Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations
Bottaro, Sandro; Bussi, Giovanni; Kennedy, Scott D.; Turner, Douglas H.; Lindorff-Larsen, Kresten

Published in:
Science Advances

DOI:
10.1126/sciadv.aar8521

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):

Download date: 27. apr., 2019
Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations

Sandro Bottaro,1* Giovanni Bussi,2 Scott D. Kennedy,3 Douglas H. Turner,4 Kresten Lindorff-Larsen1*

RNA molecules are key players in numerous cellular processes and are characterized by a complex relationship between structure, dynamics, and function. Despite their apparent simplicity, RNA oligonucleotides are very flexible molecules, and understanding their internal dynamics is particularly challenging using experimental data alone. We show how to reconstruct the conformational ensemble of four RNA tetranucleotides by combining atomistic molecular dynamics simulations with nuclear magnetic resonance spectroscopy data. The goal is achieved by reweighting simulations using a maximum entropy/Bayesian approach. In this way, we overcome problems of current simulation methods, as well as in interpreting ensemble- and time-averaged experimental data. We determine the populations of different conformational states by considering several nuclear magnetic resonance parameters and point toward properties that are not captured by state-of-the-art molecular force fields. Although our approach is applied on a set of model systems, it is fully general and may be used to study the conformational dynamics of flexible biomolecules and to detect inaccuracies in molecular dynamics force fields.

INTRODUCTION

Many biomolecules are highly dynamic systems that undergo significant conformational rearrangements during their function. Experimental techniques such as nuclear magnetic resonance (NMR) spectroscopy, fluorescence spectroscopy, and small-angle x-ray scattering are well suited to probe the dynamics of molecules in solution. However, obtaining a full description of structure and dynamics of biomolecules using experiments alone can be highly nontrivial because the measured quantities are generally time and ensemble averages over conformationally heterogeneous states.

In this perspective, maximum entropy (1–3) (MaxEnt) and Bayesian (4–6) approaches have emerged as powerful theoretical tools for integrating simulations with experiments. These approaches typically generate a structural ensemble for the system of interest using molecular dynamics (MD) or Monte Carlo simulations. This ensemble, however, may not necessarily agree with available experimental data because of limited sampling or inaccuracies in the used model describing the physics and chemistry of the system (that is, the force field). The underlying idea behind MaxEnt is to minimally perturb a simulation ensemble to match the experimental data. Random and systematic errors can be taken explicitly into account. These approaches have been successfully used to study protein systems (6), whereas applications to nucleic acids have been so far limited (7, 8).

Here, we consider the conformational ensembles of four RNA tetranucleotides by integrating available NMR data (9–11) with extensive atomistic MD simulations. Despite the lack of a biological relevance, RNA tetranucleotides serve as challenging model systems both from the experimental and computational point of view. First, they display a higher complexity because the optimal ensembles consist of a mixture of A-form-like structures is confirmed in our X_{O,L} simulations, as shown in the eRMSD histogram from an ideal A-form in Fig. 1B (yellow line). To measure distances between three-dimensional structures, we here use the eRMSD, an RNA-specific

RESULTS

Agreement between experiments and simulations

We first consider the tetranucleotide with sequence CCCC. Previous NOE measurements for r(CCCC) were found to be consistent with a conformational ensemble mostly composed of A-form–like structures, with a minor population (13%) of conformations with cytosine at position 4 (C4) inverted (see Fig. 1A) (10). Extensive MD simulations with the standard Assisted Model Building with Energy Refinement (AMBER) force field (X_{O,L} described in Materials and Methods) showed the presence of highly populated intercalated structures in which C1 is interposed between C3 and C4 (11, 13), whereas C2 is either stacked on C3 or solvent-exposed. The lack of A-form–like structures is confirmed in our X_{O,L} simulations, as shown in the eRMSD histogram from an ideal A-form in Fig. 1B (yellow line). To measure distances between three-dimensional structures, we here use the eRMSD, an RNA-specific
metric distance based on the relative orientation and position of nucleobases (14). It has recently been reported (15) that corrections to oxygen van der Waals radii (16) in conjunction with the optimal 3-charge, 4 point (OPC) water model (17) (hereafter referred to as VdW-OPC) significantly disfavor the presence of intercalated structures in r(GACC) and r(CCCC) tetranucleotides, thereby stabilizing A-form–like conformations. When using the VdW-OPC force field (Fig. 1B, blue line), we observe a small, yet significant population of A-form–like structures (eRMSD < 0.75) and C4-inverted conformations (0.75 to 1.0 eRMSD from A-form).

The higher accuracy of VdW-OPC with respect to \( \chi_{\text{OL3}} \) is further confirmed by the improved agreement between calculated and experimental data. Figure 1C reports the \( \chi^2 \) for backbone-3J scalar couplings (H3-P, H5'/H5″-P, and H4-H5'/H5″), sugar-3J couplings (H1′-H2′, H2′-H3′, and H3′-H4′), and NOE intensities (10, 11). It is evident from Fig. 1C that the conformational ensemble predicted by simulations alone is not in complete agreement with experiments. We therefore generate a conformational ensemble that satisfies the experimental constraints using the MaxEnt/Bayesian approach with the inclusion of error treatment (5, 7). In MaxEnt approaches, one seeks the minimal perturbation of the simulated ensemble (that is, the prior distribution) that satisfies a set of known experimental averages. This can be achieved (2, 7) by minimizing the function

\[
\Gamma = \log(Z(\lambda)) + \sum_{i=1}^{m} \lambda_i F_{i}^{\text{EXP}} + \frac{1}{2} \sum_{i=1}^{m} \lambda_i^2 \sigma_i^2 \tag{1}
\]

with respect to the set of Lagrange multipliers \( \lambda = \lambda_1 \ldots \lambda_m \). Here, the index \( i \) runs over the \( m \) experimental averages \( F_{i}^{\text{EXP}} \) with associated normally distributed and uncorrelated errors \( \sigma_i \). \( Z \) is the partition function \( Z(\lambda) = \sum_{\{w_i\}} \exp[-\sum_{i=1}^{m} \lambda_i F_{i}(x_i)] \), where \( F_i(x_i) \) is the function used to back-calculate the experimental observable from the atomic coordinates \( x_i \), and \( \{w_1 \ldots w_N\} \) corresponds to the weights of the \( N \) frames in the prior distribution. Note that this approach is completely equivalent to a Bayesian ensemble refinement approach (5, 19) in which one seeks the optimal weights \( \{w_1 \ldots w_N\} \) minimizing the negative log posterior \( L \)

\[
L(w_1 \ldots w_N) = \frac{m}{2} \chi^2 + \theta S_{\text{REL}} \tag{2}
\]

where \( \chi^2 = \sum_{i=1}^{m} \left(F_i(x_i) - F_i^{\text{EXP}}\right)^2/m \sigma_i^2 \) is the deviation from the experimental averages, and the relative entropy \( S_{\text{REL}} = \sum_{i=1}^{m} w_i \log(w_i/w_i^0) \) quantifies the deviation from the prior distribution. \( \theta \) sets the relative weight between these two quantities and needs to be chosen by considering how \( \chi^2 \) and \( S_{\text{REL}} \) vary for different values of this parameter (5), as described below.

A few items are worth highlighting. First, the number of experimental constraints, \( m \), is typically much smaller compared to the number of samples, \( N \), and it is therefore in practice easier to minimize the function in Eq. 1 rather than Eq. 2. Second, \( \theta \) enters the MaxEnt formulation (Eq. 1) as a global scaling factor of all Gaussian errors \( \sigma_i \). Third, heterogeneous data (NOE, 3J couplings, chemical shifts, etc.)

Fig. 1. Conformational ensemble of r(CCCC) simulations and agreement with experimental data. (A) Three-dimensional structures of r(CCCC) discussed in the main text. (B) eRMSD from the A-form histogram for \( \chi_{\text{OL3}} \) and VdW-OPC simulations. Solid lines indicate the average calculated using a blocking procedure, whereas the area between minimum and maximum is shown in shade. The histogram displays three peaks corresponding to different conformations: A-form–like (eRMSD < 0.75), C4-inverted (0.75 < eRMSD < 1), and intercalated/C2 unstacked (eRMSD > 1.0). Thresholds are shown as dashed lines. (C) Agreement between simulations and experiments quantified using the \( \chi^2 \) statistic for backbone scalar couplings (3J bb), sugar scalar couplings, NOE, and uNOE. Error bars in black show the SEM. The value of \( \chi^2 \) relative to NOE with \( \chi_{\text{OL3}} \) is out of scale; the corresponding value with error is therefore reported in the figure.
can be used simultaneously in the reweighting procedure, both averages and inequality constraints (7).

**Choosing the data and the confidence parameter**
Before proceeding to the analysis of the optimized ensemble, we study the dependence of the results on (i) the type of experimental data used for reweighting and (ii) the tunable parameter $\theta$. Given the better initial agreement with experimental data, we here consider the VdW-OPC simulations. Figure 2A (solid lines) shows $\chi^2$ as a function of $\theta$ when using scalar couplings as the only input for reweighting. As expected, small $\theta$ corresponds to a better fit, whereas in the limit of large $\theta$ we approach the original, unreweighted $\chi^2$ value (dotted-dashed line). We can also monitor the behavior of $\chi^2$ relative to data that were not used in the reweighting (Fig. 2A, dashed line). In the limit of $\theta \to 0$, the violations of uNOE become very small. Conversely, the agreement with NOE distances has a clear minimum around $\theta = 3$. When using only NOEs for reweighting (Fig. 2B), we observe improved agreement with respect to all other experimental sources of data. This effect is more pronounced when using uNOE only (Fig. 2C), demonstrating the importance and the validity of this type of data. Note that, at least for r(CCCC), the reweighted $\chi^2$ values are always smaller compared to the original, unreweighted values, indicating that the different types of data are consistent. Given the cooperative effect of the different types of data, we finally consider the case in which $^3J$ couplings, NOE, and uNOE are all used at the same time for reweighting (Fig. 2D). This combination provides the best accord both for r(CCCC) and for the other tetranucleotides (figs. S1 to S3).

When considering $\chi^2$ alone, one would choose a small $\theta$ so as to attain the best fit. In the limit $\theta \to 0$, however, the original ensemble can be substantially distorted to the point that the physicochemical information contained in the force field is lost (Eq. 2). In addition, this has a detrimental effect on the statistical errors because the number of effective frames contributing to the ensemble decreases significantly (fig. S4). To strike a good balance between fit and proximity to the prior distribution, we scan different values of $\theta$ until a further decrease of this parameter leads to an increase in the relative entropy without substantially improving the fit (5). Although this procedure does not provide a unique $\theta$, it makes it possible to identify a range of reasonable values (fig. S4). We here use a pragmatic approach and set $\theta = 2$, the largest value for which $\chi^2 < 2$ for all tetranucleotides and all types of experimental data. Note that the relative weight of different experiments might be modulated by changing the corresponding values of $\sigma$. Scatter plots comparing individual experimental averages against simulations before/after reweighting are shown in figs. S5 to S8.

**Conformational ensemble of r(CCCC)**
The set of optimized weights can be now used to calculate the full probability distribution of any observable (for example, distances, torsion angles, etc.). To appreciate the properties of the optimized ensemble, it is again interesting to consider the distribution of the distance from A-form (Fig. 3A).

The original VdW-OPC MD ensemble consists of $\approx$18% A-form structures (eRMSD from A-form < 0.75) and 9% with C4 either inverted or unstacked (eRMSD from A-form in the 0.75 to 1.0 range). From the histogram of eRMSD relative to intercalated structure (Fig. 3B), the initial ensemble estimates a 53% population of intercalated structures that can be subdivided into fully stacked intercalation (13%, eRMSD < 0.4) and intercalated structures with C2 unstacked ($\approx$ 40%, eRMSD in the 0.4 to 0.8 range).

Upon reweighting, A-form represents the major conformation (54%) followed by C4 inverted (22%). The population of intercalated structures is significantly reduced in the reweighted ensemble to $\approx$7% (Fig. 3B). This result is not surprising because it is consistent with the picture proposed in the original experimental paper (10). The ensemble obtained here, however, did not require expert interpretation of the individual NOE distances. The reweighting approach takes into account general properties encoded in the force field and makes it possible to monitor degrees of freedom that were not measured by NMR. Two significant examples are reported in Fig. 3 (C and D). Figure 3C shows the distribution of the distance between the atom OP2 in C3 and the hydrogen at the 5’ terminus in C1 (H5T), where we observe the presence of a stable hydrogen bond between these two atoms (associated with the intercalated conformation) that is almost absent after reweighting. The reweighting also markedly affects the distribution of the $\alpha$ angle in C2, because we find that gauche$^-$ ($^g$) is the preferred rotameric state in the reweighted ensemble (Fig. 3D). A similar behavior is observed for $\alpha$ in C3 and $\xi$ in C2 and in C3, in accordance with previous simulation studies that have shown the importance of these two torsion angles in tetranucleotides and tetraloop simulations (20, 21). We highlight that the backbone
might require manual inspection. These cases can be treated by using
highlight data points that are inconsistent with the others and hence
data points. Evidently, the reweighting procedure can be used to
overlaps. The corresponding NOEs were thus removed from the list of
experimental data, we discovered two previously undetected spectral
satisfied in a preliminary reweighting. After careful checking of the
(complete range (figs. S5 to S8). In the case of r(GACC), three
NOEs reported in the original experimental work (11)
A-form, (B) eRMSD from an intercalated conformation, (C) distance
between OP2 in C3 and H5T in C1, and (D) the α torsion angle of
c2. Peaks in (A) and (B) can be associated to the structures shown in Fig. 1: A-form
(eRMSD from A-form below 0.75), C4-inverted (eRMSD from A-form 0.75 to 1.0),
iclarated (eRMSD from intercalated <0.4), and intercalated with C2 unstacked
(eRMSD from intercalated 0.4 to 0.8). eRMSD boundaries are shown as dashed lines.

Fig. 3. Distribution of different observables before and after reweighting
r(CCCC) simulations using VdW-OPC. Solid lines indicate the average calculated
using a blocking procedure; minima and maxima are shown in shade. (A) eRMSD
from ideal A-form, (B) eRMSD from an intercalated conformation, (C) distance
between OP2 in C3 and H5T in C1, and (D) the α torsion angle of
c2. Peaks in (A) and (B) can be associated to the structures shown in Fig. 1: A-form
(eRMSD from A-form below 0.75), C4-inverted (eRMSD from A-form 0.75 to 1.0),
iclarated (eRMSD from intercalated <0.4), and intercalated with C2 unstacked
(eRMSD from intercalated 0.4 to 0.8). eRMSD boundaries are shown as dashed lines.

3J scalar couplings used in the reweighting procedure report on ε and
γ angles, but not on α/ζ.

Conformational ensemble of r(AAAA), r(GACC), and r(UUUU)
The same procedure described above was applied to r(AAAA), r(GACC),
and r(UUUU) tetranucleotides. In all cases, VdW-OPC is considerably
better compared with the χ_{OL3} force field (Fig. 4, left panels). The re-
weighting procedure further improves agreement with experimental
data. However, we do observe a residual discrepancy in some cases
(χ^2 > 1) that stems from predicted NOE distances falling outside the
experimental range (figs. S5 to S8). In the case of r(GACC), three
NOEs reported in the original experimental work (11) were not
satisfied in a preliminary reweighting. After careful checking of the
experimental data, we discovered two previously undetected spectral
overlaps. The corresponding NOEs were thus removed from the list of
data points. Evidently, the reweighting procedure can be used to
highlight data points that are inconsistent with the others and hence
might require manual inspection. These cases can be treated by using
error models suitable to describe outliers (7, 22).

The r(AAAA) ensemble is composed of ≈30% A-form–like
structures and 16% A4-inverted/unstacked (Fig. 4, middle panels). In this case, the available experimental data could not completely
rule out the presence of intercalated structures, which represent the
13% of the optimized ensemble (Fig. 4, right panels). The re-
main­ing 40% is composed of other structures that exhibit one or
more sugar puckers in C2′-endo and/or the A1-γ angle in syn con-
formation (Table 1 and fig. S9).

r(GACC) behaves very similarly to r(CCCC), with ≈60% A-form–
like structures and 20% C4-inverted/unstacked. The similarity be-
 tween r(GACC) and r(CCCC) can also be appreciated by considering
the sugar pucker and γ angle preferences reported in Table 1 and
figs. S10 and S11. Intercalation is almost completely absent in the re-
weighted ensembles.

Among all the systems studied here, r(UUUU) has the lowest
population of A-form-like structures (9%). The rest of the ensemble
is composed of a variety of diverse structures that cannot be easily
clustered. This can be seen from the low percentage of sugar pucker in
C3′-endo conformation (Table 1 and fig. S12) and from the rela-
tively flat distribution of eRMSD from A-form in Fig. 4. Among this
set of diverse conformations, a very small fraction of intercalated
structures are present.

Note that the percentages reported here depend on two important
choices: on the reference structures and on the choice of θ. Whereas
the geometry of the ideal A-form can be unambiguously defined (23),
the intercalated structures are obtained by performing a cluster anal-
ysis of the χ_{OL3} simulation as described previously (24). Although this
choice has a degree of arbitrariness, we found it as useful and intu-
itive manner to define an order parameter complementary to the dis-
tance from A-form. As for θ, we verified that the population of the
different states do not depend critically on this parameter in the rele-
vant range 2 < θ < 5 (fig. S13).

DISCUSSION
Here, we have described the structural ensembles of four RNA tetra-
nucleotides at the atomistic level. The characterization of these systems
represents a first step in understanding the ensembles and internal dy-
namics of larger oligonucleotides and other RNA molecules under-
going significant conformational changes. Despite their apparent
simplicity, tetranucleotides are particularly challenging systems: Be-
cause of their conformational heterogeneity, NMR experimental data
need to be interpreted as ensemble averages. For this reason, standard
procedures for NMR structure determination cannot be easily applied
(25). In addition, it is not possible to predict the properties of these
systems using simulations alone, because of known force-field inac-
curacies (Fig. 1). Only the combination of experiment with compu-
tation makes it possible to provide an atomic-detailed description of their
conformational ensembles. In this context, the MaxEnt/Bayesian
approach serves as a fundamental theoretical ingredient for using
the two techniques in conjunction.

We find that r(CCCC) and r(GACC) are ≈60% in A-form–like
conformations and ≈20% with the 3′ terminal base either unstacked
or inverted (Fig. 1A). r(AAAA) tetranucleotide is characterized by a
lower A-form content (≈30%) and displays a larger variability in
terms of sugar conformations. Our analysis shows that the presence
of intercalated structures cannot be excluded in this case. Among the
four systems considered here, r(UUUU) displays the highest disorder
(Table 2), with a percentage of A-form conformation of ≈10%.
From a technical perspective, the combination of experiments and simulations can be seen as a regularization problem in which a small set of experimental data is used to gain insights into a highly dimensional, complex set of molecular conformations. The problem is underdetermined and has to be regularized by using a suitable prior distribution, here provided by MD simulations. This interpretation becomes transparent in the Bayesian ensemble refinement formula in Eq. 2. These approaches have the advantage of enhancing the sampling in relevant regions of the conformational space. On the other hand, the reweighting procedure can be applied a posteriori to

**Table 1.** Percentage of C3′-endo (δ < 115°) and anti (γ > 120°) of reweighted VdW-OPC simulations. The statistical error calculated using block averaging is below 1%.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
</tr>
</thead>
<tbody>
<tr>
<td>% C3'-endo</td>
<td>AAAA 70.6 75.1 84.5 66.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCC 90.7 88.9 88.8 71.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GACC 86.9 87.1 88.3 71.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUUU 61.4 49.5 50.5 63.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Anti</td>
<td>AAAA 65.2 96.5 98.4 97.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCC 98.0 98.5 99.8 99.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GACC 89.2 99.9 98.9 99.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUUU 88.5 97.1 96.8 96.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2.** Number of experimental averages.

<table>
<thead>
<tr>
<th>NOE</th>
<th>3J sugar</th>
<th>3J backbone</th>
<th>uNOE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAA 36</td>
<td>11</td>
<td>17</td>
<td>243</td>
</tr>
<tr>
<td>CCC 27</td>
<td>11</td>
<td>15</td>
<td>245</td>
</tr>
<tr>
<td>GACC 20</td>
<td>12</td>
<td>17</td>
<td>284</td>
</tr>
<tr>
<td>UUUU 9</td>
<td>10</td>
<td>15</td>
<td>282</td>
</tr>
</tbody>
</table>

**Fig. 4.** Comparison between reweighted and unreweighted ensembles for r(AAAA), r(GACC), and r(UUUU) tetranucleotides. (Left) Agreements between calculated and experimental averages for X_{OL3}, VdW-OPC, and reweighted VdW-OPC simulations. (Middle) Histograms of eRMSD from ideal A-form. (Right) Histograms of eRMSD from intercalated structure. The dashed lines indicate the thresholds used for calculating the percentage of A-form–like (middle) or intercalated structures (right) upon reweighing.
existing simulations whenever new experimental data are available (27). Because reweighting only requires a cheap post-processing of existing trajectories, it is straightforward to perform multiple cross-validation tests. In addition, reweighting is very convenient when the forward model calculation is particularly demanding, because in biased methods the back calculation of averages from structures has to be performed at least every few time steps (28).

Here, we have found that combining experimental data with simulations had mutual beneficial effects. On the one hand, simulations helped identifying spurious experimental data points. On the other hand, we have used experimental data to identify inaccuracies in MD force fields. Modern atomistic force fields consist of hundreds of parameters, and even finding the relevant interactions that can potentially improve their accuracy is a time-consuming and nontrivial task. Our approach substantially simplifies this search (Fig. 3, C and D), because the probability distribution over any degree of freedom before and after reweighting can be readily compared. We find that hydrogen bonds to nonbridging oxygens are significantly destabilized upon reweighting, in accordance with previous simulation studies (11, 29). At the same time, the population of τ and γ torsion angles is, in some cases, shifted from gauche+ to gauche−. As molecular mechanics force fields improve, the approach described here should require less experimental data to provide reliable determination of structural ensembles (30, 31).

MATERIALS AND METHODS

MD simulations

We performed MD simulations on r(AAAA), r(CCCC), r(UUUU), and r(GACC) tetrarnucleotides. Each system was simulated with two different force fields: (i) the AMBER 99 force field (32) with parameters that were used in previous studies (33) and the χOL corrections to χ torsion angles (34) in TIP3P water. We refer to this combination as χOL3L. These simulations were taken from our previous studies (20, 35). (ii) χOL3L with corrections to van der Waals oxygen radii (16) (atom types O2, OH, and OS) and using the OPC water model (17). We refer to this combination as VdW-OPC. Parameters are available at http://github.com/srnas/ff. MD simulations were performed using the GROMingen MAchine for Chemical Simulations (GROMACS) 4.6.7 software package (36). Ideal A-form, fully stacked initial conformations were generated using the Make-NA web server. The oligonucleotides were solvated in a truncated dodecahedric box and neutralized by adding Na+ counterions (37). Initial conformations were minimized in vacuum first, followed by a minimization in water and equilibration in NPT ensemble at 300 K and 1 bar for 1 ns. Production runs were performed in the canonical ensemble using a stochastic velocity rescaling thermostat (38). All bonds were constrained with the Linear Constraint Solver algorithm, and equations of motion were integrated with a time step of 2 fs. Tetrarnucleotides were simulated using temperature replica exchange (39) using 24 replicas in the temperature range of 278 to 400 K for 1.0 μs per replica. All the analyses presented here were performed for the 300 K replica and using 20,000 frames. Averages and SEMs were calculated using four blocks of 5000 samples each. Sampling was sufficient to achieve similar eRMSD distributions for each block (fig. S14) and to obtain populations of different substates in agreement with multidimensional replica exchange MD simulations (12, 13, 15).

NMR data

Experimental NOE and scalar couplings have previously been published (10, 11). We used Gaussian-distributed experimental errors of 1.5 Hz for scalar couplings and of 0.1 Å for uNOE. The error for NOE was estimated as min(3J exp − 3J min, 3J max − 3J exp). The number of experimental averages for each NMR parameter and for each tetrarnucleotide sequence is reported in Table 2. The complete list of experimental data is available in the Supplementary Materials. NOE intensities from simulations are calculated as averages over the N samples NOEcalc = (ΣN wib exp−6). 3J scalar couplings were calculated using the Karplus relationships described in fig. S15 and table S1 using the software baRNAba https://github.com/srnas/barnaba. Note that in some cases, the error introduced by the forward model is significant. As an example, 3J scalar couplings calculated using Karplus relationships can introduce errors up to 2 Hz (fig. S16). Care should also be taken when calculating NOE intensities from proton-proton distances because the simple r−6 averaging does not take spin diffusion into account, and it is only valid in the limit of slow internal motion compared to the tumbling time (40).

Unobserved NOE

NMR spectra were inspected for the presence of NOESY cross peaks between every pair of protons in the tetramer. If no cross peak was observed, then the potential contact was classified as a uNOE. If the spectral position of a potential cross peak did not overlap any other observed cross peak, then the minimum detectable cross-peak volume was assumed to be two times the SD of spectral noise (Verr). Scalar coupling results in NOE cross peaks that are split into multiplets of two, four, or more peaks, resulting in accordingly reduced peak heights and increased minimum detectable volume. For a cross peak consisting of M multiplets, the minimum detectable volume is 2MVerr. Verr and a scaling factor, c, obtained in the original work (10, 11) from NOESY spectra with a 200-ms mixing time, are used to associate a distance, R, with the minimum detectable volume: R = (c/2Mverr)1/6. The analysis of uNOEs was carried out here with 800-ms NOESY spectra, where cross peaks are typically 2.5- to 3-fold greater than at 200 ms, so the minimum detectable NOE volume was reduced by a factor of 2.5 (after correcting for any difference in the number of NMR scans). If the spectral position of a potential cross peak partially overlapped one or more observed cross peaks, then the minimum detectable volume of the potential cross peak was determined by the magnitude of the observed cross peak and exact details of the overlap (instead of spectral noise). Typically, if the partially overlapped observed cross peak was medium or weak, respectively, then a potential cross peak exhibiting no apparent intensity was classified as unobserved with a volume that corresponded to an internuclear distance of greater than 3.3 or 4.0 Å. If the overlapping observed cross peak was strong or the potential cross peak was close to the diagonal, then the potential cross peak was not classified as unobserved.
fig. S10. Torsion angle distribution before (blue) and after (gray) reweighting r(GGCU) simulations with θ = 2.

fig. S11. Torsion angle distribution before (blue) and after (gray) reweighting r(GCCG) simulations with θ = 2.

fig. S12. Torsion angle distribution before (blue) and after (gray) reweighting r(UUUU) simulations with θ = 2.

fig. S14. Histogram of eRMSD from A-form and intercalated in four simulation blocks of 5000 s per block.

fig. S15. Karplus equations listed in table S1 (vide infra) overlayed on experimental data from previous studies (41–44).

fig. S16. Root mean square error between calculated and experimental 1J couplings.

REFERENCES AND NOTES


Acknowledgments: We thank S. Olsson for helpful comments on the manuscript. Funding: The research is funded by a grant from The Velux Foundations (S.B. and K.L.-L.), a Hallas-Møller Stipend from the Novo Nordisk Foundation (K.L.-L.), and the Lundbeck Foundation BRAINSTRUC initiative (K.L.-L.). G.B. has received funding from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC grant agreement no. 306662 (5-RNA-S). D.H.T. was supported by NIH grant R01 GM02939. Author contributions: S.B. performed simulations, created figures, and led the project. G.B. contributed with theoretical tools for combining simulations with experiments. S.D.K. and D.H.T. collected the experimental data. K.L.-L. and S.B. conceived the study and supervised the project. All authors analyzed data and contributed to writing the paper. Competing interests: The authors declare that they have no competing interests. Data and materials Availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Experimental data are available at https://github.com/sbottaro/tetr nucleotides_data. The Python script used to reweight trajectories is available at https://github.com/sbottaro/rr.

Submitted 23 December 2017
Accepted 5 April 2018
Published 18 May 2018
10.1126/sciadv.aar8521

Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations
Sandro Bottaro, Giovanni Bussi, Scott D. Kennedy, Douglas H. Turner and Kresten Lindorff-Larsen

Sci Adv 4 (5), eaar8521
DOI: 10.1126/sciadv.aar8521

ARTICLE TOOLS http://advances.sciencemag.org/content/4/5/eaar8521
SUPPLEMENTARY MATERIALS http://advances.sciencemag.org/content/suppl/2018/05/14/4.5.eaar8521.DC1
REFERENCES This article cites 47 articles, 3 of which you can access for free http://advances.sciencemag.org/content/4/5/eaar8521#BIBL
PERMISSIONS http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service