not systematic. Seven datasets (category IIb) have at least two reference values correctly referenced that were recorded in one spectrum. For example, two C8 shifts within the expected region strongly indicate that also the other C8/C6/C2 shifts of the RNA are likely to be correct, independent of whether or not the C1′ shifts are consistent. In 8 datasets (category IIc and IId), all the reference values are shifted by approximately the same value. The offset of the five datasets of category IIc can be explained by the improper calibration to TMS instead of DSS (blue boxes). While these datasets can be easily re-calibrated by adding 2.66 ppm to all ^{13}C chemical shifts, datasets of category IId require recalibration by a different offset. For 10 datasets (category III), the origin of the inconsistency is not clear from the reference values. Therefore, we did not attempt to recalibrate these datasets. 15 RNA datasets lacked our internal reference values (category IV) and could not be evaluated. This was either due to the absence of chemical shifts or the RNA termini differed from Fig. 2b. Comments for each individual case can be found in Supplementary Table 2.

To demonstrate the benefit of proper calibration, we show in Fig. 1b the corrected ^{13}C chemical shift values of the datasets of category I, IIa, consistent parts of category IIb and the recalibrated values of category IIc. The filtering and recalibration significantly improved the quality of the data, resulting in a much improved correlation between the C6/C8 and H6/H8 chemical shifts (Fig. 1b). The higher reliability and accuracy of the data revealed additional systematic inconsistencies that were not detected earlier. In one case we detected a systematic offset of C6/C8 chemical shifts of Ura/Ade that was not observed for Cyt/Gua bases (BMRB entry 5834). In another case (BMRB entry 15656, category IIb) in which the C5 reference resonance of 3′C was outside the expected range, all C5 chemical shifts are systematically shifted by ~2 ppm as illustrated in Fig. 5. For details, see footnotes of Supplementary Table 2.

Correct referencing of the ^{13}C chemical shift database results in better structure–chemical shift relationships: sugar pucker–^{13}C chemical shift correlations

It was shown earlier that the sugar pucker conformation influences the sugar ^{13}C chemical shift values (Ohlenschlager et al. 2008; Varani and Tinoco 1991). We wanted to determine whether we could now get a good correlation using our ensemble of corrected chemical shift data. For 29 datasets, we could also identify pdb files from which dihedral angles could be extracted. We first investigated the correlation between C1′ chemical shifts and the sugar pucker conformation. Purines and pyrimidines are treated separately because the type of base attached to the sugar affects the C1′ chemical shift. As shown in Fig. 6, purines and pyrimidines show clearly different C1′ ^{13}C chemical shifts depending on the sugar pucker. However, there is still some overlap between the different pucker states. Nucleotides in an exchange between the pucker states typically have intermediate chemical shifts (Varani and Tinoco 1991). This agrees with the observed chemical shifts of the C1′ shift of the 5′G and the C1′ shift of the 3′C, which are known to be in equilibrium between C2′- and C3′-endo conformations. The separation for the two sugar pucker conformations is similar to the ones found in a previous study using a linear combination of chemical

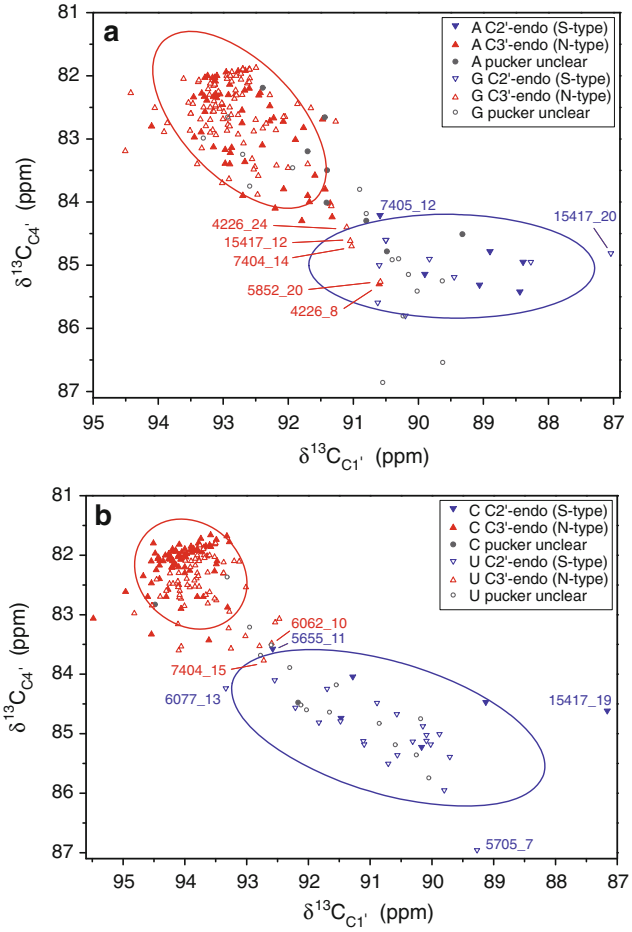

Fig. 7 2D ribose chemical shift correlations in dependence of the sugar pucker. C1′–C4′ chemical shift correlations of purines (**a**) and pyrimidines (**b**), respectively. C2′-endo conformers are colored *blue*, C3′-endo conformers are colored *red* and nucleotides with unclear pucker conformation are colored *grey*, respectively. Covariance ellipses at 2 standard deviations of a bivariate normal distribution (86 percent of the data points are supposed to lie in the corresponding ellipses) show a clear separation between C2′-endo and C3′-endo conformations. Labels indicate the BMRB and residue number of outlier data points

shifts optimized to get maximal separation (Ohlenschlager et al. 2008). In contrast to this mentioned study only one chemical shift is required here. An even better separation of the sugar puckers can be obtained by considering C1′–C4′ 2D correlations (Fig. 7). By assuming an underlying 2D Gaussian distribution we calculated the corresponding covariance ellipses at two standard deviations in which 86% of the data points are supposed to lie. A clear separation of the different sugar puckers for purines (Fig. 7a) and pyrimidines (Fig. 7b) was obtained. The chemical shifts of C3′ show also an obvious dependence on the sugar pucker whereas the C2′ does not (Fig. 8). Altogether the sugar puckers appear to be predictable on the basis of the C1′, C3′ or the C4′ chemical shifts. In addition the C2′–C3′ 2D plots allowed us to detect potential swapped assignments in the C2′ and C3′ chemical shifts of some sugar resonances (Fig. 8).

Discussion

The splitting of the ^{13}C chemical data into two clusters (Fig. 1a), as well as previously described problems caused by improper ^{13}C chemical shift calibration (Ohlenschlager et al. 2008), illustrate the importance of a validation procedure for deposited RNA ^{13}C chemical shifts. For validating proper referencing of ^{13}C resonances in RNA, we propose five internal chemical shift standards that are found in most RNA structures studied by NMR, and do not vary with solution conditions, two from guanines at the RNA 5′-end (C8 of 5′G and 5′GG) and three from a cytosine at the RNA 3′-end (C1′, C3′ and C5). Using these references, we found that only 22 datasets were correctly referenced and contained exclusively correct reference values. We were able to increase the number of usable datasets from 22 to 45 after corrections of several datasets and by adding six (Table 1). Among those, 8 datasets were recalibrated, 9 datasets were partially recalibrated (inconsistent parts were omitted) and 6 additional datasets were contributed from our laboratory. Improper calibration was the main source of errors. In a few cases a more detailed evaluation was necessary to distinguish systematic from non-systematic errors (Fig. 5). Overall, more than 50% of the published ^{13}C chemical shift data of RNAs are not properly calibrated, or contain obvious errors. This is much more than we expected since about 25% of wrongly calibrated datasets were reported for protein ^{13}C

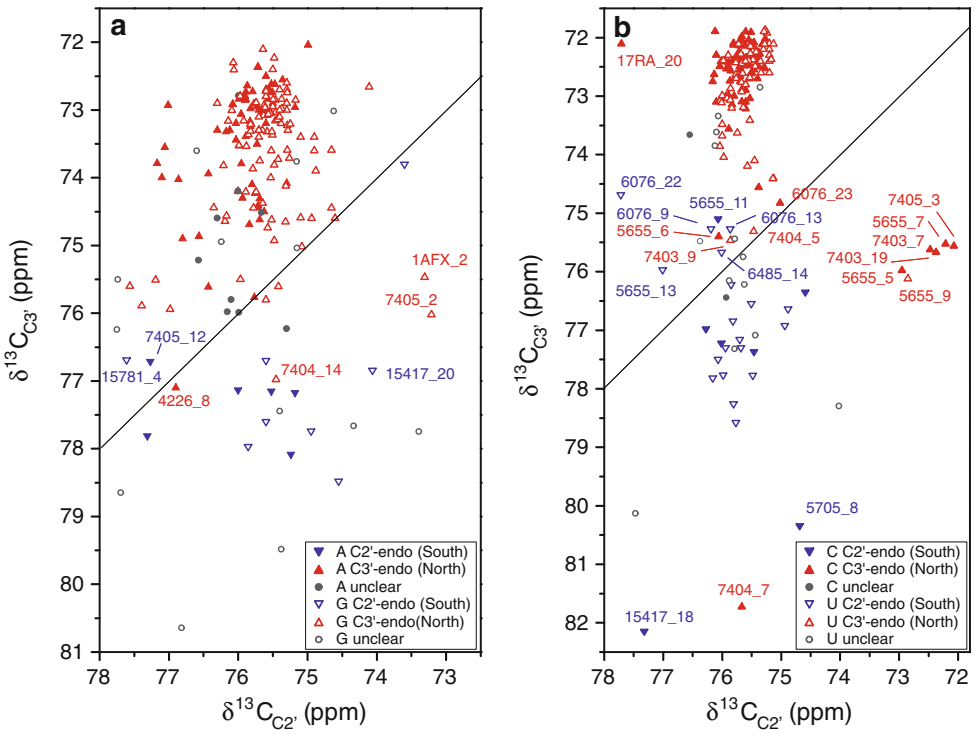

Fig. 8 2D ribose chemical shift correlations in dependence of the sugar pucker. C2′–C3′ chemical shift correlations of purines (**a**) and pyrimidines (**b**), respectively. The different sugar pucker conformations are nicely separated along the C3′-axis. No clear correlation with the C2′ chemical shift can be seen. Six pyrimidines show unexpected C2′ values between 72 and 73.5 ppm. Revision of the secondary structure reveals that most of these outliers can be found in an A-RNA helix environment where C3′ conformation is expected and swapping C2′ and C3′ chemical shifts values would bring both chemical shift values in the expected ranges. Since it is often complicated to unambiguously assign C2′ and C3′ chemical shifts it is very likely that these shifts were swapped during the assignment process

shifts (Zhang et al. 2003). Each individual dataset is mentioned in Supplementary Table 2. In contrast to the initial data, the ensemble of correctly calibrated and corrected data shows a clear clustering of chemical shifts depending on the residue type (Fig. 1) suggesting that the entire database can now be used to systematically analyze the dependence of ^{13}C chemical shift values on RNA sequence and structure. So far the presented method is limited to a subset of RNAs containing specific bases at the 3′ and 5′ ends that need to be base-paired. However, further analysis of the corrected ^{13}C database will reveal other typical chemical shifts suitable as internal reference values that could then be used to validate the ^{13}C calibration of RNAs with different termini or lacking assignments of the terminal nucleotides.

As a first application, we could use this corrected database by showing a clear correlation between the conformation of the sugar pucker and the C1′, C3′ or C4′ ^{13}C chemical shifts (Figs. 6, 7 and 8). In a previous study, Ohlenschlager et al. needed to use a linear combination of several ^{13}C ribose chemical shifts (Ebrahimi et al. 2001) to predict the sugar pucker conformations yielding ~95% correct predictions (Ohlenschlager et al. 2008). With our corrected database, we can obtain equally high prediction rates for the sugar pucker conformation by directly using ^{13}C C1′, C3′ or C4′ chemical shifts with no need of linear combinations. This method is simpler, and not dependant on a full assignment of the sugar. Furthermore, the three values can be used for independent confirmation.

In order to prevent the publication of improperly referenced RNA chemical shifts in the future, we suggest that the five internal reference shifts proposed here should be used as a method for validation of future depositions. We nevertheless would like to emphasize the importance of correct referencing according to the recommendations for biomolecules (Markley et al. 1998). Since proper indirect chemical shift referencing seems to be less established in the RNA-NMR community, we provide a detailed calibration procedure in the Supplementary Material to ensure proper referencing for future depositions into the BMRB. Improving the quality of the ^{13}C chemical shift data within the BMRB database should lead to more structure–chemical shift relationships for RNA that could be exploited to help resonance assignments, and facilitate RNA structure determination with NMR.

Acknowledgment We like to thank Olivier Duss for providing spectra of the two stem-loops FZL2 and FZL4, Wolfgang Bermel and Peter Schmieder for helpful discussions concerning chemical shift referencing. Further we are grateful to Peter Lukavsky for beneficial discussions of the C1′ chemical shift dependence on the ribose pucker and Fred Damberger for his comments on the manuscript. We thank Ryan Mackay and Lawrence P. McIntosh for their help regarding chemical shift calibration with Varian software. This work was supported by SNF-NCCR structural biology.

References

Butcher SE, Dieckmann T, Feigon J (1997) Solution structure of the conserved 16 S-like ribosomal RNA UGAA tetraloop. J Mol Biol 268:348–358

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620

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

Duarte CM, Pyle AM (1998) Stepping through an RNA structure: a novel approach to conformational analysis. J Mol Biol 284:1465–1478

Ebrahimi M, Rossi P, Rogers C, Harbison GS (2001) Dependence of 13C NMR chemical shifts on conformations of rna nucleosides and nucleotides. J Magn Reson 150:1–9

Fares C, Amata I, Carlomagno T (2007) 13C-detection in RNA bases: revealing structure-chemical shift relationships. J Am Chem Soc 129:15814–15823

Findeisen M, Brand T, Berger S (2007) A 1H-NMR thermometer suitable for cryoprobes. Magn Reson Chem 45:175–178

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

Goddard TD, Kneller DG (1999) SPARKY 3. University of California, San Francisco

Grzesiek S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15 N-enriched proteins. J Biomol NMR 3:185–204

Jucker FM, Pardi A (1995) Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. Biochemistry 34:14416–14427

Lam SL, Chi LM (2010) Use of chemical shifts for structural studies of nucleic acids. Prog Nucl Magn Reson Spectrosc 56:289–310

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wuthrich K (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. J Biomol NMR 12:1–23

Meyer SL (1975) Data analysis for scientists and engineers. Wiley, New York

Morcombe CR, Zilm KW (2003) Chemical shift referencing in MAS solid state NMR. J Magn Reson 162:479–486

Mulder FA, Filatov M (2010) NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination. Chem Soc Rev 39:578–590

Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH (2006) Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. Nat Struct Mol Biol 13:160–167

Ohlenschlager O, Haumann S, Ramachandran R, Gorlach M (2008) Conformational signatures of 13C chemical shifts in RNA ribose. J Biomol NMR 42:139–142

SantaLucia J Jr, Turner DH (1993) Structure of (rGGCGAGCC)2 in solution from NMR and restrained molecular dynamics. Biochemistry 32:12612–12623

Schubert M, Labudde D, Oschkinat H, Schmieder P (2002) A software tool for the prediction of Xaa-Pro peptide bond conformations in proteins based on 13C chemical shift statistics. J Biomol NMR 24:149–154

Schubert M, Lapouge K, Duss O, Oberstrass FC, Jelesarov I, Haas D, Allain FH (2007) Molecular basis of messenger RNA recognition by the specific bacterial repressing clamp RsmA/CsrA. Nat Struct Mol Biol 14:807–813

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

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

Sich C, Ohlenschlager O, Ramachandran R, Gorlach M, Brown LR (1997) Structure of an RNA hairpin loop with a 5′-CGUUUCG-3′ loop motif by heteronuclear NMR spectroscopy and distance geometry. Biochemistry 36:13989–14002

Smith JS, Nikonowicz EP (1998) NMR structure and dynamics of an RNA motif common to the spliceosome branch-point helix and the RNA-binding site for phage GA coat protein. Biochemistry 37:13486–13498

Szewczak AA, Moore PB (1995) The sarcin/ricin loop, a modular RNA. J Mol Biol 247:81–98

Varani G, Tinoco I (1991) Carbon assignments and heteronuclear coupling-constants for an Rna oligonucleotide from natural abundance C-13-H-1 correlated experiments. J Am Chem Soc 113:9349–9354

Varani G, Aboulela F, Allain FHT (1996) NMR investigation of RNA structure. Prog Nucl Magn Reson Spectrosc 29:51–127

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

Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Methods Enzymol 338:3–34

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

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD (1995) 1H, 13C and 15 N chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502

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