Chemical shifts-based similarity restraints improve accuracy of RNA structures determined via NMR

CHAD LAWRENCE1 and ALEXANDER GRISHAEV1,2
1Institute for Bioscience and Biotechnology Research, Rockville, Maryland 20850, USA
2Biomolecular Measurement Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

ABSTRACT

Determination of structure of RNA via NMR is complicated in large part by the lack of a precise parameterization linking the observed chemical shifts to the underlying geometric parameters. In contrast to proteins, where numerous high-resolution crystal structures serve as coordinate templates for this mapping, such models are rarely available for smaller oligonucleotides accessible via NMR, or they exhibit crystal packing and counter-ion binding artifacts that prevent their use for the chemical shifts analysis. On the other hand, NMR-determined structures of RNA often are not solved at the density of restraints required to precisely define the variable degrees of freedom. In this study we sidestep the problems of direct parameterization of the RNA chemical shifts/structure relationship and examine the effects of imposing local fragmental coordinate similarity restraints based on similarities of the experimental secondary ribose $^{13}$C/$^1$H chemical shifts instead. The effect of such chemical shift similarity (CSS) restraints on the structural accuracy is assessed via residual dipolar coupling (RDC)-based cross-validation. Improvements in the coordinate accuracy are observed for all of the six RNA constructs considered here as test cases, which argues for routine inclusion of these terms during NMR-based oligonucleotide structure determination. Such accuracy improvements are expected to facilitate derivation of the chemical shift/structure relationships for RNA.

Keywords: RNA structure; chemical shifts; accuracy; cross-validation

INTRODUCTION

Biomolecular NMR plays an important role in structural studies of oligonucleotides, accounting for ~40% of RNA-only models in the Protein Data Bank (PDB; Berman et al. 2000). However, determination of RNA structure via NMR presents a number of challenges compared to proteins. One of major contributors is the large number of variable torsion angles ($\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, $\zeta$, $\chi$) per nucleotide compared to the ($\phi$, $\psi$) pair for each amino acid. Another is that commonly used interproton distance restraints are less effective for the often elongated oligonucleotide geometries. In protein NMR, structure definition challenges were ameliorated via tight restraints on backbone dihedral angle from the experimental chemical shifts, in combination with precise orientational restraints such as residual dipolar couplings (RDCs; Tolman et al. 1995; Tjandra and Bax 1997). This approach led to the increase in coordinate accuracy while also extending applicability to larger proteins (Raman et al. 2010). The caveat of the procedure is the necessity of establishing a precise relationship between the experimental chemical shifts and three-dimensional coordinates. This challenge has been addressed via knowledge-based analyses of chemical shifts from proximal sites, taking into account both structural and primary sequence similarities, relying on extensive databases of high-resolution protein crystal structures and curated complete sets of $^1$H/$^{13}$C/$^{15}$N chemical shifts. Modern software tools are capable of backbone torsion angle prediction with precision and accuracy approaching 10° (Berjanski et al. 2006; Cheung et al. 2010; Shen and Bax 2013; Hafsa et al. 2015), with such restraints now nearly universal for NMR protein structure determination.

As chemical shifts represent one of the most readily accessible and precisely defined NMR observables, similar developments would be advantageous for oligonucleotides. Coordinate-based prediction of the RNA/DNA $^1$H, $^{13}$C and $^{15}$N chemical shifts is indeed achievable via...
quantum chemistry tools (Fonville et al. 2012; Sahakyan and Vendruscolo 2013; Suardiaz et al. 2013; Swails et al. 2015; Jin et al. 2016). In complement, a number of semi-empirical approaches have been developed using additive contributions of ring currents, local magnetic anisotropies, and induced electric fields (Wijmenga et al. 1997; Sahakyan and Vendruscolo 2013; Suardiaz et al. 2013), with applications for both validation and refinement of RNA structures (Frank et al. 2013a; van der Werf et al. 2013; Sripakdeepong et al. 2014). Prediction of the RNA $^1$H and $^{13}$C chemical shifts from structure is also possible via machine learning empirical methods (Frank et al. 2013b, 2014), with model validation applications. Increases in the amount of RNA NMR data also led to development of empirical chemical shift prediction tools based solely on the primary sequence and the secondary structure (Barton et al. 2013; Brown et al. 2015), finding use for validation and automation of NMR resonance assignments (Aeschbacher et al. 2013).

**Chemical shifts remain underutilized in NMR structural studies of RNA**

In light of the considerable body of work aimed at interpretation of the nucleic acids’ chemical shifts, it could seem surprising that they were almost never used quantitatively for determining RNA structure. To a degree, the challenges of inverting the structure/chemical shift relationship for RNA reflect high density of aromatic rings in nucleic acids and widespread occurrence of chemical shift referencing imperfections (Aeschbacher et al. 2012). Ultimately, the main challenge is severe shortage of accurate structural models that could be used for chemical shifts analysis. Out of approximately 240 RNA-only chemical shift data sets in the Biological Magnetic Resonance Bank (BMRB) (Ulrich et al. 2008), fewer than 10 can be associated with crystal structures in the PDB, in result limiting the analysis to models determined via solution NMR. However, even though approximately 200 such correspondences can be established from the BMRB and PDB databases, a majority of those do not exhibit restraint density sufficient for precise specification of the variable torsion angles, with average respective restraint uncertainties of ~30°. In result, empirical RNA structure-based chemical shift parameterizations so far have been optimized from sparse sets of only 20–30 structures (Cromsigt et al. 2001; Frank et al. 2013b, 2014), compared to the case of proteins with primary databases of up to 600 crystal structures and booster data sets containing approximately 9500 additional models (Shen and Bax 2010). In consequence, limited accuracies of modeling both the ring current effects for the $^1$H shifts and the torsion angle dependencies of the $^{13}$C and $^{31}$P chemical shifts have been noted as factors adversely impacting their use for structural analysis (Frank et al. 2013a; Brown et al. 2015; Swails et al. 2015).

Information contained in the NMR chemical shifts of RNA can be processed directly if converted to structural similarity restraints

Considering the challenges of linking NMR chemical shifts with RNA structure, increasing coordinate accuracy of the NMR-determined RNA models and, thus expanding the set of reliable chemical shift/structure pairs, is an issue of utmost importance. Accuracy of the NMR-determined structures can be readily assessed via RDCs due to their steep dependence on the orientations of the respective inter-nucleotri vectors (Simon et al. 2005). For the structures refined against the RDC data, cross-validation statistics averaged over subsets of RDCs excluded from the refinement (Clore and Garrett 1999) can be monitored via free Q- or R-factor metrics, analogous to the free R-factors used in crystallography (Brünger 1992), allowing reliable discrimination between closely related models (Chen and Tjandra 2011).

We explore an idea of directly exploiting patterns of chemical shifts similarities to improve RNA structural accuracy independently of any preestablished structure/chemical shifts relationships. Specifically, we refine RNA structures applying pseudoenergy terms that enforce agreements between the internal coordinates of local fragments exhibiting similar $^{13}$C/$^1$H ribose chemical shifts, in essence lowering the number of model’s degrees of freedom according to the observed chemical shifts. The accuracy of the resulting models can then be gauged against reference calculations in which such terms are absent. The notion of a link between the similarity of the oligonucleotide chemical shifts and the related structural parameters was originally expressed in the NMR analysis of the Dickerson dodecamer B-DNA (Tjandra et al. 2000), and subsequently used in a study of an A-form RNA helix (PDB deposition 2GBH; O’Neil-Cabello et al. 2004). This logic is also the driving force behind TALOS programs for protein chemical shifts-based torsion angle prediction (Cornilescu et al. 1999). Noncrystallographic symmetry (NCS) restraint terms that can enforce fragmental similarities are commonly used for NMR refinement of multimeric assemblies. However, they can be configured in a more flexible manner after removal of source code safeguards restricting them to matching primary sequences. We set out to determine whether introduction of the chemical shift-based fragment similarity restraints produces consistent favorable effects on the RNA structural accuracy. As this approach does not involve direct refinement against chemical shifts data, the resulting models would remain perfectly appropriate for subsequent derivations of the chemical shifts/structure relationships.

To establish a set of test cases, we analyzed all of the approximately 90 NMR-determined structures of RNA in the PDB that include ribose C–H RDCs among the deposited restraints, selecting those associated with close-to-
such corrections were consistent within 10%. We have also
determined the average $^1$H/$^1$C ribose chemical shift val-
ues for each of the four common RNA nucleotide types
swapped the $2/3'$ assignments for a small fraction of sites
that exhibited $C_3'/C_5'$ resonance flips. For the RNA $H_2'/H_5''$
chemical shifts reported in the BMRB, only 30% corre-
spond to the resolved stereo-specifically assigned pairs.
This subset forms two clusters in the 3.9–4.7 parts per mil-
ion (p.p.m.) range, indicating possible stereo-misassign-
ments (Supplemental Fig. S1). Approximately two thirds
of such resonance pairs with $^1$H shifts differing by more
than 0.35 p.p.m. constitute the dominant downfield/up-
field $H_2'/H_5''$ cluster. In light of this likely stereo-ambiguity,
the analyses were carried out for the downfield and upfield
5' proton signals instead. Chemical shifts data for the A/G/
C/U nucleotides selected according to the above criteria
were further subjected to two cycles of outlier removal
outside of three standard deviations. The resulting accu-
mulated statistics are reported in Table 1. As expected,
substantial nucleotide type dependence is exhibited by
the 1' sites, with approximately 1 p.p.m. and 0.3 p.p.m.
differences between the $C_3'$ and $H_1'$ shifts in purines versus
pyrimidines.

### RESULTS AND DISCUSSION

**Analysis of the database-deposited RNA chemical shifts reveals nucleotide type dependencies and allows corrections for referencing and assignment imperfections**

Chemical shifts-based structure analysis commonly in-
volves corrections for the residue type dependence, as-
suming additive contributions of the chemical bonding
and the tertiary structure effects. Such “secondary” chem-
ical shifts are naturally formulated for RNA setting the
canonical A-form as the reference state. Therefore, we have
determined the average $^1$H/$^1$C ribose chemical shift val-
ues for each of the four common RNA nucleotide types
from the BMRB-deposited data. Our analysis was restrict-
ed to the A-form Watson–Crick paired bases similarly
flanked on both sides. In agreement with earlier assess-
ments (Aeschbacher et al. 2012), we find ~40% of the
RNA-only chemical shift data sets in the BMRB misrefer-
cenced in the carbon dimension. In line with earlier rec-
ommendations, we have corrected the $^1$C data using the 5
GC and 3'-CC ribose marker chemical shift patterns when
such corrections were consistent within 10%. We have also

| Nt type | $C_3'$ | $C_2'$ | $C_1'$ | $C_4'$ | $C_5'$ | $H_1'$ | $H_2'$ | $H_3'$ | $H_4'$ | $H_5'$/$H_5''$ upfield | $H_5'$/$H_5''$ downfield |
|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------------------------|------------------------|
| A       | 92.95  | 75.65  | 72.83  | 82.13  | 65.20  | 5.93   | 4.59   | 4.58   | 4.47   | 4.13                   | 4.49                   |
| G       | 92.84  | 75.45  | 72.91  | 82.11  | 65.75  | 5.74   | 4.56   | 4.51   | 4.48   | 4.12                   | 4.45                   |
| C       | 93.94  | 75.54  | 72.35  | 82.04  | 64.59  | 5.52   | 4.40   | 4.42   | 4.40   | 4.06                   | 4.50                   |
| U       | 93.64  | 75.47  | 72.45  | 82.22  | 64.62  | 5.53   | 4.43   | 4.47   | 4.44   | 4.08                   | 4.51                   |

Statistics were extracted from the RNA data deposited at the BMRB.

### Patterns of the $C_5'$ chemical shifts support differences in strengths of the purine and the pyrimidine C–H…O$_5$' hydrogen bonds in A-RNA

Dependence of the C/H shifts on the nucleotide type
generally decreases from the 1' to the 4' sites, consistent
with the increasing distances to the nucleobases and
the diminishing chemical bonding and ring current effects.
Surprisingly, $C_5'$ resonances exhibit purine/pyrimidine differ-
ences nearly equaling those observed for the $C_1'$ sites,
while $H_2'/H_5''$ chemical shifts show little nucleotide type
dependence. We propose that this phenomenon reflects
shorter and stronger intranucleotide $C_6'-H_6...O_5'-C_5'$
C–H/O hydrogen bonds for the A-RNA pyrimidines com-
pared to the $C_6'-H_6...O_5'-C_5'$ hydrogen bonds in purines.
NMR evidence for such purine/pyrimidine differences was previously suggested based on the marked differen-
ces between the chemical shift tensors for the $C_6$ and $C_5$
base sites in A-RNA versus B-DNA (Ying et al. 2006).
Shortening of the pyrimidine $H_6...O_5'$ distances in A-
RNA compared to the purine $H_6...O_5'$ is apparent from
the analysis of crystal structures as well, becoming more
pronounced with increasing resolution of the diffraction
Therefore, our observation of C₅′ purine/pyrimidine chemical shift differences can be considered supporting evidence of the systematic differences in the corresponding C–H…O hydrogen bonds observable via the O₅′-C₅′ sites.

Analysis of the nucleotide-type corrected chemical shifts for the six tested RNA constructs produces significant numbers of novel fragmental similarity restraints

Raw ribose ¹³C/¹H chemical shifts for the six tested RNA constructs were corrected for the nucleotide type dependence as specified in Table 1. Pairwise similarities of the resulting secondary chemical shift vectors were calculated according to Equation 1 with the α parameter set to 1.0, placing emphasis on carbon data’s decreased dependence on the interresidue ring currents. Nucleotide-type corrected chemical shift similarity maps for the tested RNA constructs are shown in Figure 1. The active NCS restraints were selected as those below the CSS threshold of 0.2 p.p.m., compared to the maximum observed CSS values of 0.8 p.p.m.–3.3 p.p.m. The majority of the CSS-paired sites with similar local geometries involve canonical Watson–Crick base-paired nucleotides (Fig. 2). Overall, selected sets of similarity restraints include a small fraction of all possible nucleotide pairs in the sequences, ranging between 2% and 7%.

Using the structure of the UUCG tetraloop as an example (PDB entry 2KOC), the following nucleotides are paired up based on the similarity of the secondary chemical shifts (see Supplemental Fig. S3): 3/4, 3/11, 3/12, 3/13, 4/11, 4/12, 4/13, 11/13, and 12/13. These pairings enforce close geometric similarities within a cluster of nucleotides including 3, 4, 11, 12, and 13 (see Fig. 2; Supplemental Fig. S3). Compared to the ≈0.1 p.p.m. r.m.s. deviations between the secondary chemical shifts within this cluster, all of the remaining nucleotides exhibit unique patterns of chemical shifts differing by as much as 3.3 p.p.m. r.m.s. As they are determined by the chemical connectivity, these patterns need to be subtracted from the chemical shifts data in order to reveal the structural effects we are aiming to capture.

Reference calculations excluding chemical shift similarity restraints suggest variation of structural accuracy depending on the density of NMR restraints and the extent of conformation dynamics

The initial rounds of reference calculations against the distance and the torsion angle restraints starting from the PDB-deposited coordinates for the six test cases produced structures exhibiting poor agreement with the excluded RDCs (Q-factors of 0.7–1.4). Such results are generally representative of the NOE-based models (Bax and Grishaev 2005), and demonstrate that the preparatory simulation stages accomplish their purpose. Refinements including RDC restraints for the reference sets of calculations without the chemical shifts-based NCS terms resulted in structures exhibiting lower free Q-factors ranging between 0.78 for
the loop domain of 1S34 and 0.37 for 2RN1 (Table 2). High average $Q_{\text{free}}$ for the 1S34 loop relative to the respective value of 0.59 for the stem, as well as the twofold difference between their $Q_{\text{fit}}$ values are consistent with previous findings of the loop’s increased conformational dynamics (Cabello-Villegas et al. 2004), suggesting inadequacy of the single-structure representation for this part of the construct. The difference between the eigenvalues of the alignment tensors for the stem and loop domains of 1S34, with the respective magnitudes and rhombicities $(D_a/R)$ of $-16.8$ Hz/0.47 and $-13.1$ Hz/0.55, is consistent with conformational variability of the interfacial flip-out base of A7. Aside from the 1S34 entry, cross-validation statistics of the reference NCS-free structure calculations appear fairly uniform with the free $Q$-factors ranging from 0.37 to 0.52. Among the two best cross-validating entries, 2KOC and 2RN1, the former is characterized by an exceptional density of high-precision torsion angle restraints derived from an extensive set of cross-correlated spin relaxation rates and homo- and hetero-nuclear scalar couplings, while the latter exhibits the highest density of the fitted 1-bond C–H RDCs (4.2 per nt) of all tested cases. These observations suggest that RNA coordinate accuracy is correlated with the ratio of the number of NMR observables to the number of the degrees of freedom necessary to describe the structural model.

**Introduction of the chemical shifts-based similarity restraints significantly improves structural accuracy as measured via residual dipolar coupling cross-validation**

Except for the loop domain of the 1S34 entry, introduction of the CSS-based restraints brings substantial improvement in cross-validation statistics with relative decreases in $Q_{\text{free}}$ ranging between 14% and 38%, and ending with the free $Q$-factors varying between 0.26 and 0.44 (Table 2; Fig. 3). Negligible cross-validation improvement for the 1S34 loop domain likely reflects low density of CSS-based restraints acting upon it (one out of six total). In addition to the previously noted well-cross-validating 2KOC and 2RN1 constructs which, with NCS restraints included, exhibit free $Q$-factors of 0.26 and 0.32, respectively, 1N8X and 1XHP now also cross-validate with low $Q_{\text{free}}$ values of 0.34. Significant accuracy improvements in these cases are likely aided by particularly high numbers of close chemical shift matches and the corresponding active CSS restraints (see Table 3). Application of the CSS-derived pairwise NCS terms leads to effectively identical geometries for the corresponding nucleotide pairs (coordinate r.m.s.d. values below 0.01 Å). Superimposed structures calculated with and without CSS restraints are shown in Supplemental Figure S5.

**TABLE 2.** RDC fit and cross-validation statistics for the tested RNA constructs

| PDB ID | $Q_{\text{fit}}$ without similarity restraints | $Q_{\text{fit}}$ with similarity restraints | $Q_{\text{free}}$ without similarity restraints | $Q_{\text{free}}$ with similarity restraints |
|--------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 1N8X   | 0.009 ± 0.001                   | 0.037 ± 0.004                   | 0.474 ± 0.045                   | 0.344 ± 0.046                   |
| 1S34-stem | 0.028 ± 0.001                   | 0.028 ± 0.001                   | 0.593 ± 0.050                   | 0.374 ± 0.030                   |
| 1S34-loop | 0.056 ± 0.006                   | 0.059 ± 0.007                   | 0.779 ± 0.113                   | 0.762 ± 0.087                   |
| 1XHP   | 0.025 ± 0.001                   | 0.033 ± 0.001                   | 0.480 ± 0.048                   | 0.345 ± 0.034                   |
| 2KOC   | 0.047 ± 0.001                   | 0.047 ± 0.004                   | 0.419 ± 0.078                   | 0.261 ± 0.047                   |
| 2O33   | 0.116 ± 0.007                   | 0.109 ± 0.004                   | 0.516 ± 0.067                   | 0.440 ± 0.059                   |
| 2RN1   | 0.036 ± 0.001                   | 0.038 ± 0.001                   | 0.373 ± 0.052                   | 0.320 ± 0.036                   |
Comparison of the empirical validation parameters for the structures refined with and without the CSS-derived geometric similarity terms (Supplemental Table S3) indicates that the predominant effect of enforcing the CSS lies in improvement of the fraction of nucleotide “suites” within well-recognized rotameric states (Murray et al. 2003; Williams et al. 2018). Clash statistics remain largely unchanged, dominated by the specifications of the non-bonded force field and the overall quality of the experimental distance restraints. Similarly, no systematic changes are observed for the coordinate precision.

CSS restraints act to limit the number of the degrees of freedom in the RNA structural model in agreement with the experimental chemical shifts

Improvement of the RNA coordinate accuracy with the introduction of the fragment similarity restraints suggests that restriction of the models’ degrees of freedom in this manner reflects real responses of chemical shifts to the structural changes. These observations are consistent with early analyses of the effect of NCS restraints on the structural accuracy of medium-resolution protein crystal structures (Kleywegt and Jones 1995). Application of the chemical shift-based similarity restraints is straightforward and comes with no additional experimental or computational costs, as any NMR-based structure analysis implies chemical shift availability. Moreover, poor ribose chemical shift dispersion that generally complicates formulation of the NMR restraints becomes an advantage with this approach, as it increases the number of enforceable close chemical shift matches. Since the NCS-based restraints involve similarities rather than the specific chemical shift values, the resulting structures remain appropriate for studying chemical shifts/torsion angles relationships, simply representing more accurate variants of the original models. As the procedure used for obtaining the similarity restraints compares chemical shifts within the same spectral data set, it is unaffected by any possible imperfections of the absolute chemical shift referencing or pulse sequence effects, again facilitating its routine use. The price for this simplicity of application is the necessarily limited amount of information associated with such chemical shift processing, when compared to direct mapping of the chemical shifts to the torsion angles currently achievable for proteins.

Further improvements of structural accuracy are likely achievable with expansion and continued curation of the BMRB, the public database of RNA chemical shifts

Interpretation of the RNA chemical shifts through the proximal torsion angles as done here neglects the impact of the remote nucleobases. Earlier studies concluded that such effects are generally small for the ribose carbon sites (Frank et al. 2013b), while dependences of the ribose 1H shifts on types of the neighboring nucleotides are certainly

| PDB ID | Total nts | Total 1-bond C–H RDCs/nt | Fitted 1-bond C–H RDCs/nt | Fitted lower-precision RDCs/nt | Lower-precision RDC types | Fitted distance restraints/nt | Fitted dihedral angle restraints/nt | Number of CSS-based similarity restraints | Number of fitted/cross-validated RDC sets |
|--------|-----------|--------------------------|---------------------------|-------------------------------|----------------------------|--------------------------------|-------------------------------|------------------------------------------|----------------------------------------|
| 1N8X   | 36        | 0.72                     | 0.67                      | N.A.                          | N.A.                       | 10.6                           | 3.3                           | 45                                        | 13                                     |
| 1S34   | 23        | 3.61                     | 3.17                      | 1.04                          | 1-bond C–C                | 18.4                           | 4.3                           | 6                                         | 8                                      |
| 1XHP   | 32        | 2.38                     | 2.22                      | N.A.                          | N.A.                       | 15.7                           | 7.6                           | 25                                        | 15                                     |
| 2KOC   | 14        | 2.07                     | 2.00                      | N.A.                          | N.A.                       | 18.4                           | 11.9                          | 6                                         | 29                                     |
| 2O33   | 20        | 2.95                     | 2.65                      | 0.55                          | 1-bond C–C, 2-bond C–H   | 24.7                           | 6.2                           | 13                                        | 10                                     |
| 2RN1   | 32        | 2.78                     | 2.50                      | 1.69                          | 1-bond C–H                | 14.2                           | 5.3                           | 13                                        | 10                                     |
possible (Barton et al. 2013). Indeed, ribose H1′ and H2″ sites (PDB nomenclature) are positioned closest to the bases of their 3′ or 5′ neighbors in A-RNA. However, inspection of the sites included in the CSS-based restraints reveals that their H1′ atoms are generally located 5–6 Å away from the centers of the nearest aromatic rings of their 5′ neighbors, relatively far considering the r−3 distance dependence of the ring current terms. This observation is consistent with earlier findings of independence of the H1′ shifts on the nature of both 3′ and 5′ neighboring bases in A-RNA (Cromsigt et al. 2001). On the other hand, H2″ sites in the NCS-restrained residues are positioned ~3 Å from the centers of their 3′ neighboring nucleobases and could be impacted by the corresponding ring currents. Earlier studies, however, have not identified clear correlation of the H2″ shifts with the nature of the 3′ neighbors in A-RNA (Cromsigt et al. 2001; Barton et al. 2013). Fine-tuning of the CSS terms to suppress these effects should be possible via additional corrections of the reference state chemical shifts for the nearest neighbor effects.

As demonstrated by the results for 1N8X and 1XHP entries, increases in the density of the CSS restraints likely bring additional improvements in the structural accuracy. The general idea of the approach outlined here offers ample potential for further development, such as further fine-tuning of the contributions of the individual carbon and proton sites within the CSS score. We also propose that the CSS-based restraints may be applicable even between independently refined constructs, similar to the manner they were used here for the two domains of 1S34. Refinement of a cluster of RNA structures linked by the CSS restraints would be a novel concept that has not been previously considered in biomolecular NMR. A significant advantage of this approach is that it could bring additional increases in the density of the active CSS-based restraints proportionally to the number of structures within the co-refined set. Since the CSS-based strategy relies on the detection of close chemical shift matches, development of the interconstruct fragmental similarity restraints would require establishing precise chemical shift referencing for the separate entries, as well as corrections for small effects brought by pulse sequence variations (Aeschbacher et al. 2012; Brinson et al. 2019). For the six test cases considered here, ribose 1H chemical shifts are consistent with BMRB statistics. In contrast, two of the studied constructs—1N8X and 1XHP exhibit sizeable shifts in ribose 13C referencing, in agreement with earlier findings (Aeschbacher et al. 2012; Frank et al. 2013b). For the 1XHP entry, the offset appears correctable as the five ribose carbon resonance types are shifted by similar amounts (2.6 p.p.m.–1.9 p.p.m.). In contrast, resonance offsets of the five ribose carbon types for 1N8X are not consistent as they range between 6.2 p.p.m. and 13.7 p.p.m., thus precluding data re-referencing. This issue does not impact restraint generation as carried out in this study but would need to be addressed before any joint analysis of the chemical shifts data can be performed. The strategy outlined here does not appear directly extendable to the 13C/1H chemical shifts of the nucleobases due to a complex interplay of the effects of the nearby aromatic ring systems. However, structural accuracy improvements resulting from application of the ribose CSS terms coupled with better descriptions of the base-base and base-backbone interactions may facilitate better empirical parameterization of the ring currents, opening up a way toward more routine use of the nucleobase chemical shifts in RNA structure determination.

**Future avenues for improvement include better characterization of lower-populated conformers**

Even with the improved structural accuracy brought by the CSS restraints, derivation of the torsion angle/chemical shifts mapping for RNA is expected to be challenging. Variation of the torsion angles within the main A-form rotamer that encompasses 75% of the nucleotides is limited (r.m.s. of 8° or less for the main chain torsions and 3° or less for the ring torsions in sub-2.7 Å resolution RNA crystal structures in the PDB). Such tight definition of the major rotameric cluster (3emmtp3 in the δ−1/ε−1/ζ−1/α/β/γ/δ suite notation of Murray et al. 2003) would require very high precision of chemical shift modeling in order to yield nontrivial structural restraints; however, this analysis may be aided by the large fraction of data included in it. Ribose 1H/13C chemical shifts are likely to be more informative for separating the discrete RNA backbone rotamers, in which case the main challenge is the low population of such states relative to the dominant A-form. The small number of suitable RNA constructs considered in this study does not allow for making conclusions regarding the mapping of the chemical shifts to structural parameters; however, we can evaluate sensitivity of the secondary chemical shifts to variation of the local structure.

Analysis of the ribose secondary 13C shifts for the two best-cross-validating constructs (2KOC and 2RN1) reveals that one-third of the outliers constitute 3′- or 5′- terminal nucleotides, associated with chain termination effects and likely increased conformational dynamics (Cromsigt et al. 2001). Canonical A-form nucleotides with ribose secondary 13C chemical shifts near zero exhibit tight clustering with standard deviations from 0.1 p.p.m. to 0.6 p.p.m. In order to investigate structural origins of the remaining outliers, we have carried out structure refinement with active CSS terms while fitting all experimental NMR restraints for the 2KOC and 2RN1 constructs, followed by the torsion angle analysis.

For the 2KOC construct, nonterminal ribose secondary chemical shift outliers are observed for nucleotides U7–G10 (Supplemental Fig. S3), corresponding to the last three nucleotides of the UUCG tetraloop and the 3′-
flanking Watson–Crick base-paired G10. Considering tight definition of the respective backbone torsion angles (Supplemental Table S1), chemical shift deviations from the values typical for the A-form RNA reflect the 3′emtmp2′, 2′emtmp2′, and 2′epttp3′ rotamers of the three torsion angle suites within the tetraloop (Murray et al. 2003). C2′-endo pucker of the nucleotides U7 and C8 is thus associated with upfield 4–5 p.p.m. shifts for the C1′, downfield 5–8 p.p.m. shifts for the C3′, and downfield 2–5 p.p.m. shifts for the C4′ resonances. C2′ resonances are least affected by the unusual rotamers of the UUCG tetraloop, with 2 p.p.m. downfield secondary shifts for C8 and G9, while C5′ resonances exhibit fairly uniform downfield secondary shifts of 3–4 p.p.m. for U7–G10. These outlier 13C chemical shifts are also expected to reflect altered base stacking and unusual glycosidic χ torsions suggesting that ribose 13C resonance positions respond to RNA backbone rotamer switches in a complex and coordinated manner, which would likely preclude simple analysis of the isolated torsion angles based on small contiguous clusters of local chemical shifts.

For the 2RN1 construct, nonterminal secondary ribose 13C chemical shift outliers are observed for the nucleotides C5–U7, G20–C23, and G27 (Supplemental Fig. S4). Out of these, stretches C6–U7 and G21–C23 make up the domain interface of the TAR-TAR*GA kissing complex. The original report of the 2RN1 structure (Van Melckebeke et al. 2008) highlighted the unusual C4′-exo pucker for A11 and G21, as well as the C2′-endo pucker for G20, with the rest of the nonterminal sites confined to the common C3′-endo conformation. Analysis of the structure bundles generated with the CSS restraints indicates that, rather than corresponding to the unusual but precisely determined pucker states, nucleotides C5–C6, A11–G12, G21–U22, and A28–29 exhibit greatly increased backbone torsions angle uncertainties (Supplemental Table S2). Inspection of the corresponding torsion angle restraints reveals that all of these sites are associated with either completely unrestrained α, β, and γ torsions, or their restraint uncertainties of approximately ±100° (G6, G12, G20–U22, A28–C29). Compared to the remaining ±30° torsion angle restraints representing the canonical A-RNA conformation, these omissions were meant to allow greater conformational freedom for both the noncanonical G21–A28 base pair and the interface between the TAR and TAR*GA domains. However, the resulting decreased structural precision near these sites complicates the analysis of the nearby unusual chemical shifts. On the other hand, structure/chemical shifts analysis for 2RN1 highlights two discrepancies that appear readily addressable. Nucleotide U7 belongs to the canonical A-form suite 3′emtmp3′, inconsistent with its extreme ~12 p.p.m. downfield C4′ resonance and warranting further investigation. Similarly, nucleotide G12, while poorly defined structurally as stated above, displays ribose chemical shifts entirely consistent with the canonical A-form rotamer 3′emtmp3′. These examples highlight benefits attainable with closer integration of the RNA chemical shifts data in structure refinement.

Concluding remarks

We propose routine use of the chemical shifts-based fragment similarity restraints described here in all structural studies of RNA using solution NMR data. Such applications are best performed while monitoring cross-validation statistics, most straightforwardly accomplished in cases when RDC restraints are available. Keeping in mind that one of the payoffs of the better accuracy of the NMR-derived RNA models is the possibility of improved NMR structure-based mapping between the torsion angles and the chemical shifts, further applications can be selected from approximately 90 structures of RNA constructs in the PDB that include RDCs as deposited restraints. Such study would require access to the corresponding data sets of chemical shifts, ~40% of them not deposited to the BMRB, and a joint effort of the RNA NMR community.

The strategy outlined in this study readily lends itself to a number of extensions outlined above, paving a way toward better integration of experimental ribose and nucleobase chemical shifts in the RNA structure determination.

MATERIALS AND METHODS

Ribose 13C and 1H chemical shifts corresponding to the PDB depositions 1N8X, 1S34, 1XHP, 2KOC, 2RN1, and 2033 were obtained from the BMRB database (entries 5773, 6062, 6320, 5705, 11,014, and 15,080, respectively). RNA-only data entries deposited to the BMRB were also analyzed to construct the nucleotide-type specific chemical shift distributions for the ribose H1′, H2′, H3′, H4′, H5′, H6′, C1′, C2′, C3′, C4′, and C5′ sites based on the data recorded at temperatures between 15°C and 35°C. In total, the assembled set included 537 A, 781 G, 577 C, and 548 U nucleotides from the 78 entries listed in the Supplemental Material. Raw ribose chemical shifts for the six tested RNA constructs were corrected for the residual systematic differences between the nucleotide types calculated from the BMRB-extracted statistics. The agreement between the vectors of such secondary proton and carbon ribose chemical shifts was quantified for all nucleotide pairs via the chemical shift similarity (CSS) index:

\[
\text{CSS} = \sqrt{\frac{\sum_{m=1}^{N_{1H}} (\Delta \delta_{1H}^m)^2 + \sum_{j=1}^{N_{13C}} (\alpha^{-1} \Delta \delta_{13C}^j)^2}{N_{1H} + N_{13C}}}
\]

where \(N_{1H}\) and \(N_{13C}\) are the numbers of 1H and 13C ribose chemical shifts of the same type reported for both nucleotides within a given pair, \(\Delta \delta\) values denote the differences between the respective secondary chemical shifts, and the factor \(\alpha\) balances the impact of the 1H and 13C data. Nucleotides at the 3′ and 5′ chain ends were excluded from the CSS index evaluation due to differences in chemical bonding and base stacking, as well as likely
increased conformational dynamics that would preclude their representation by a single conformation. In cases when \( H_5' \) and \( H_8'' \) resonances were reported as unambiguously assigned, their chemical shifts were used as deposited. Otherwise, the shifts were flipped to result in the downfield \( H_5' \) and the upfield \( H_8'' \). Nucleotide pairs with the lowest CSS scores were selected for application of the respective similarity restraints via the NCS terms.

RNA structure refinements were carried out with the experimental NMR distance, torsion angle, and RDC restraints associated with the PDB depositions 1N8X, 1S34, 1XHP, 2KOC, 2RN1, and 2O33. The most precise 1-bond C–H and N–H RDC restraints were split into multiple exhaustive sets of fitted and cross-validated couplings (Table 3). In the cases when the experimental RDCs included less precisely determined coupling types, the corresponding restraints were fitted but not cross-validated against, with the corresponding fit force constants scaled down to reflect their higher uncertainties. Such examples included 1-bond base and ribose C–C RDCs for the 1S34 entry, 1-bond C–C and 2-bond C–H base RDCs for the 2O33 entry, and lower-precision 1-bond C–H RDCs for the 2RN1 entry (see Table 3 for details). Flexible linkage of the stem and loop domains previously noted only terms representing nonbonded interactions with the atomic reality-specifying improper dihedrals, as well as quartic repulsive–attractive potentials for the paired nucleotides, \( P \) atoms were removed from the corresponding NCS terms. Similarly, when neither \( C_5' \)/\( H_8'' \)/\( H_8'' \) chemical shifts were not reported for at least one of the paired nucleotides, \( P \) atoms were removed from the NCS term. The above-defined NCS term atom selections are illustrated in Supplemental Figure S6.

RDC values in each cross-validated set were predicted from the final structures using the alignment tensors determined from the fitted RDC set. Cross-validation statistics are reported via free Q-factors calculated as

\[
Q_{\text{free}} = \sqrt{\frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} (\text{RDC}_{i,\text{val,obs}} - \text{RDC}_{i,\text{val,pred}})^2}
\]

The numerator of the expression contains the sum over the cross-validated (inactive) set of RDCs with the predicted values calculated from the alignment tensor fitted to the active RDCs, and the denominator—the corresponding magnitude \( D_{\text{shift}} \) and rhombicity \( R_{\text{fit}} \). Fitted and free Q-factor statistics are reported as the averages and the uncertainties of the averages over the multiple sets of fitted/cross-validated RDCs used for each entry.

Supplemental Material includes the list of RNA entries from BMRB used for extraction of the chemical shift statistics, distributions of the backbone torsion angles in the entries 2KOC and 2RN1 with the structures refined against the CSS-based terms and all RDC data fitted, structure statistics for the models refined with and without the CSS terms (Supplemental Tables S1, S2, S3, respectively), clustering of the stereo-specifically assigned \( H_5' \) and \( H_8'' \) chemical shifts from the BMRB statistics (Supplemental Fig. S1), distributions of the nucleobase/backbone C–H…O hydrogen bond donor/acceptor distances in the crystal structures of RNA in the PDB (Supplemental Fig. S2), the secondary chemical shifts of the ribose carbon sites in the 2KOC and 2RN1 entries (Supplemental Figs. S3, S4, respectively), overlays of the structures calculated with and without the CSS-based terms (Supplemental Fig. S5), and illustration of the atom sets used for specification of the NCS terms (Supplemental Fig. S6).

Disclaimers

Certain commercial equipment, instruments, and materials are identified in this paper in order to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.
SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

Received December 30, 2019; accepted August 19, 2020.

REFERENCES

Aeschbacher T, Schubert M, Allain F. 2012. Database proton NMR chemical shifts for RNA signal assignment and validation. J Biomol NMR 55: 33–46. doi:10.1007/s10858-012-9683-9

Bax A, Grishaev A. 2005. Weak alignment NMR: a hawk-eyed view of RNA chemical shifts. Nucleic Acids Res 41: e172. doi:10.1093/nar/gkt665

Barton S, Heng X, Johnson BA, Summers MF. 2013. Crystallography and NMR System: a new software suite for biomolecular structure determination. J Mol Biol 425: 12175. doi:10.1016/j.jmb.2011.09.052

Berjanski M, Neal S, Wishart D. 2006. PREDITOR: a web server for predicting protein backbone dihedral angles. Nucleic Acids Res 34: W63–W69. doi:10.1093/nar/gkl341

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28: 235–242. doi:10.1093/nar/28.1.235

Brünger AT, Adams P, Clore GM, DeLano W, Gros P, Grosse-Kunstleve R, Bickelhaupt FM, Wijmenga SS. 2012. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. Nucleic Acids Res 40: 6271–6282. doi:10.1093/nar/gks596

Brünger AT. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355: 472–475. doi:10.1038/355472a0

Brünger AT, Adams P, Clore GM, DeLano W, Gros P, Grosse-Kunstleve R, Jiang JS, Kuszewski J, Nilges M, Pannu N, et al. 1998. Crystallography and NMR System: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54: 905–921. doi:10.1107/S0907444998003254

Cabello-Villegas J, Giles K, Soto A, Yu P, Mougin A, Beemon K, Wang YX. 2004. Systematic study of RNA NMR chemical shifts: a structure-based approach. J Mol Biol 340: 1217–1237. doi:10.1016/j.jmb.2004.01.033

Chen K, Tjandra N. 2011. Use of residual dipolar coupling in studying proteins by NMR. Top Curr Chem 326: 47–68. doi:10.1007/128_2011_215

Cheung M, Maguire M, Stevens T, Broadhurst R. 2010. DANGLE: a Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. J Magn Reson 202: 223–233. doi:10.1016/j.jmr.2009.11.008

Clare GM, Garrett D. 1999. R-factor, free R, and complete cross-validation for dipolar coupling refinement of NMR structures. J Am Chem Soc 121: 6836–6837.

Comisescu 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:10.1007/A1008 392405740

Cromsigt J, Hilbers CW, Wijmenga SS. 2001. Prediction of proton chemical shifts in RNA. J Biomol NMR 21: 11–29. doi:10.1023/A :1011914132531

Fonville JM, Swart M, Vokacova Z, Syrchovskv Y, Sponer JE, Sponer J, Hilbers CW, Bickelhaupt FM, Wijmenga SS. 2012. Chemical shifts in nucleic acids studied by density functional theory calculations and comparison with experiment. Chem Eur J 18: 12372–12387. doi:10.1002/chem.201103593

Frank AT, Horowitz S, Andricioaei I, Al-Hashimi HM. 2013a. Utility of 1H NMR chemical shifts in determining RNA structure and dynamics. J Phys Chem B 117: 2045–2052. doi:10.1021/jp308636c

Frank AT, Bae SH, Stelzer AC. 2013b. Prediction of RNA 1H and 13C chemical shifts: a structure-based approach. J Phys Chem B 117: 13497–13506. doi:10.1021/jp407254m

Frank AT, Law SM, Brooks CL. 2014. A simple and fast approach for predicting 1H and 13C chemical shifts: towards chemical shift-guided simulations of RNA. J Phys Chem B 118: 12168–12175. doi:10.1021/jp08342

Hafsa N, Arndt D, Wishart D. 2015. CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucl. Acids Res 43: W370–W377. doi:10.1093/nar/gkv494

Jin X, Zhu T, Zhang J, He X. 2016. A systematic study of RNA NMR chemical shift calculation based on the automated fragmentation QM/MM approach. RSC Adv 6: 108590–108602. doi:10.1039/C6RA22518G

Kleywegt GJ, Jones TA. 1995. Where freedom is given, liberties are taken. Structure 3: 535–540. doi:10.1016/S0969-2126(01)00187-3

Lawrence C, Stover C, Noznitsky J, Wu Z, Summers M. 2003. Structure of the intact stem and bulge of HIV-1 Psi-RNA stem loop SL1. J Mol Biol 326: 529–542. doi:10.1016/S0022-2836(02)01305-0

Murray L, Arendall WB, Richardson DC, Richardson JS. 2003. RNA backbone is rotameric. Proc Natl Acad Sci 100: 13904–13909. doi:10.1073/pnas.1835769100

Nozinski S, Fürtig B, Jonker H, Richter C, Schwalbe H. 2009. High-resolution NMR structure of an RNA model system: the 14-mer cUUCCgg tetraloop hairpin RNA. Nucleic Acids Res 38: 683–694. doi:10.1093/nar/gkp956

O’Neil-Cabello E, Bryce D, Nikonorwicz E, Bax A. 2004. Measurement of five dipolar couplings from a single 3D NMR multiplet applied to the study of RNA dynamics. J Am Chem Soc 126: 66–67. doi:10.1021/ja033814k

Raman S, Lange O, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot T, Eletsky A, Szypersky T, et al. 2010. NMR structure determination for larger proteins using backbone-only data. Science 327: 1014–1018. doi:10.1126/science.1183649

Sahakyan AB, Vendruscolo M. 2013. Analysis of the contributions of ring current and electric field effects to the chemical shifts of RNA bases. J Phys Chem B 117: 1989–1998. doi:10.1021/jp3057306

Sashital D, Comisescu G, McManus C, Brow D, Butcher S. 2004. U2-U6 RNA folding reveals a group II intron-like domain and a four-helix junction. Nat Struct Mol Biol 11: 1237–1242. doi:10.1038/nmb863

Sashital D, Venditti V, Angers C, Comisescu G, Butcher S. 2007. Structure and thermodynamics of a conserved U2 snRNA domain from yeast and human. RNA 13: 328–338. doi:10.1261/rna .418407

Schwieters C, Kuszewski J, Tjandra N, Clore GM. 2003. The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160: 66–74. doi:10.1016/S1090-7807(02)00014-9

Schwieters C, Kuszewski J, Clore GM. 2006. Using Xplor-NIH for NMR molecular structure determination. Progr NMR Spectrosc 48: 47–62. doi:10.1016/j.pnmrs.2005.10.001

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

Simon K, Xu J, Kim C, Skrynnikov N. 2005. Estimating the accuracy of protein structures using residual dipolar couplings. *J Biomol NMR* 33: 83–93. doi:10.1007/s10858-005-2601-7

Sripakdeepong P, Cevc M, Chang A, Erat M, Ziegler M, Zhao Q, Fox G, Gao X, Kennedy S, Kierzek R, et al. 2014. Structure determination of noncanonical RNA motifs guided by $^1$H NMR chemical shifts. *Nat Methods* 11: 413–418. doi:10.1038/nmeth.2876

Suardiaz R, Sahakyan A, Vendruscolo M. 2013. A geometrical parametrization of C1′–C5′ ribose chemical shifts calculated by density functional theory. *J Chem Phys* 139: 034101. doi:10.1063/1.4811498

Swails J, Zhu T, He X, Case DA. 2015. AFNMR: automated fragmentation quantum mechanical calculation of NMR chemical shifts for biomolecules. *J Biomol NMR* 63: 125–139. doi:10.1007/s10858-015-9970-3

Tjandra N, Bax A. 1997. Direct measurement of distance and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278: 1111–1114. doi:10.1126/science.278.5340.1111

Tjandra N, Tate S, Ono A, Kainosho M, Bax A. 2000. The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. *J Am Chem Soc* 122: 6190–6200. doi:10.1021/ja000324n

Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH. 1995. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc Natl Acad Sci* 92: 9279–9283. doi:10.1073/pnas.92.20.9279

Van der Werf R, Tessari M, Wijmenga SS. 2013. Nucleic acid helix structure determination from NMR proton chemical shifts. *J Biomol NMR* 56: 95–112. doi:10.1007/s10858-013-9725-y

Van Melckebeke H, Devany M, Di Primo C, Beurain F, Bryce D, Boisbouvier J. 2008. Liquid-crystal NMR structure of HIV TAR RNA bound to its SELEX RNA aptamer reveals the origins of the high stability of the complex. *Proc Natl Acad Sci* 105: 9210–9215. doi:10.1073/pnas.0712121105

Wijmenga SS, Kruthof M, Hilbers SW. 1997. Analysis of $^1$H chemical shifts in DNA: assessment of the reliability of $^1$H chemical shift calculations for use in structure refinement. *J Biomol NMR* 10: 337–350. doi:10.1023/A:1018348123074

Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB, et al. 2018. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci* 27: 293–315. doi:10.1002/pro.3330

Ying J, Grishaev A, Bryce DL, Bax A. 2006. Chemical shift tensors of protonated base carbons in helical RNA and DNA from NMR relaxation and liquid crystal measurements. *J Am Chem Soc* 128: 11443–11454. doi:10.1021/ja061984g