Prediction of RNA H-1 and C-13 Chemical Shifts: A Structure Based Approach

ARTICLE *in* THE JOURNAL OF PHYSICAL CHEMISTRY B · SEPTEMBER 2013

Impact Factor: 3.3 · DOI: 10.1021/jp407254m · Source: PubMed

3 AUTHORS:

Aaron T Frank
University of Michigan
18 PUBLICATIONS **250** CITATIONS
SEE PROFILE

Sung-Hun Bae
Chong Kun Dang Pharmaceutical Corporation
22 PUBLICATIONS **426** CITATIONS
SEE PROFILE

Andrew Stelzer
Nymirum
14 PUBLICATIONS **398** CITATIONS
SEE PROFILE

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

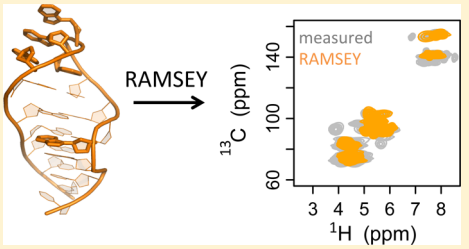
Prediction of RNA ^{1}H and ^{13}C Chemical Shifts: A Structure Based Approach

Aaron T. Frank,*,† Sung-Hun Bae, and Andrew C. Stelzer

Nymirum, 3510 West Liberty Road, Ann Arbor, Michigan 48103, United States

Supporting Information

ABSTRACT: The use of NMR-derived chemical shifts in protein structure determination and prediction has received much attention, and, as such, many methods have been developed to predict protein chemical shifts from three-dimensional (3D) coordinates. In contrast, little attention has been paid to predicting chemical shifts from RNA coordinates. Using the random forest machine learning approach, we developed RAMSEY, which is capable of predicting *both* ^{1}H and protonated ^{13}C chemical shifts from RNA coordinates. In this report, we introduce RAMSEY, assess its accuracy, and demonstrate the sensitivity of RAMSEY-predicted chemical shifts to RNA 3D structure.



INTRODUCTION

Over the past decade, the central dogma of molecular biology has undergone a significant revision; once thought of as passive translators of information from the genome to the proteome, ribonucleic acids (RNA) have emerged as important regulators of cellular activity.[1,2] Astonishingly, the recently concluded Encyclopedia of DNA Elements (ENCODE) project[3] estimates that as much as 80% of the human genome is transcribed into functional noncoding ribonucleic acids (ncRNAs), many of which are presumed to be actively involved in regulating cellular activity. Deciphering the structure−dynamics−function relationship of these RNAs is of immediate importance, as ncRNAs have been implicated in a number of diseases[4] including myotonic dystrophy type 1 (DM1), prostate cancer, spinal muscular atrophy (SMA), Huntington's disease-like 2 (HDL2), and autism. A recent study by Stelzer et al. demonstrates the importance of accurately characterizing the structural dynamics when carrying out a small molecule virtual screen against a target RNA. Using a combination of nuclear magnetic resonance (NMR) spectroscopy and computer simulation, they accurately determined the "dynamical" ensemble for the HIV-1 trans-activating response (TAR) RNA element, which is essential to HIV-1 replication. By performing a virtual screen against the *dynamical* ensemble, they were able identify a novel small molecule that was shown to inhibit HIV-1 replication in vivo.[5]

Characterizing an RNA's three-dimensional (3D) structure is a prerequisite for understanding its structural dynamics. Unfortunately, using traditional techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine RNA structure can be prohibitively time-consuming and expensive. Recently, much effort has been expended to develop methods to efficiently extract RNA structural information from readily accessible experimental observables. Examples of such readily accessible experimental observables include molecular envelopes, derived from small-angle X-ray scattering,[6,7] and contact mapping, derived from hydroxyl radical cleavage experiments.[8,9] Recent work suggests that NMR chemical shifts, the most accessible and accurately measured NMR observable, may also be capable of resolving RNA 3D structure.[10,11] Indeed, NMR chemical shifts have shown great promise as a reporter of protein secondary and tertiary structure, and are now routinely used to accurately predict,[12,13] validate,[14,15] and refine[16,17] protein models.

In order for chemical shifts to be of utility in resolving structure, efficient chemical shift prediction methods are needed. Nucleic acid chemical shifts can be predicted using *sequence*-based or *structure*-based approaches. *Sequence*-based approaches utilize secondary structure information to predict chemical shifts. One such approach that has shown great promise is RNAshifts,[18] which uses a database approach to predict ^{1}H chemical shifts within A-form helical regions of RNA. RNAshifts exhibited excellent agreement between measured and predicted chemical shifts, with a root-mean-square-error (RMSE) and Pearson correlation coefficient (R) of ∼0.06 ppm and ∼0.81, respectively.[18] *Structure*-based approaches on the other hand predict chemical shifts using 3D atomic coordinates. Among *structure*-based approaches, ab initio calculations, which utilize quantum mechanics (e.g., density functional theory, or DFT) to predict chemical shielding effects and thus chemical shifts, have shown tremendous promise in delineating the relationship between chemical shifts and nucleic acid structure,[19−23] as highlighted by recent studies by Wijmenga and co-workers[24] and Vendruscolo and co-workers.[25] One drawback of ab initio approaches is the significant computational cost of performing these calculations. *Empirical-structure*-based approaches, which utilize classical approximations to describe chemical shielding effects are

Received: July 22, 2013
Revised: September 11, 2013

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 1. Chemical Shift−Structure Database[a]

				sample conditions		
PDBID	BMRB	residues	% helical	T (K)	pH	description
1LDZ	4226	30	66.7	298	5.5	stem-loop (6-nt asymmetrical internal loop)
1LC6	5371	24	66.7	303	7.0	stem-loop
1NC0	5655	24	50.0	303	6.5	stem-loop (hyper-stable)
2KOC	5705	14	71.4	298	6.4	stem-loop (UUCG tetraloop)
1PJY	5834	22	81.8	303	6.8	stem-loop (ACAA tetraloop, A:A pair)
1OW9	5852	23	60.9	298	7.0	stem-loop (sheared G:A pairs)
1R7W	6076	34	70.5	298	6.5	stem-loop (6-nt asymmetrical internal loop)
1R7Z	6077	34	70.6	298	6.5	stem-loop (6-nt asymmetrical internal loop)
1YSV	6485	27	81.5	303	6.0	stem-loop (GCU(A/C)A pentaloop)
1Z2J	6543	45	84.4	303	6.8	stem-loop (ACAA tetraloop and 3 purine-bulge)
2GM0	7098	70	77.1	303	6.4	duplex (A rich 1−2 internal loops, A-bulges)
2JTP	15417	34	64.7	277	7.0	stem-loop (CCC triloop, G:A pair, A:A cross stacking)
2JXQ	15571	20	100.0	298	6.8	duplex
2JXS	15572	21	95.2	298	6.8	duplex (single A-bulge)
2K3Z	15780	17	94.1	298	6.8	duplex (single A-bulge, G:C-A triple)
2K41	15781	17	94.1	298	6.8	duplex (single U-bulge)
2L3E	17188	35	68.6	283	6.4	stem-loop (asymmetric 5-nt internal bulge loop)
2LDL	17671	27	66.7	303	5.5	stem-loop (GAUUAGU heptaloop, A:C wobble)
1UUU[b]	NA	19	63.2	293	6.0	stem-loop (GUUUC pentaloop)

[a]Listed are the PDBID and BMRB accession codes used to construct the chemical shift−structure database for training RAMSEY predictors. Also listed for each RNA are the number of residues, % helical bases, sample conditions (temperature and pH) under which ^{1}H and ^{13}C chemical shifts data were acquired, and a structural description. [b]Chemical shifts obtained from the literature.[51]

computationally less demanding and, thus, provide an attractive complement to ab initio calculations. Currently, however, there are few *empirical-structure*-based methods capable of predicting nucleic acid chemical shifts.[26,27] Additionally, there are currently no *empirical-structure*-based methods for predicting RNA ^{13}C chemical shifts, despite the wealth of structural information that ^{13}C chemical shifts provide.

Machine-learning techniques have been successfully employed to model the nonlinear relationships between chemical shifts and structure of proteins. The protein chemical shift predictor PROSHIFT[28] was generated using neural networks, as was the program TALOS-N,[29] which predicts backbone dihedral angles directly from chemical shifts. Similarly, boosting[30] and bagging,[31] two popular ensemble-based machine learning approaches, were combined to generate the accurate chemical shift predictor SHIFTX2.[32] Here we employ the random forest[33] approach to generate RAMSEY, a set of empirical predictors that efficiently and accurately compute *both* ^{1}H and protonated ^{13}C chemical shifts from RNA 3D coordinates (http://nymirum.com/RAMSEYWebService/). Random forest is ideal for establishing chemical shift−structure relationship because it is well suited to handle multidimensional data sets and modeling nonlinear relationships. In addition, the random forest approach is resilient against overfitting, and its algorithmic simplicity makes it computationally inexpensive. In this report, we assess the accuracy of the predictors and demonstrate the sensitivity of RAMSEY-predicted chemical shifts to structure using the sarcin-ricin loop (SRL) RNA[34,35] as a model system. We show (i) that RAMSEY predicts ^{1}H and ^{13}C chemical shifts with good accuracy, and (ii) that the predicted chemical shifts exhibit significant sensitivity to 3D structure, affording RNA structural information to be extracted directly from NMR chemical shifts.

METHODS AND MATERIALS

RAMSEY Chemical Shift Predictors. We compiled a chemical shift−structure database consisting of 19 RNAs for which both experimental ^{1}H and protonated ^{13}C chemical shifts were available and NMR structure(s) deposited in the PDB (www.pdb.org) (Table 1). Care was taken to ensure that we only included data sets for which the ^{13}C chemical shifts were verified as being correctly and consistently referenced.[36] After examining the corresponding ^{1}H chemical shifts in these data sets, we verified that they were also correctly referenced. The training database contained 3150 ^{1}H chemical shifts and 2270 ^{13}C chemical shifts.

In the chemical shift-structure database, each chemical shift entry was mapped to local 3D descriptors calculated from PDB coordinates. Prior to calculating the 3D descriptors, the average structure was calculated for each NMR ensemble. The ensemble model closest to the average structure was selected and then energy minimized using the AMBERff99χ_{OL} force field[37] and the generalized Born surface area (GBSA) implicit solvent. 3D descriptors were then calculated from the energy-minimized structure. 3D descriptors include: the type of base (ADE, GUA, CYT, URA); type of nucleus (H1′, H2′, H3′, H4′, H5′, H5″, H2, H5, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6, C8); ring-current effect; local magnetic anisotropy effect; polarization effect by an electric field of the electron density along the chemical bond(s); close-contact; number of base−base, base−backbone, and backbone−backbone hydrogen bonding bonds; total stacking interactions; and complete set of dihedral angles (α, β, γ, δ, ε, ζ, χ, ν, ν_1, ν_2, ν_3, ν_4). Ring-current, local magnetic anisotropy, and polarization effects were calculated as described previously for the NUCHEMICS method.[27] Close-contact (CC) [$CC = \Sigma_{i,j} \exp(-(2/5)r_{ij})$] was calculated for all heavy atoms i and j which do not belong to a same residue and their distance (r_{ij}) is less than 5 Å. Hydrogen bonding (D−H···A) was counted if the distance

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

between donor and acceptor (r_{DA}) is less than 3.5 Å and their angle ($\theta_{D\text{-}H\text{-}A}$) is less than 45°. Atoms O3P, P, O1P, O2P, O5′, C5′, C4′, C3′, and O3′ were considered as backbone atoms. Stacking was counted if the centers of two rings are within 5.0 Å and the angle between two norms of the ring planes is less than 30°. We did not take into account experimental conditions such as the temperature or buffer under which NMR spectra were collected.

Random forest[33] is a robust and efficient approach in which ensembles of independent random decision trees are generated and used to predict new data by aggregating the predictions of the individual trees. Each decision tree is generated using a random subset of the training database, and a random subset of descriptors are selected and used to determine the splits at each node in a given tree. By applying the random forest approach to our chemical shift-structure database, we generated a set of independent H1′, H2′, H3′, H4′, H5′, H5″, H2, H5, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6 and C8 chemical shift predictors that we collectively refer to as RAMSEY. The RAMSEY predictors were generated using the random forest library implemented in the randomForest[38] R package (http://www.r-project.org). The tree size (i.e., number of trees in a given predictor) was set to 1000, and the sample size (fraction of the training database randomly select to grow each tree) was set to 1/3. Default values were utilized for all other parameters. Because each decision tree in the forest is grown using a randomly selected subset of the training database, the accuracy of a given tree can be determined by assessing the ability of that tree to predict the subset of target data not used to grow that tree (i.e., via cross-validation). For each tree, the accuracy can be determined by calculating the root mean squared error (RMSE) and Pearson correlation coefficient (R) between measured chemical shifts and predicted chemical shifts. Using this cross-validation approach, the accuracy of each RAMSEY chemical shift predictor was determined by averaging the RMSE and R over the set of 1000 trees in each predictor forest.

To compare the accuracy of RAMSEY's ^{1}H predictors to that of SHIFTS and NUCHEMICS, we first used SHIFTS and NUCHEMICS to predict ^{1}H chemical shifts for the nuclei corresponding to those in our chemical shift-structure database. The predictions utilized the same energy-minimized structures used to build the chemical shift-structure database. For a given proton nucleus type (e.g., H1′, H2′, H3′, H4′, H5′, H5″, H2, H5, H6, or H8), one-third of the available measured chemical shifts data (the same as the RAMSEY sample size) were randomly selected and the RMSE and R relative to the SHIFTS- and NUCHEMICS-predicted chemical shifts were computed. This procedure was repeated 1000 times (the same as the number of trees in each RAMSEY predictor) and the RMSE and R for that nucleus were then averaged. The SHIFT and NUCHEMICS RMSE and R estimated using this procedure were then directly compared to those of the RAMSEY ^{1}H predictors.

Sensitivity of RAMSEY-Predicted Chemical Shifts to RNA 3D Structure. To test the sensitivity of RAMSEY-predicted ^{1}H and protonated ^{13}C chemical shifts to RNA 3D structure, we determined the chemical shift−structure relationship for a benchmark system. We selected the sarcin-ricin loop (SRL) RNA[34,35] as a model because: (i) both ^{1}H and ^{13}C chemical shifts were available; (ii) SRL RNA contains complex structural features (such as base-triples, nonbase-paired apical loop and noncanonical base-pairs) that deviate from the typical A-form RNA helices; (iii) the rigid structure simplifies interpretation of the chemical shifts, and NMR chemical shifts can be modeled from a single static structure rather than from an ensemble of conformational substates; and (iv) the structure of the SRL RNA has been extensively studied and solved by both NMR and X-ray crystallography. We examined the relationship between the chemical shift prediction error (i.e., error between measured and RAMSEY-predicted chemical shifts) and the structural agreement between the experimentally determined structures and models taken from a diverse conformational pool.

The diverse conformational pool for the SRL was generated using the MC-Sym RNA structure prediction program, which generates native-like models of a target RNA directly from sequence. The MC-Sym models were generated as follows: First, secondary structures were predicted from the SRL sequence using MC-Fold.[39] A total of 8038 MC-Sym 3D models were generated for the top nine secondary structures using the MC-Sym web server.[39] Finally, using the AMBERff99 χ_{OL} force field paired with the generalized born surface area (GBSA), the MC-Sym models were energy-minimized to relieve steric clashes. Following MC-Sym model generation, RAMSEY was used to back-predict chemical shifts from the individual models. The agreement between measured and predicted chemical shifts from a given model with 3D coordinates $\boldsymbol{r}$ was calculated by a weighted mean absolute error wMAE($\boldsymbol{r}$) defined as

$$\mathrm{wMAE}(\boldsymbol{r}) = \sum_{n=1}^{N}\sum_{m=1}^{M} w_n \left|\delta_{m,\mathrm{pred}}(\boldsymbol{r}) - \delta_{m,\mathrm{meas}}\right| \tag{1}$$

Here, n is the index that runs over the individual RAMSEY chemical shift predictor types (H1′, H2′, H3′, H4′, H5′, H5″, H2, H5, H6, H8, C1′, C2′, C3′, C4′, C5′, C2, C5, C6, and C8); m is the subset of chemical shifts for each type; $\delta_{m,\mathrm{pred}}(\boldsymbol{r})$ and $\delta_{m,\mathrm{meas}}$ are the predicted and measured chemical shifts, respectively; and w_n is a weight factor that scales prediction error for nucleus type n. w_n is defined as

$$w_n = \frac{R_n^{\,2}}{\mathrm{RMSE}_n} \tag{2}$$

where $R_n^{\,2}$ is the squared Pearson correlation coefficient and RMSE_n is the estimated root-mean-squared error of the nth RAMSEY predictor. The w_n makes wMAE($\boldsymbol{r}$) comparable between RAMSEY predictors that have different prediction error ranges depending on nucleus and atom types, and weighs contributions to wMAE($\boldsymbol{r}$) based on accuracy of RAMSEY predictors (i.e., chemical shifts predicted from the more accurate predictors contribute more to wMAE($\boldsymbol{r}$)). For SRL, the wMAE($\boldsymbol{r}$) was calculated relative to the ^{1}H and ^{13}C chemical shifts over residues from 7 to 23. The ^{13}C chemical shifts for SRL were verified (see Results and Discussion) as containing a systematic referencing error of ∼2.6 ppm.[36] Therefore, prior to computing wMAE($\boldsymbol{r}$), the ^{13}C chemical shifts were rereferenced.

The structural agreement between models in the conformational pool and the known structure were quantified using the heavy atom root-mean-square-difference (RMSD). Specifically, the models were compared to the solved NMR (PDBID 1SCL)[34] and X-ray (PDBID 430D)[35] structures. In plots showing a correlation between wMAE($\boldsymbol{r}$) and RMSD, data were binned along the RMSD using bin widths of 0.5 Å within which mean wMAE($\boldsymbol{r}$) and the standard deviation was computed.

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 2. Accuracy of RAMSEY ^{1}H and ^{13}C Predictors[a]

nucleus	#shifts	database range (ppm)	RMSE (ppm)	(RMSE/range) %	R
H1′/C1′	455/381	2.43/8.44	0.19/0.89	7.8/10.5	0.77/0.77
H2′/C2′	456/315	1.13/6.02	0.13/0.78	11.5/13.0	0.68/0.34
H3′/C3′	408/288	1.70/7.80	0.14/0.93	8.2/11.9	0.64/0.76
H4′/C4′	390/287	1.68/10.90	0.09/0.79	5.4/7.2	0.70/0.76
H5′/C5′	347/236	1.00/9.15	0.19/0.98	19.0/10.7	0.48/0.70
H5″	316	1.91	0.19	9.9	0.56
H2/C2	104/107	2.26/9.40	0.21/1.27	9.3/13.5	0.86/0.40
H5/C5	217/209	1.20/10.66	0.15/0.73	12.5/6.8	0.85/0.97
H6/C6	217/209	0.93/6.69	0.13/0.75	14.0/11.2	0.76/0.74
H8/C8	240/238	1.62/9.03	0.20/0.97	12.3/10.7	0.83/0.88
mean		1.59/8.68	0.16/0.90	11.0/10.6	0.71/0.69

[a]Listed are the number of chemical shifts in the training database, width of the chemical shift distribution in the training database, the cross-validation root-mean-square-error (RMSE), the ratio of the RMSE to the width of the chemical shift distribution, and the Pearson correlation coefficient (R) for each ^{1}H and ^{13}C RAMSEY predictor.

Finally, we performed receiver-operator-characteristic (ROC) analysis to compare the sensitivity of RAMSEY-predicted chemical shifts to 3D structure when using ^{1}H chemical shifts only, ^{13}C chemical shifts only, or both ^{1}H and ^{13}C chemical shifts. ROC curves report on the ability of a predictor to classify a set of objects by comparing sensitivity (the proportion of objects correctly classified as positive) and specificity (the proportion of objects correctly classified negative). To quantify the degree to which a predictor successfully classifies a set of objects, the area-under-the-curve (AUC) can be computed. In the application of ROC curve analysis to the sensitivity of RAMSEY-predicted chemical shifts to 3D structure, the predictor variable is the wMAE($\boldsymbol{r}$) and the objects are the MC-Sym models that are classified positively ("native") or negatively ("non-native") when the RMSD relative to the solved structure is less than or larger than 2.0 Å, respectively. An AUC of 1 indicates perfect classification. An AUC of 0.5 suggests that the classification using a given predictor is no better than random classification. Typically, AUCs of greater than ∼0.85 are indicative of a good predictor; that is, the predictors are able to accurately partition positive and negative classes. The ROC curves and the area under the curve (AUC) were calculated using wMAE($\boldsymbol{r}$) calculated from ^{1}H chemical shifts only, ^{13}C chemical shifts only, or both ^{1}H and ^{13}C chemical shifts.

RESULTS AND DISCUSSION

While widely applied in the resolution of protein structures, the use of chemical shifts in resolving RNA structure has only recently begun to receive attention. There is a dearth of available methods capable of accurately and efficiently predicting chemical shifts based on 3D coordinates.[40] The empirical methods currently available can only predict ^{1}H chemical shifts with an accuracy of ∼0.3 ppm, which is large relative to the chemical shift range for the typical RNA protons (∼1.7 ppm).[10] Importantly, there is a conspicuous lack of empirical methods capable of predicting ^{13}C chemical shifts, which represent an additional source of potentially rich and complementary structural information. We therefore set out to generate chemical shift predictors for *both* ^{1}H and ^{13}C chemical shifts.

Accuracy of RAMSEY ^{1}H Chemical Shift Predictors. For proton nuclei, RAMSEY predicted chemical shifts with an RMSE that ranged between 0.09 and 0.21 ppm (mean = 0.16 ppm), and an R that ranged between 0.48 and 0.86 (mean = 0.71) (Table 2). The largest discrepancies were observed for H5′ and H5″ nuclei, which exhibited R values of 0.48 and 0.56, respectively. Typically, H5′ and H5″ are ambiguously assigned and reported due to difficulties in stereospecific assignment, and thus the training sets for H5′ and H5″ experimental chemical shifts are not completely reliable. When neglecting these nuclei, the mean R increased to 0.76 (Table 2). A head-to-head comparison of RAMSEY with SHIFTS and NUCHEMICS (the only other currently available empirical methods capable of predicting RNA ^{1}H chemical shifts from 3D coordinates) revealed RAMSEY to be more accurate. Not only did RAMSEY exhibit a lower RMSE (RAMSEY mean RMSE was 0.16 ppm, compared to 0.41 ppm and 0.26 ppm for SHIFTS and NUCHEMICS, respectively), it also exhibited higher mean R values (R = 0.71 for RAMSEY compared to 0.45 for SHIFTS and 0.56 for NUCHEMICS).

An alternative to the *structure*-based approach for predicting chemical shifts is the *sequence*-based approach, which utilizes secondary structure information rather than atomic coordinates to predict chemical shifts. RNAshifts,[18] which was designed to predict chemical shifts within A-form helical regions is an example of one such *sequence*-based approach. In order to compare RAMSEY with the *sequence*-based approach employed by RNAshifts, we generated a new set of predictors using a small training data set (604 ^{1}H chemical shift entries) containing only chemical shifts within A-form helical regions (PDBID 2GM0, 2JXQ, 2JXS, 2K3Z, and 2K41; Table 1). For the new set of predictors, we observed modest improvements in the accuracy of ^{1}H chemical shift predictions compared to predictors generated using chemical shifts from both helical and nonhelical regions. Specifically, the mean cross-validation RMSE decreased from 0.16 to 0.13 and the mean R increased from 0.71 and 0.73 (Table S1, Supporting Information). By comparison, the RMSE and R reported for RNAshifts was ∼0.06 ppm and 0.81, respectively, indicating that when applied exclusively to A-form helical regions, the *structure*-based approach did not attain the same level of accuracy observed for RNAshifts. It is conceivable that as the chemical shift-structure database expands, the accuracy of the *structure*-based predictions for chemical shifts within A-form helical region will converge to those observed for *sequence*-based approaches like RNAshifts.

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

Accuracy of RAMSEY ^{13}C Chemical Shift Predictors. For carbon nuclei, RAMSEY predicted chemical shifts with an RMSE that ranged between 0.66 and 1.27 ppm (mean = 0.90 ppm) and R that ranged between 0.37 and 0.98 (mean = 0.69) (Table 2). As was the case of the ^{1}H proton, there were a few ^{13}C nuclei for which the predicted chemical shifts exhibited significantly poor agreement with measured chemical shifts. In particular, C2′ and C2 nuclei had R values of 0.27 and 0.40, respectively. When neglecting these nuclei the mean R increased to 0.78. One possible reason for the poor performance for C2′ and C2 nuclei is that the simple 3D descriptors used to train the predictors failed to capture local structure features that are most important for these nuclei. It is also possible that the number of chemical shifts used to train the predictors for these nuclei were insufficient to establish the complex shift-structure relationships that govern their respective chemical shift dispersions. In the case of proteins, ^{13}C chemical shifts predictors have been trained with a chemical shift database that contains ∼65 000 chemical shift entries.[28] Expansion of the chemical shift-structure database would likely improve the accuracy of these predictors.

To compare the accuracy of ^{1}H and ^{13}C chemical shift predictors, we determined the ratio of the RMSE to the chemical shift range spanned in our training database. We found that the RMSE of both ^{1}H and ^{13}C predictors are approximately 11% of the width of chemical shifts distribution (Table 2). This ratio provides an alternative measure of the accuracy of the individual predictors and suggests that that RAMSEY is able to predict chemical shifts within ∼11% of the expected chemical shift range for a given nucleus. The results indicate that the accuracy of the ^{1}H and ^{13}C predictors were comparable.

Importance of Individual 3D Descriptors. A particularly important aspect of generating predictive models is assessing the importance of the variables used to characterize the feature space, as they can lead to insights into the fundamental rules governing the phenomenon of interest. Previous attempts have been made to interrogate the relationship between chemical shifts and structure using ab initio techniques at different levels of theory.[19−25] These studies revealed that nucleic acid chemical shifts are sensitive to both the local structure within nucleosides (e.g., dihedral angles), as well as inter-residue interactions (e.g., base pairing). Along similar lines, we attempted to assess the importance of the 3D descriptors (variables) used to train RAMSEY. We calculated the "*importance score*" of a given descriptor as the mean increase in error that is observed when that variable was randomly permuted throughout the training data set. The expectation is that important descriptors, when permuted, should demonstrate a large increase error and vice versa. We utilized this approach to compare the importance of the set of 3D descriptors used to train RAMSEY. For each ^{1}H and ^{13}C chemical shift predictor, we determined the *importance score* for each 3D descriptor. The *importance score* was then normalized so that the scores ranged between 0 and 1, with 0 indicating that the feature had no importance, and 1 indicating that the feature was the most important in the set. To compare the collective influence of the 3D descriptors on the ^{1}H and ^{13}C chemical shifts predictors, the average *importance scores* were determined separately for ^{1}H and ^{13}C predictors.

For ^{1}H predictors, we found that the most important 3D descriptors were ring current effects, magnetic anisotropy, residue type, and base−base hydrogen bonding, which exhibited relative *importance scores* of 1.00, 0.48, 0.35, and 0.24, respectively (Table 3). Interestingly, stacking interactions exhibited low relative importance (0.02). This was most likely due to the fact that stacking interactions were highly correlated with ring current effects, and so represented a redundant descriptor. This correlation, coupled with the fact that stacking interactions were described using a discrete count variable (rather than a continuous variable, as in the case of the ring current effects), most likely rendered the stacking descriptor uninformative relative ring current effects when deciding splits in the decision trees. In the case of ^{13}C, the most important descriptors were residue type, base−base hydrogen bonding, the ribose dihedral ν_3, the backbone dihedral ε, and ring current effects, which had relative *importance scores* of 1.00, 0.84, 0.81, 0.67, and 0.67, respectively (Table 3). Interestingly, the relative *importance scores* for dihedral angles were much larger for carbons than protons. Taken together, these results suggest that while both ^{1}H and ^{13}C chemical shifts are sensitive to inter-residue structures as captured base−base interactions and ring current effects, ^{13}C chemical shifts seem to have an additional dependence on local intraresidue structure, as captured by dihedrals angles.

Table 3. Relative *Importance Scores* of Individual 3D Descriptors[a]

variables	relative importance ^{1}H	relative importance ^{13}C
residue type (A, G, C, U)	0.35	1.00
close contact	0.16	0.66
ring current	1.00	0.67
bond polarization	0.12	0.15
magnetic anisotropy	0.48	0.44
stacking	0.02	0.01
base−base hydrogen bond	0.24	0.84
base-backbone hydrogen bond	0.01	0.11
backbone−backbone hydrogen bond	0.01	0.02
α	0.05	0.18
β	0.05	0.18
γ	0.03	0.14
δ	0.13	0.27
ε	0.13	0.67
ζ	0.12	0.31
χ	0.07	0.39
ν_0	0.08	0.35
ν_1	0.17	0.59
ν_2	0.18	0.64
ν_3	0.15	0.81
ν_4	0.08	0.38

[a]Listed are the mean relative *importance scores* for RAMSEY's ^{1}H and ^{13}C chemical shifts predictors. The scores are the normalized mean difference between the cross-validation error obtained for a given tree and that obtained when a given variable is permuted prior to generating the decision tree. The score ranges between 0 and 1, and the feature with the largest percentage increase in error upon permutation has a score of 1.

Application of RAMSEY. We next examined how well RAMSEY predicted chemical shifts for an independent set of RNAs not used in its training. We first compiled a list of RNAs whose NMR structures and ^{1}H and/or ^{13}C chemical shifts were available. We excluded RNAs bound to proteins and RNAs containing mutated residues. The final list contained 56 RNAs

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

Table 4. Application of RAMSEY[a]

nucleus	RMSE (ppm)	MAE (ppm)	R
H1′/C1′	0.23/3.30/1.11/1.11	0.23/1.79/0.74/0.74	0.64/0.23/0.66/0.66
H2′/C2′	0.20/2.88/1.44/0.94	0.20/1.70/0.62/0.57	0.55/0.12/0.14/0.22
H3′/C3′	0.22/2.67/1.33/1.40	0.22/1.82/1.08/1.07	0.46/0.53/0.62/0.65
H4′/C4′	0.16/2.68/1.15/1.05	0.16/1.77/0.72/0.71	0.50/0.41/0.55/0.59
H5′/C5′	0.27/3.81/1.79/1.26	0.27/2.15/0.97/0.93	0.42/0.30/0.41/0.52
H5″	0.20/NA/NA/NA	0.20/NA/NA/NA	0.43/NA/NA/NA
H2/C2	0.35/10.68/1.41/1.40	0.35/4.18/0.94/0.94	0.64/−0.11/0.34/0.33
H5/C5	0.20/6.03/3.48/1.15	0.20/2.58/1.15/0.85	0.71/0.30/0.63/0.95
H6/C6	0.18/8.20/2.26/1.07	0.18/2.91/0.86/0.77	0.58/0.13/0.34/0.60
H8/C8	0.24/8.73/3.10/1.24	0.24/3.58/1.03/0.86	0.77/0.15/0.47/0.83
mean	0.22/5.44/1.91/1.18	0.22/2.50/0.91/0.83	0.57/0.22/0.47/0.60

[a]Listed are the root-mean-square-error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (R) for the application of RAMSEY to the 55 RNAs listed in Table S1. The four elements in each cell correspond to results for ^{1}H chemical shifts, ^{13}C chemical shifts, re-referenced ^{13}C chemical shifts, and re-referenced ^{13}C chemical shifts with prediction outliers removed, respectively.

(Table S2). For each RNA, the representative model selected from the NMR ensemble was energy minimized (see Methods and Materials) and used as an input to RAMSEY. Comparisons were then made between measured and RAMSEY-predicted chemical shifts. We found that for ^{1}H the RMSE and R values obtained over the independent test set were comparable to those obtained via cross-validation from the training set. The ^{1}H RMSE ranged between 0.16 and 0.35 ppm (mean = 0.22 ppm), and the R values ranged between 0.43 and 0.77 (mean = 0.57) (Table 4).

In contrast, for ^{13}C, the RMSE and R values obtained over the independent test set were significantly different from the training set. In this case, the RMSE ranged between 2.24 and 8.86 ppm (mean = 4.58 ppm), and the R values ranged between −0.11 and 0.53 (mean = 0.23) (Table 4). Because ^{13}C chemical shifts frequently contain inconsistencies[36] (e.g., systematic referencing errors), we carried out a closer examination of the RAMSEY ^{13}C prediction errors and their distributions. Of the 56 RNAs, we were able to identify 14 (PDBID 1SCL, 2QH2, 2QH3, 2QH4, 1ZC5, 1XHP, 2M21, 1Z30, 2M8K, 2JYM, 1S9S, 2LQZ, 2LC8, and 1N8X) whose absolute median errors (AME) were significantly shifted from 0.00 ppm (Table 5). Interestingly, of the 14, 6 (PDBID 1SCL, 2QH2, 2QH3, 2QH4, 1ZC5, and 1XHP) were previously identified as containing systematic referencing errors, and the AME of 1SCL (2.78 ppm), 2QH2 (2.45 ppm), 2QH3 (2.58 ppm), and 2QH4 (2.61 ppm) were in excellent agreement with the ∼2.7 ppm estimated by Aeschbacher et al. We compared the chemical shifts of the terminal G:C base pair to a set of reference values (following a similar approach used by Aeschbacher et al. to validate ^{13}C data sets), and discovered that chemical shift data sets for PDBID 2M21, 1Z30, 2M8K, 2JYM, and 1S9S also appeared to contain systematic referencing errors. In these cases, the AME were 2.78, 1.49, 1.52, 1.61, and 1.98 ppm for 2M21, 1Z30, 2M8K, 2JYM, and 1S9S, respectively (Table 5). After rereferencing the ^{13}C chemical shifts of the 11 RNAs that contained systematic referencing errors, the mean RMSE for the 11 RNAs were reduced from 2.42 to 1.16 ppm, indicating that, in part, the initial poor agreement observed between the measured and predicted ^{13}C chemical shifts was due to referencing errors (Table 5).

Table 5. RNA Exhibiting Large ^{13}C AME[a]

PDBID	AME (ppm)	RMSE (ppm)	MAE (ppm)
1SCL	2.78	3.07/1.29	2.86/0.79
2QH2	2.45	2.65/0.96	2.47/0.66
2QH3	2.58	2.63/1.09	2.51/0.56
2QH4	2.61	2.80/1.33	2.54/0.95
1ZC5	2.12	2.25/1.00	2.04/0.71
1XHP	2.09	2.16/0.79	2.04/0.55
2M21	2.78	2.80/0.71	2.71/0.53
1Z30	1.49	1.51/0.77	1.43/0.50
2M8K	1.52	2.13/1.69	1.73/1.38
2JYM	1.61	1.67/0.83	1.54/0.57
1S9S	1.98	2.95/2.28	2.04/0.73
2LQZ	34.80	33.06/27.40	32.78/13.29
2LC8	21.90	25.71/24.05	23.12/16.80
1N8X	9.65	9.02/4.11	8.16/3.08

[a]Listed are RNAs that exhibited absolute median error (AME) of prediction > 1.0 ppm. Also listed are the RMSE and MAE calculated before and after re-referencing by the median error. RNAs are identified by their PDBID.

While the large RMSE obtained when calculating chemical shifts for 11 of the 14 structures could be attributed to referencing errors, the remaining three RNAs 2LQZ, 2LC8, and 1N8X had AME of 34.80, 21.90, and 9.65 ppm, respectively, even after rereferencing, and continued to give rise to RMSEs of 27.40, 24.05, and 4.11 ppm, respectively (Table 5). These results suggest that the errors were not a result of systematic errors in the original BMRB chemical shifts data. In the case of 1N8X, Aeschbacher et al. identified the chemical shift data set as containing unknown and inconsistent errors, which is consistent with what we have found here. Though we cannot rule out the fact that these errors are due to RAMSEY, these results warrant a closer examination of the chemical shifts deposited in the BMRB.

Excluding 2LQZ, 2LC8, and 1N8X, the average RMSE for individual ^{13}C nuclei ranged between 1.11 and 3.48 ppm (mean = 1.91 ppm) after rereferencing (Table 4). By comparison, the mean MAE, which is less sensitive to outliers, was 0.91 ppm. The large difference between the mean RMSE and the mean MAE is an indication that the RMSE may be dominated by a small number of large prediction outliers. To confirm this, we calculated the 99.9% confidence interval of the prediction errors, and identified prediction that fell outside the interval as outliers. Using this stringent criterion, 11 out of the 5988 predicted chemical shifts were identified as outliers (Table S4). After removing these outliers, the mean RMSE for individual

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

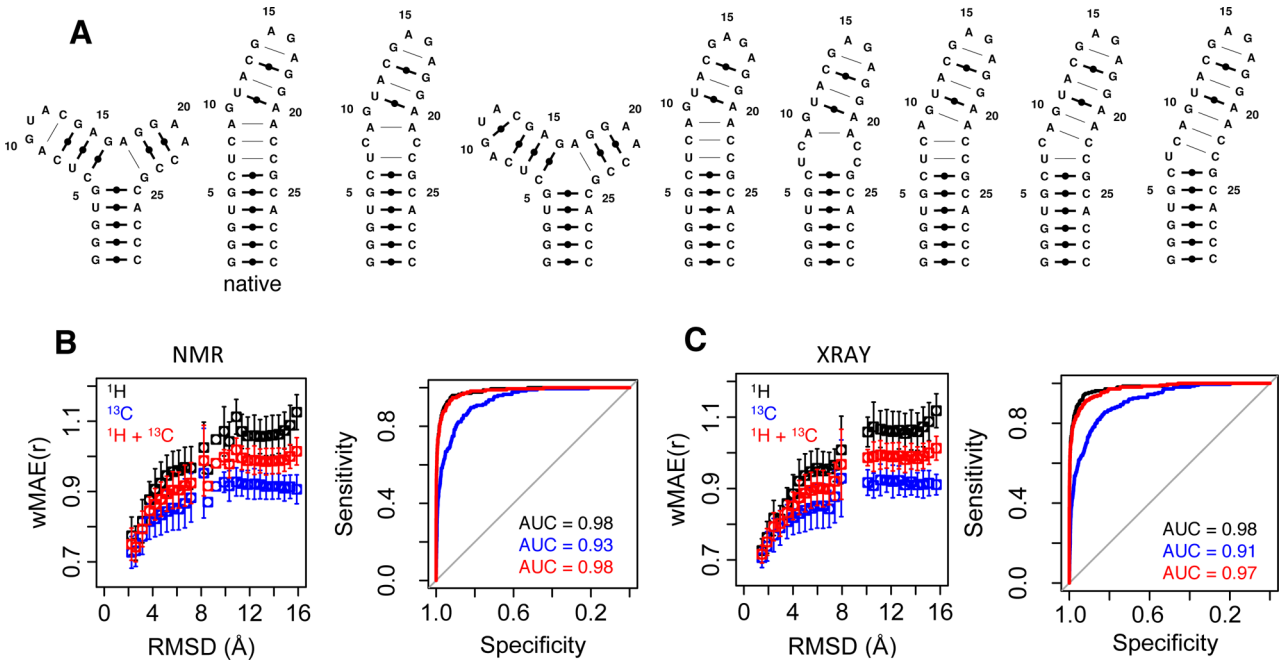


Figure 1. Sensitivity of RAMSEY-predicted chemical shifts. (A) A structurally diverse conformational pool was constructed for the SRL RNA by generating MC-Sym models for the top 9 MC-Fold predicted secondary structures. (B,C) Correlation plots of wMAE(r) and RMSD relative to NMR (B, left) and solved X-ray (C, left) structures, respectively. ROC curves using wMAE(r) as the predictor to classify MC-Sym models as *native* and *non-native* relative to the NMR (B, right) and X-ray (C, right) structure, respectively, are also shown. For both correlation and ROC curve plots, results are shown when wMAE(r) was calculated using ^{1}H chemical shifts (black), ^{13}C chemical shifts (blue), or both ^{1}H and ^{13}C chemical shifts (red).

^{13}C nuclei was reduced from 1.91 to 1.18 ppm (Table 4), comparable to the accuracy estimated from cross-validation of the training database (~0.90 ppm).

Given that the training database used to generate RAMSEY contained mostly stem-loop RNAs (Table 1), we were interested in examining how well RAMSEY predicted chemical shifts for RNAs with more complicated structural features such as kissing interactions, pseudoknots, and quadruplexes. While ^{1}H chemical shifts are predicted with similar accuracy in both set of RNAs, RAMSEY ^{13}C predictions for more structurally complex RNAs exhibited larger error. The mean ^{1}H and ^{13}C RMSE for RNAs containing stem-loops were ~0.20 and 0.98 ppm (Table S3), respectively, as compared to ~0.25 and 1.44 ppm (Table S3), respectively, for RNAs with kissing interactions, pseudoknots, and quadruplexes (PDBID 2ADT, 2LU0, 1A60, 1YMO, 2M8K, 2LC8, 2L1V, 2M18, 2L1F, and 2RN1). One of the outstanding differences between stem-loop RNAs and the more complicated RNAs like pseudoknots and quadruplexes is the unique backbone structure found in the latter. Indeed, ^{13}C chemical shifts are thought to be more sensitive to backbone geometry, and in contrast to ^{1}H, reasonably accurate ^{13}C chemical shift predictors can be generated using only backbone dihedral angles (data not shown). Thus, ^{13}C predictors trained with stem-loop RNAs could fail in more complicated RNA structures. Expansion of the chemical shift−structure database to include such RNAs would likely improve the accuracy of ^{13}C RAMSEY predictors.

We also determined the applicability of RAMSEY to ligand-bound RNAs. Surprisingly, for the three ligand-bound RNAs in our list (PDBID 2LWK, 2L94, and 2M4Q), the mean ^{1}H and ^{13}C RMSE were ~0.22 and 0.93 ppm (Table S3), indicating that chemical shifts for these RNA can be reliably predicted with only RNA structure taken into account. We do not expect this to be universal, as some ligands, depending on binding pose, will significantly perturb the electronic environment of nuclei near the binding site, and so must be explicitly accounted for in any chemical shift prediction method. Nonetheless, these results suggest that, at least in part, the chemical shift dispersion associated with ligand binding is due to change in RNA structure, and not simply the presence of the ligand near NMR active nuclei.

Taken as a whole, these results demonstrate that RAMSEY was able to accurately predict both ^{1}H and ^{13}C chemical shifts in the independent test set, especially for RNAs containing stem-loop architecture. In the case of ^{13}C, however, careful analysis was needed as some of the data sets contained significant errors. This latter point underscores the importance of carefully analyzing predictions errors when assessing the accuracy of chemical shift predictors, as summary statistics like RMSE can be skewed by the presence of a few prediction outliers. The results also highlight the ability of RAMSEY to detect errors in ^{13}C chemical shifts data sets and to provide reliable estimates for the magnitude of these errors. This ability will find utility when adding new chemical shifts data and structures to the chemical shift-structure database where maintaining consistency within the database will be essential.

Sensitivity of RAMSEY-Predicted Chemical Shifts to RNA 3D Structure. We next examined the sensitivity of RAMSEY-predicted ^{1}H and protonated ^{13}C chemical shifts to RNA 3D structure. It has long been recognized that NMR-derived chemical shifts are sensitive to local RNA structure.[19−23] In the case of ^{1}H chemical shifts, recent evidence has highlighted their ability to resolve RNA 3D structure.[10,11] For a model RNA system, we therefore generated an extensive and diverse conformational pool and then determined the relationship between the structural agreement of models in the pool to

dx.doi.org/10.1021/jp407254m | J. Phys. Chem. B XXXX, XXX, XXX−XXX

the solved structures and the error between measured chemical shifts and the chemical shifts predicted from those models using RAMSEY. Ideally, such an analysis should yield a strong positive correlation between structural agreement and error, with the models that most resemble the native structure having the lowest prediction errors, and those that deviate significantly from the native structure having higher prediction errors.

We carried out this analysis on the sarcin-ricin loop (SRL) RNA, which contains a highly structured loop region (residues 7−23) that is stabilized by series of noncanonical base pairs and capped by a GAGA tetraloop. Additionally, SRL contains a bulged-out guanosine (G10) and U11-A20-G10 base triple.[34,35] We began by generating a conformational pool for the SRL RNA using MC-Sym (see Methods and Materials). Briefly, MC-Sym models were generated for the top nine secondary structures predicted by MC-Fold based on the SRL sequence. The heavy atom RMSD was then calculated between the energy-minimized models and the solved NMR (PDBID 1SCL) and X-ray (PDBID 430D) structures. Next, we determined the agreement between the measured chemical shifts and the chemical shifts predicted from the MC-Sym models. Here, the agreement between measured and predicted shifts was quantified using the wMAE(r) (eq 1).

We observed strong positive correlation between wMAE(r) and the RMSD calculated relative to NMR (Figure 1B) and X-ray (Figure 1C) structures, respectively. The models that most closely resemble the native structures have the lowest wMAE(r) (Figure 2), and those that deviated significantly from the native structures demonstrate higher wMAE(r). The results were similar regardless of whether wMAE(r) was calculated using ^{1}H chemical shifts only, ^{13}C chemical shifts only, or both ^{1}H and ^{13}C chemical shifts. The ^{1}H chemical shifts, however, appear to exhibit slightly greater sensitivity to RMSD than ^{13}C chemical shifts. For lower RMSD values, the slope of the ^{1}H wMAE(r) is slightly higher than that of the ^{13}C wMAE(r) (Figure 1B and C). This was confirmed by ROC analysis, in which the AUCs obtained using ^{1}H wMAE(r) as the predictor were slightly higher than those obtained using ^{13}C wMAE(r) as the predictor. When comparing the MC-Sym models to the solved NMR structures, the ^{1}H and ^{13}C AUCs were 0.98 and 0.93, respectively; when comparing the MC-Sym models to the solved X-ray structures, the ^{1}H and ^{13}C AUCs were 0.98 and 0.91, respectively (Figure 1B and C). These results indicate that ^{1}H chemical shifts give rise to slightly more accurate predictions than ^{13}C chemical shifts when using RAMSEY to determine *native* and *non-native* structures. Similar results were obtained when combining ^{1}H and ^{13}C chemical shifts (AUCs = 0.98 and 0.97 for the NMR and X-ray, respectively). The fact that RAMSEY-predicted ^{1}H chemical shifts appear slightly more sensitive to RNA structure than protonated ^{13}C may be because for carbon nuclei, the 3D descriptors extracted from the solved NMR structure used to train the predictors failed to capture certain aspects of the local electronic environment that are relevant to chemical shift dispersion. Another possibility is that the observed differential in sensitivity is reflective of the actual physical reality: ^{1}H chemical shifts are more sensitive to local RNA structure that distinguish native-like models (e.g., base pairing and stacking) than protonated ^{13}C chemical shifts. In the future, it will be interesting to determine the sensitivity of RAMSEY-predicted chemical shifts to RNA 3D structure for nuclei such as ^{15}N, which are sensitive to base pairing interactions and ^{31}P, which report directly on backbone conformation.

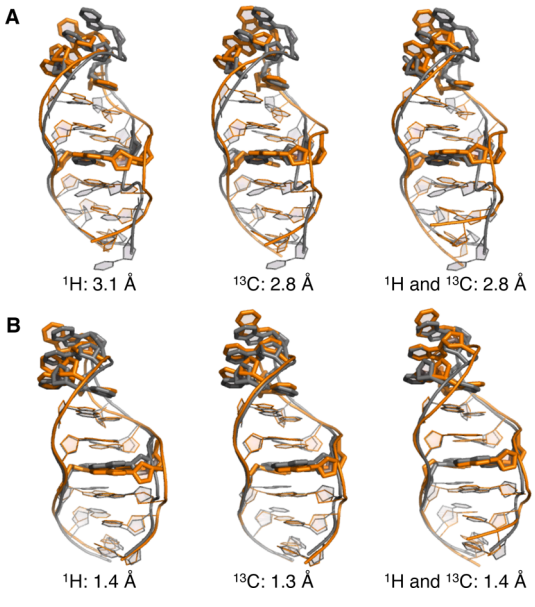


Figure 2. Structural Comparison between solved and low wMAE(r) model. Comparison between the (A) NMR and (B) X-ray solved structures (gray) and the lowest wMAE(r) model (orange) selected from MC-Sym generated conformational pool by using only ^{1}H (left), only ^{13}C (middle), or both ^{1}H and ^{13}C (right) RAMSEY-predicted chemical shifts. Indicated below each structure is the heavy atom RMSD between the solved structures and the MC-Sym models.

The results presented above demonstrate how chemical shifts could aid in RNA structure prediction. Starting from sequence, prediction methods like iFoldRNA,[41] MMB,[42] NAST,[43] FARNA,[44] Assemble,[45] BARNACLE,[46] RNA2D3D,[47] 3DRNA,[48] and MC-Sym[39] can be used to generate models that closely resemble the native structure for a target.[49] However, identifying and selecting representative models generated by these methods is still a significant challenge. A possible solution to this challenge would be to use experimental NMR chemical shifts in combination with RAMSEY-predicted chemical shifts as structural filters against the pool of predicted structures. The results for SRL RNA strongly suggest that RAMSEY-based chemical shift filters can be used to accurately select representative models from prediction pools. Extensive benchmarking on a diverse set of RNA targets will be needed to verify the feasibility of utilizing RAMSEY-predicted chemical shifts to aid RNA structure prediction. The possibility is particularly exciting because it would open up an opportunity of characterizing the structures of new RNA targets. Of particular interest is the use of ^{13}C chemical shifts to resolve RNA structure for low-populated excited states, which are believed to play key roles in RNA structure and function. New NMR methods are emerging that allow ^{13}C shifts for these transient states to be determined.[50] Currently, only ^{13}C chemical shifts can be determined for these excited state species, which further highlights the utility of generating chemical shift predictors for ^{13}C chemical shifts.

The results presented here are quite promising and bode well for the use of RAMSEY-predicted chemical shifts to aid in extracting RNA structural information directly from NMR chemical shifts. Admittedly, the accuracy of RAMSEY can be further improved: the R values for the predictors were ∼0.7, which is significantly lower than the accuracy observed for *sequence*-based approaches that predict ^{1}H chemical shifts for

helical regions ($R \sim 0.81$).[18] An immediate route to improving the accuracy of RAMSEY would be to expand the chemical shift-structure database used to train the predictors. Undoubtedly, some of the errors in the predictors are the result of errors in the NMR structures from which the structural features used to train the predictors are extracted. We anticipate that chemical shift predictors trained on a chemical shift-structure database consisting of high-resolution X-ray rather than NMR structures should exhibit superior accuracy to the predictors presented here. SHIFTX2,[32] a protein chemical shift predictor trained exclusively on high-resolution X-ray structures achieved ^{13}C RMSE of ∼0.5 ppm, demonstrating the importance of utilizing high-resolution structures when generating chemical shift predictors. Unfortunately, only a handful of RNAs have both NMR chemical shifts and corresponding high-resolution X-ray structures.

CONCLUSION

Because of the immense interest in RNA function and the unmet demand to rapidly generate realistic RNA 3D structures, new methods that enable RNA structural information to be efficiently extracted from readily accessible experimental observables are required. Here, we used the random forest machine learning technique to develop a method that efficiently and accurately predicts both ^{1}H and protonated ^{13}C NMR chemical shifts from RNA 3D coordinates. The set of predictors, which we collectively refer to as RAMSEY, were able to predict ^{1}H and protonated ^{13}C chemicals shifts with an RMSE of ∼0.16 and ∼0.90 ppm, respectively. In the case of ^{13}C, the predictors were shown to be accurate enough to detect data sets with known systematic referencing errors, and to provide reliable estimate of the magnitude of the error. Finally, using the SRL RNA as test system, we demonstrated that RAMSEY-predicted ^{1}H and ^{13}C chemical shifts were extremely sensitive to 3D structure and thus well suited to aid in RNA structure prediction and determination. Further work is required to systematically expand and curate the current chemical shift−structure database, which should improve the accuracy of RAMSEY, as well as enable the development of chemical shifts predictors for ^{15}N and ^{31}P nuclei.

ASSOCIATED CONTENT

Supporting Information

Tables S1−S4: (S1) Accuracy of A-form helical ^{1}H chemical shift predictors; (S2) compilation of RNA in the independent test set; (S3) results of the application of RAMSEY to the independent test set; (S4) list of RAMSEY chemical shift prediction outliers. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: afrank@nymirum.com or afrankz@umich.edu.

Present Address

[†]Department of Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, MI 48109-1055, United States

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Professor Hashim Al-Hashimi for many helpful discussions, and for reading the initial versions of this manuscript.

REFERENCES

(1) Eddy, S. R. Non-Coding RNA Genes and the Modern RNA World. *Nat. Rev. Genet.* **2001**, *2*, 919−929.
(2) Collins, L. J.; Penny, D. The RNA Infrastructure: Dark Matter of the Eukaryotic Cell? *Trends Genet.* **2009**, *25*, 120−128.
(3) ENCODE Project Consortium The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **2004**, *306*, 636−640.
(4) Cooper, T.; Wan, L.; Dreyfuss, G. RNA and Disease. *Cell* **2009**, *136*, 777−793.
(5) Stelzer, A. C.; Frank, A. T.; Kratz, J. D.; Swanson, M. D.; Gonzalez-Hernandez, M. J.; Lee, J.; Andricioaei, I.; Markovitz, D. M.; Al-Hashimi, H. M. Discovery of Selective Bioactive Small Molecules by Targeting an RNA Dynamic Ensemble. *Nat. Chem. Biol.* **2011**, *7*, 553−559.
(6) Doniach, S.; Lipfert, J. Use of Small Angle X-Ray Scattering (SAXS) to Characterize Conformational States of Functional RNAs. *Methods Enzymol.* **2009**, *469*, 237−251.
(7) Xia, Z. Z.; Bell, D. R. D.; Shi, Y. Y.; Ren, P. P. RNA 3D Structure Prediction by Using a Coarse-Grained Model and Experimental Data. *J. Phys. Chem. B* **2013**, *117*, 3135−3144.
(8) Tullius, T. D.; Greenbaum, J. A. Mapping Nucleic Acid Structure by Hydroxyl Radical Cleavage. *Curr. Opin. Chem. Biol.* **2005**, *9*, 127−134.
(9) Das, R.; Baker, D.; Herschlag, D. Structural Inference of Native and Partially Folded RNA by High-Throughput Contact Mapping. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4144−4149.
(10) Frank, A. T.; Horowitz, S.; Andricioaei, I.; Al-Hashimi, H. M. Utility of 1H NMR Chemical Shifts in Determining RNA Structure and Dynamics. *J. Phys. Chem. B* **2013**, *117*, 2045−2052.
(11) van der Werf, R. M.; Tessari, M.; Wijmenga, S. S. Nucleic Acid Helix Structure Determination From NMR Proton Chemical Shifts. *J. Biomol. NMR* **2013**, *56*, 95−112.
(12) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. Protein Structure Determination From NMR Chemical Shifts. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615−9620.
(13) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. Consistent Blind Protein Structure Generation From NMR Chemical Shift Data. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685−4690.
(14) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. Validation of Protein Structure From Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.
(15) Sahakyan, A. B.; Cavalli, A.; Vranken, W. F.; Vendruscolo, M. Protein Structure Validation Using Side-Chain Chemical Shifts. *J. Phys. Chem. B* **2012**, *116*, 4754−4759.
(16) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure* **2010**, *18*, 923−933.
(17) Jakovkin, I.; Klipfel, M.; Muhle-Goll, C.; Ulrich, A. S.; Luy, B.; Sternberg, U. Rapid Calculation of Protein Chemical Shifts Using Bond Polarization Theory and Its Application to Protein Structure Refinement. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12263.
(18) Barton, S.; Heng, X.; Johnson, B. A.; Summers, M. F. Database Proton NMR Chemical Shifts for RNA Signal Assignment and Validation. *J. Biomol. NMR* **2012**, *55*, 33−46.
(19) Giessner-Prettre, C.; Pullman, B. Quantum Mechanical Calculations of NMR Chemical Shifts in Nucleic Acids. *Q. Rev. Biophys.* **1987**, *20*, 113−172.
(20) Ghose, R.; Marino, J.; Wiberg, K.; Prestegard, J. Dependence of 13C Chemical Shifts on Glycosidic Torsional Angles in Ribonucleic Acids. *J. Am. Chem. Soc.* **1994**, *116*, 8827−8828.

(21) Farès, C.; Amata, I.; Carlomagno, T. 13C-Detection in RNA Bases: Revealing Structure−Chemical Shift Relationships. *J. Am. Chem. Soc.* **2007**, *129*, 15814−15823.
(22) Rossi, P.; Harbison, G. S. Calculation of 13C Chemical Shifts in RNA Nucleosides: Structure-13C Chemical Shift Relationships. *J. Magn. Reson.* **2001**, *151*, 1−8.
(23) Ebrahimi, M.; Rossi, P.; Rogers, C.; Harbison, G. S. Dependence of 13C NMR Chemical Shifts on Conformations of RNA Nucleosides and Nucleotides. *J. Magn. Reson.* **2001**, *150*, 1−9.
(24) Fonville, J. M.; Swart, M.; Vokáčová, Z.; Sychrovský, V.; Šponer, J. E.; Sponer, J.; Hilbers, C. W.; Bickelhaupt, F. M.; Wijmenga, S. S. Chemical Shifts in Nucleic Acids Studied by Density Functional Theory Calculations and Comparison with Experiment. *Chem.—Eur. J.* **2012**, *18*, 12372−12387.
(25) Suardíaz, R.; Sahakyan, A. B.; Vendruscolo, M. A Geometrical Parametrization of C1′-C5′ RNA Ribose Chemical Shifts Calculated by Density Functional Theory. *J. Chem. Phys.* **2013**, *139*, 034101.
(26) Dejaegere, A.; Bryce, R. A.; Case, D. A. An Empirical Analysis of Proton Chemical Shifts in Nucleic Acids. *ACS Symp. Ser.* **1999**, *732*, 194−206.
(27) Cromsigt, J. A.; Hilbers, C. W.; Wijmenga, S. S. Prediction of Proton Chemical Shifts in RNA. Their Use in Structure Refinement and Validation. *J. Biomol. NMR* **2001**, *21*, 11−29.
(28) Meiler, J. J. PROSHIFT: Protein Chemical Shift Prediction Using Artificial Neural Networks. *J. Biomol. NMR* **2003**, *26*, 25−37.
(29) Shen, Y.; Bax, A. Protein Backbone and Sidechain Torsion Angles Predicted From NMR Chemical Shifts Using Artificial Neural Networks. *J. Biomol. NMR* **2013**, *56*, 227−241.
(30) Schapire, R. E. The Strength of Weak Learnability. *Machine learning* **1990**, *5*, 197−227.
(31) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(32) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. SHIFTX2: Significantly Improved Protein Chemical Shift Prediction. *J. Biomol. NMR* **2011**, *50*, 43−57.
(33) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(34) Szewczak, A. A. A.; Moore, P. B. P.; Chang, Y. L. Y.; Wool, I. G. I. The Conformation of the Sarcin/Ricin Loop From 28S Ribosomal RNA. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9581−9585.
(35) Correll, C. C. C.; Munishkin, A. A.; Chan, Y. L. Y.; Ren, Z. Z.; Wool, I. G. I.; Steitz, T. A. T. Crystal Structure of the Ribosomal RNA Domain Essential for Binding Elongation Factors. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 13436−13441.
(36) Aeschbacher, T.; Schubert, M.; Allain, F. H.-T. A Procedure to Validate and Correct the 13C Chemical Shift Calibration of RNA Datasets. *J. Biomol. NMR* **2012**, *52*, 179−190.
(37) Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E., III; Jurečka, P. Refinement of the Cornell Et Al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.
(38) Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18−22.
(39) Parisien, M.; Major, F. The MC-Fold and MC-Sym Pipeline Infers RNA Structure From Sequence Data. *Nature* **2008**, *452*, 51−55.
(40) Case, D. A. Chemical Shifts in Biomolecules. *Curr. Opin. Struct. Biol.* **2013**, *23*, 172−176.
(41) Sharma, S. S.; Ding, F. F.; Dokholyan, N. V. N. iFoldRNA: Three-Dimensional RNA Structure Prediction and Folding. *CABIOS, Comput. Appl. Biosci.* **2008**, *24*, 1951−1952.
(42) Flores, S. C. S.; Sherman, M. A. M.; Bruns, C. M. C.; Eastman, P. P.; Altman, R. B. R. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *8*, 1247−1257.
(43) Jonikas, M. A.; Radmer, R. J.; Laederach, A.; Das, R.; Pearlman, S.; Herschlag, D.; Altman, R. B. Coarse-Grained Modeling of Large RNA Molecules with Knowledge-Based Potentials and Structural Filters. *RNA* **2009**, *15*, 189−199.
(44) Das, R.; Baker, D. Automated De Novo Prediction of Native-Like RNA Tertiary Structures. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14664−14669.
(45) Jossinet, F.; Ludwig, T. E.; Westhof, E. Assemble: an Interactive Graphical Tool to Analyze and Build RNA Architectures at the 2D and 3D Levels. *Bioinformatics* **2010**, *26*, 2057−2059.
(46) Frellsen, J.; Moltke, I.; Thiim, M.; Mardia, K. V.; Ferkinghoff-Borg, J.; Hamelryck, T. A Probabilistic Model of RNA Conformational Space. *PLoS Comput. Biol.* **2009**, *5*, e1000406.
(47) Martinez, H. M. H.; Maizel, J. V. J.; Shapiro, B. A. B. RNA2D3D: a Program for Generating, Viewing, and Comparing 3-Dimensional Models of RNA. *J. Biomol. Struct. Dyn.* **2008**, *25*, 669−683.
(48) Zhao, Y.; Huang, Y.; Gong, Z.; Wang, Y.; Man, J.; Xiao, Y. Automated and Fast Building of Three-Dimensional RNA Structures. *Sci. Rep.* **2012**, *2*.
(49) Laing, C.; Schlick, T. Computational Approaches to 3D Modeling of RNA. *J. Phys.: Condens. Matter* **2010**, *22*, 283101.
(50) Dethoff, E. A.; Petzold, K.; Chugh, J.; Casiano-Negroni, A.; Al-Hashimi, H. M. Visualizing Transient Low-Populated Structures of RNA. *Nature* **2012**, 1−7.
(51) Sich, C. C.; Ohlenschläger, O. O.; Ramachandran, R. R.; Görlach, M. M.; Brown, L. R. L. Structure of an RNA Hairpin Loop with a 5"-CGUUUCG-3" Loop Motif by Heteronuclear NMR Spectroscopy and Distance Geometry. *Biochemistry* **1997**, *36*, 13989−14002.