Conformational Ensemble of RNA Oligonucleotides from Reweighted Molecular Simulations

Sandro Bottaro[a], **Giovanni Bussi**[b], **Scott D. Kennedy**[c], **Douglas H. Turner**[d], **and Kresten Lindorff-Larsen**[a]

[a]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [b]SISSA, International School for Advanced Studies. Trieste, Italy; [c]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA.; [d]Department of Chemistry, University of Rochester, Rochester, NY 14627, USA

This manuscript was compiled on December 6, 2017

We determine the conformational ensemble of four RNA tetranucleotides by using available nuclear magnetic spectroscopy data in conjunction with extensive atomistic molecular dynamics simulations. This combination is achieved by applying a reweighting scheme based on the maximum entropy principle. We provide a quantitative estimate for the population of different conformational states by considering different NMR parameters, including distances derived from nuclear Overhauser effect intensities and scalar coupling constants. We show the usefulness of the method as a general tool for studying the conformational dynamics of flexible biomolecules as well as for detecting inaccuracies in molecular dynamics force fields.

RNA | Molecular Dynamics | NMR | Maximum Entropy

1. 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 (SAXS) 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 non-trivial, 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) approaches have emerged as powerful theoretical tools for integrating simulations with experiments. Such 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, due to limited sampling or to inaccuracies in the employed model describing the physics and chemistry of the system (i.e. the force field). The underlying idea behind MaxEnt is to minimally perturb a simulation ensemble so as to match the experimental data. Random as well as systematic errors can be taken explicitly into account. The modification to the ensemble can be either performed on-the-fly, or even a posteriori by reweighting existing simulations. These approaches have been successfully employed to study protein systems (5), while applications to nucleic acids have been so far limited (6, 7).

In this paper we consider the conformational ensembles of four RNA tetranucleotides by integrating available NMR data (8–10) with extensive atomistic MD simulations. Despite their apparent simplicity, tetranucleotides are particularly challenging systems both from the experimental and computational point of view. First, they display significant dynamics: therefore one single structure cannot be representative of the entire ensemble. The conformational heterogeneity makes it non-trivial to provide a structural interpretation of average measurements using standard three-dimensional structure determination tools. Second, current state-of-the-art molecular dynamics force fields fail in predicting the properties of these tetranucleotides (11). Several studies (10, 12) have shown MD simulations to over-stabilize so-called intercalated conformations (see Fig.1), that in some cases correspond to the predicted free-energy minimum. From the experimental point of view the presence and the population of intercalated conformations is expected to be low, but cannot be accurately quantified.

Here we show that, even with the aforementioned complications, it is possible to obtain an accurate thermodynamic description for a system of interest by combining experiments and simulations. We report extensive atomistic MD simulations in explicit water for r(AAAA), r(CCCC), r(GACC) and r(UUUU)

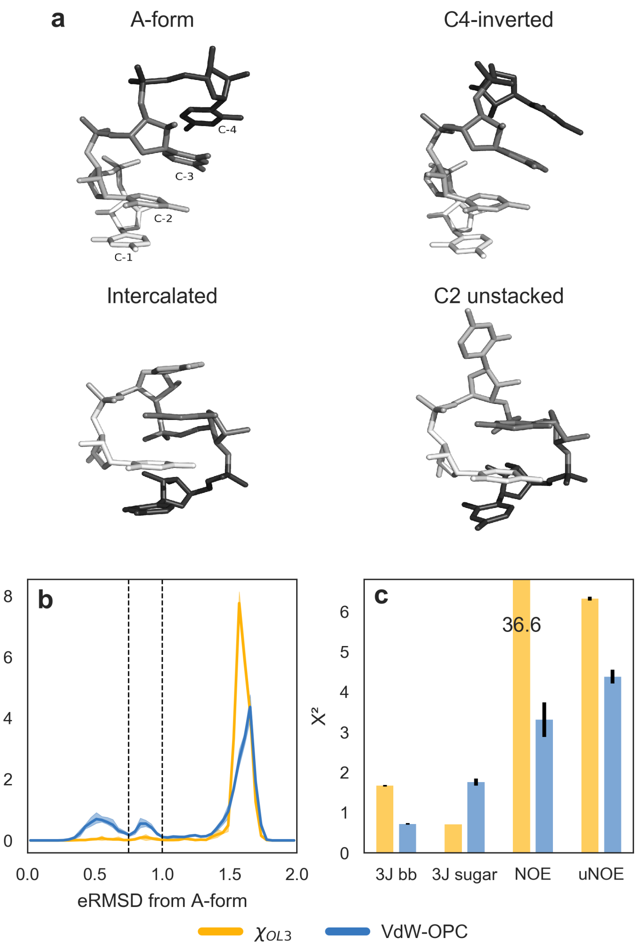

Fig. 1. **a**) three dimensional structures of r(CCCC) discussed in the main text. **b**) eRMSD from A-form histogram for $\chi_{\mathrm{OL3}}$and χ_{OL3}-VdW-OPC simulations. Solid lines indicate the average calculated using a blocking procedure, while the area between minimum and maximum is shown in 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 χ^2 statistic for backbone scalar couplings (3J bb), sugar scalar couplings, NOE and unobserved NOE (uNOE). Error bars in black show the standard error of the mean.

tetranucleotides. Except for the sequence, no other prior structural knowledge of the systems is used in simulations. We show substantial disagreement between predicted and experimental NMR data, even when using recent force-field parameters. We therefore employ the MaxEnt/Bayesian approach to refine the simulated ensembles so as to match a set of available NMR experimental data, including NOE intensities and scalar couplings. Analysis of the optimal ensembles shows that r(CCCC) and r(GACC) are mostly – but not exclusively – in A-form-like conformations. r(AAAA) and r(UUUU) display a higher complexity, as the optimal ensembles consist of a mixture of A-form with other conformationally heterogeneous structures.

2. Results

A. Agreement between experiments and simulations. We first consider the tetranucleotide with sequence CCCC. 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

Please provide details of author contributions here.

Please declare any conflict of interest here.

[2] To whom correspondence should be addressed. E-mail: sandro.bottaro@bio.ku.dk, lindorff@bio.ku.dk

at position 4 (C4) inverted (9) (see Fig. 1a). Extensive MD simulations with the standard AMBER force field (χ_{OL3} described in the Methods section) showed the presence of highly populated intercalated structures in which C1 is interposed between C3 and C4 (10, 12), while C2 is either stacked on C3 or solvent exposed. The lack of A-form-like structures is confirmed in our χ_{OL3} simulations, as shown in the eRMSD histogram from 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 (13). It has recently been reported (14) that corrections to oxygen van der Waals radii (15) in conjunction with the OPC water model (16) (here called χ_{OL3}-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 χ_{OL3}-VdW-OPC force field (Fig. 1b, blue line), we observe a small, yet significant population of A-form like structures (eRMSD<0.75) as well as C4-inverted conformations (0.75-1.0 eRMSD from A-form).

The higher accuracy of χ_{OL3}-VdW-OPC with respect to χ_{OL3} is further confirmed by the improved agreement between calculated and experimental data. Fig. 1c reports the χ^2 for backbone ^{3}J scalar couplings (H3-P, H5'/H5"-P, H4-H5'/H5"), sugar ^{3}J couplings (H1'-H2', H2'-H3', H3'-H4') and NOE intensities (9, 10). Additionally, we consider the absence of specific peaks in the NOESY spectra as a source of information. On the basis of assigned chemical shifts, NMR spectra were inspected for the presence of NOE cross-peaks between every pair of non-exchangeable protons in the tetramers. To assign unobserved NOEs (uNOE), the maximum NMR observable distance was estimated for each potential NOE from the minimum detectable cross-peak volume (see Methods). Whenever simulations predict a shorter distance between such proton-pairs, it is considered a violation of a uNOE. Note that the importance of unobserved NOE have been discussed for protein systems as well (17). Unobserved NOEs are of particular importance because several violations are present in intercalated structures (10). It can be clearly seen in Fig. 1c that the χ_{OL3}-VdW-OPC force field provides a better agreement with experimental data, especially for NOEs. We note, however, the higher χ^2 for ^{3}J sugar scalar couplings with respect to the standard χ_{OL3} force field.

B. Reweighting procedure. 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 (4, 6). In MaxEnt approaches one seeks the minimal perturbation of the simulated ensemble (i.e. the prior distribution) that satisfies a set of known experimental averages. This can be achieved (2, 6) by minimizing the function

$$\Gamma = \log(\mathrm{Z}(\lambda)) + \sum_i^m \lambda_i F_i^{\text{EXP}} + \frac{1}{2}\sum_i^m \lambda_i^2 \sigma_i^2 \qquad [1]$$

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^{EXP} with associated normally distributed and uncorrelated errors σ_i. Z is the partition function $\mathrm{Z}(\lambda) = \sum_j^N w_j^0 \exp\left[-\sum_i^m \lambda_i F_i(\mathbf{x}_j)\right]$ where $F_i(\mathbf{x}_j)$ is the function used to back-calculate the experimental observable from the atomic coordinates $\mathbf{x}$, and $\{w_1^0 \ldots w_N^0\}$ correspond to the weights of the N frames in the prior distribution. Note that this approach is completely equivalent to a Bayesian ensemble refinement approach (4, 18) in which one seeks the optimal weights $\{w_1 \ldots w_N\}$ minimizing the log posterior L

$$L(w_1 \ldots w_N) = \frac{m}{2}\chi^2 + \theta \mathrm{S}_{\text{REL}} \qquad [2]$$

where $\chi^2 = \sum_i^m \left(\sum_j^N w_j F_i(x_j) - F_i^{\text{EXP}}\right)^2 / m\sigma_i^2$ is the deviation from the experimental averages, and the relative entropy $S_{\text{REL}} = \sum_j^N w_j \log w_j / w_j^0$ quantifies the deviation from the prior distribution. θ is a parameter that sets the relative weight between these two quantities, and needs to be chosen e.g. via L-curve selection.

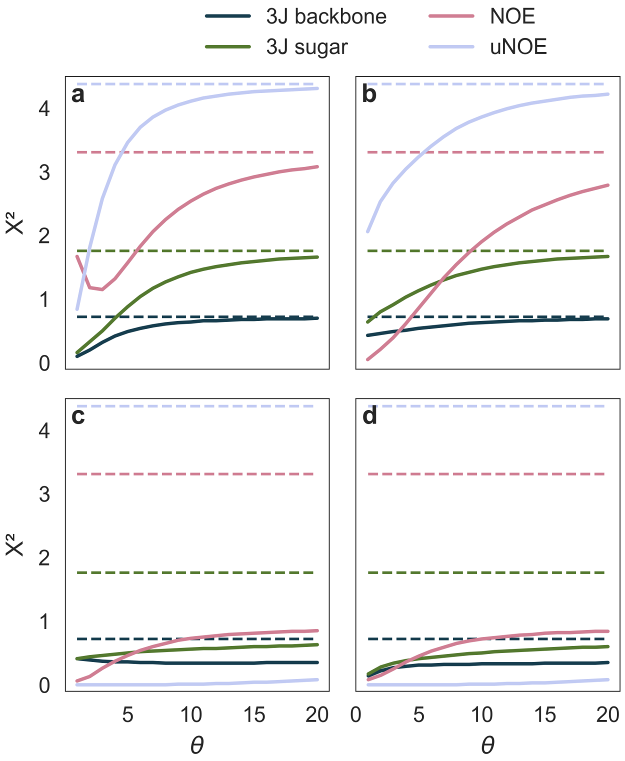

Fig. 2. Agreement between reweighted r(CCCC) simulations and experiments as a function of the parameter θ. χ^2 for different values of θ are reported when using **a**) scalar couplings only, **b**) NOE distances and **c**) uNOE distances. Results using all three types of data are shown in panel **d**. Initial, unreweighted χ^2 are shown as dashed lines.

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, θ enters the MaxEnt formulation (Eq.1) as a global scaling factor of all Gaussian errors σ_i. Third, heterogeneous data (NOE, ^{3}J couplings, chemical shifts, etc.) can be used simultaneously in the reweighting procedure, both as averages as well as inequality constraints (6).

C. 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, θ. Given the better initial agreement with experimental data, we here consider the χ_{OL3}-VdW-OPC simulations. Figure 2a shows χ^2 as a function of θ when using scalar couplings as the only input for reweighting. As expected, small θ corresponds to a better fit, while in the limit of large θ we approach the original, unreweighted χ^2 value (dashed line). We can also monitor the behavior of χ^2 relative to data that were not used in the reweighting (Fig. 2a). 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 χ^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 ^{3}J 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) as well as for the other tetranucleotides (Figs. S1-S3).

When considering χ^2 alone one would choose a small θ, 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 physico-chemical information contained in the force field is lost (Eq. 2). Additionally, this has a detrimental effect on the

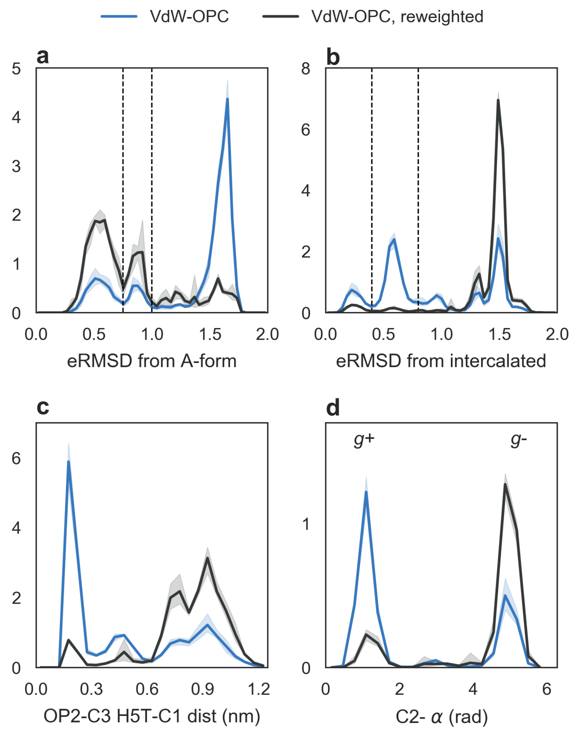

Fig. 3. Distribution of different observables before and after reweighting r(CCCC) simulations using χ_{OL3}-VdW-OPC. Solid lines indicate the average calculated using a blocking procedure, minima and maxima are shown in shade. **a**) eRMSD from ideal A-form, **b**) eRMSD from an intercalated conformation, **c**) distance distribution between OP2 in C3 and H5T in C1, and **d**) the α torsion angle of C2. Peaks in panels a-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 0.75-1.0), intercalated (eRMSD from intercalated $<$0.4), and intercalated with C2 unstacked (eRMSD from intercalated 0.4-0.8). eRMSD boundaries are shown as dashed lines.

statistical errors, as the number of effective frames contributing to the ensemble becomes very small (Fig. S4). In order to strike a good balance between fit and proximity to the prior distribution, we scan different values of θ until a further decrease of this parameter leads to an increase in the relative entropy without substantially improving the fit (4). While this procedure does not provide a unique θ, 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. Scatter plots comparing individual experimental averages against simulations before/after reweighting are shown in Fig. S5-S8.

D. Conformational ensemble of r(CCCC). The set of optimized weights can be now used to calculate the full probability distribution of any observable (e.g. distances, torsion angles, etc.). In order 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 χ_{OL3}-VdW-OPC MD ensemble consists of $\approx$18% A-form structures (eRMSD from A-form < 0.75), and 9% of structures with C4 either inverted or unstacked (eRMSD from A-form in the 0.75-1.0 range). From the histogram of eRMSD relative to intercalated structure (Fig.3b), the initial ensemble estimates a 53% population of intercalated structures, that can be subdivided into fully stacked intercalation (13%, eRMSD < 0.4) and intercalated structures with C2 unstacked ($\approx$ 40%, eRMSD in the 0.4-0.8 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, as it is consistent with the picture proposed in the original experimental paper (9). The ensemble obtained here, however, did not require expert interpretation of the individual NOE distances. More importantly, 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. 3c and d. Panel c 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 dramatically affects the distribution of α angle in C2, as we find that *gauche*⁻ (g⁻) is the preferred rotameric state in the reweighted ensemble (Fig. 3d). A similar behavior is observed for α in C3, ζ in C2 and in C3 (Fig. S10), in accordance with previous simulation studies that have shown the importance of these two torsion angles in tetranucleotides and tetraloops simulations (19, 20). We highlight that the backbone ^{3}J scalar couplings used in the reweighting procedure report on ϵ and γ angles, but not on α/ζ.

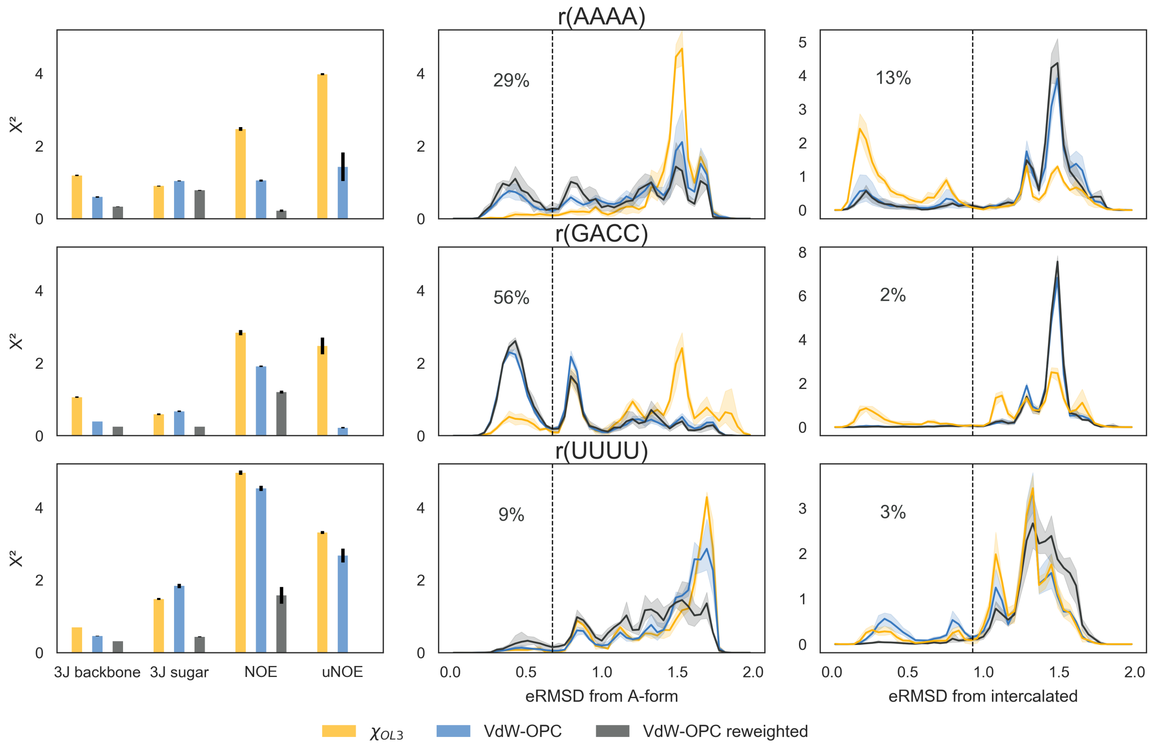

Fig. 4. Comparison between reweighted and unreweighted ensembles for r(AAAA), r(GACC) and r(UUUU) tetranucleotides. Left panels: agreement between calculated and experimental averages for χ_{OL3}, χ_{OL3}-VdW-OPC, and reweighted χ_{OL3}-VdW-OPC simulations. Central panels: histogram of the eRMSD from ideal A-form. Right panels: histogram of eRMSD from intercalated. The dashed lines indicate the tresholds used for calculating the percentage of A-form-like (middle) or intercalated structures (right) upon reweighing.

E. 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, χ_{OL3}-VdW-OPC is considerably better compared to χ_{OL3} force field (Fig.4, left panels). The reweighting procedure further improves agreement with experimental data. However, we do observe a residual discrepancy in some cases ($\chi^2 > 1$), that stems from predicted NOE distances falling outside the experimental range (Figs. S5-S8). In the case of r(GACC), three NOEs reported in the original experimental work (10) 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 datapoints that are inconsistent with the others and, as such, might require manual inspection. These cases can be treated by using error models suitable to describe outliers (6, 21).

The r(AAAA) ensemble is composed by ≈30% A-form like structures and 16% A4-inverted/unstacked (Fig.4, central panel). In this case, the available experimental data could not rule out completely the

Table 1. Percentage of C3'-endo ($\delta < 115°$) and anti ($\chi > 120°$) of reweighted χ_{OL3}-VdW-OPC simulations. The statistical error calculated using block averaging is below 1%.

	Sequence	N1	N2	N3	N4
% C3'-endo	AAAA	70.6	75.1	84.5	66.3
	CCCC	90.7	88.9	88.5	71.7
	GACC	86.9	87.1	88.3	71.1
	UUUU	61.4	49.5	50.5	63.2
% χ Anti	AAAA	65.2	96.5	98.4	97.8
	CCCC	98.0	98.5	99.8	99.7
	GACC	89.2	99.9	98.9	99.4
	UUUU	88.5	97.1	96.8	96.3

presence of intercalated structures, which represent the 13% of the optimized ensemble (Fig.4, right panel). The remaining 40% is composed of other structures that exhibit one or more sugar puckers in C2'-endo and/or the A1-χ angle in *syn* conformation (Table 1 and Fig. S9).

r(GACC) behaves very similarly to r(CCCC), with $\approx$ 60%A-form-like structures and 20% of C4-inverted/unstacked. The similarity between r(GACC) and r(CCCC) can also be appreciated by considering the sugar pucker and χ angle preferences reported in Table 1 and Figs. S10-S11. Intercalation is almost completely absent in the reweighted 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 relatively 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. While the geometry of the ideal A-form can be unambiguously defined (22), the intercalated structures are obtained by performing a cluster analysis of the χ_{OL3} simulation as described previously (23). Although this choice has a degree of arbitrariness, we found it as a useful and intuitive manner to define an order parameter complementary to the distance from A-form. As for θ, we verified that the population of the different states do not depend critically on this parameter in the relevant range $2 < \theta < 5$ (Fig. S13).

3. Discussion

In this paper we have described the structural ensembles of four RNA tetranucleotides at the atomistic level. The characterization of these systems represents a first step in understanding the ensembles and internal dynamics of larger oligonucleotides and other RNA molecules undergoing significant conformational changes.

Due to their conformational heterogeneity, RNA oligonucleotides represent prototypical cases in which NMR experimental data need to be interpreted as ensemble averages. As such, standard procedures for NMR structure determination cannot be easily applied (24). Additionally, it is not possible to predict the properties of these systems using simulations alone, because of known force-field inaccuracies (Fig.1). Only the combination of experiment with computation 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.

In a broad sense, this can be seen as a regularization problem in which a small set of experimental data are used to gain insights into a highly dimensional, complex set of molecular conformations. The problem is under-determined, 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 formulation in Eq.2 (4, 18). The balance between fit quality (χ^2) and deviation from the prior distribution (S_{REL}) is tuned by a system-dependent confidence parameter, θ, that is not known a priori. The approach used here takes explicitly into account the uncertainty σ on the experimental average. Since θ is a global scaling factor, the values of σ allow the relative weight of different heterogeneous data to be accounted for. Note that the calculation of the experimental observable from the atomic coordinates (i.e. the forward model) introduces inaccuracies that can be larger than the experimental uncertainty. For example, ^{3}J scalar couplings calculated using Karplus relationships can introduce errors up to 2 Hz (Fig.S14). Care should be also taken when calculating NOE intensities from proton-proton distances, as 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 (25).

In a number of recent MaxEnt-inspired approaches a bias deriving from the experimental data is estimated on-the-fly during the simulations (4, 6, 21, 26). 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 existing simulations whenever new experimental data are available (27). Since reweighting only requires a cheap post-processing of existing trajectories, it is straightforward to perform multiple cross validation tests. Additionally, reweighting is very convenient when the forward model calculation is particularly demanding, since in biased methods the back-calculation of averages from structures has to be performed at least every few time steps (28).

In our work the reweighting approach is also used as a tool to help identify inaccuracies in molecular dynamics force fields. Modern atomistic molecular mechanics force fields consist of hundreds of parameters, and even finding the relevant interactions that can potentially improve their accuracy is a time consuming and non-trivial task. The reweighting substantially simplifies this search (Fig. 3c-d), as the probability distribution over any degree of freedom before and after reweighting can be readily compared. We find that hydrogen bonds to non-bridging oxygens are significantly destabilized upon reweighting, in accordance with previous simulation studies (10, 29). At the same time, the population of α 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).

Materials and Methods

MD simulations. We have performed MD simulations on r(AAAA), r(CCCC), r(UUUU), and r(GACC) tetranucleotides. Each system was simulated with two different force-fields: i) the AMBER 99 force field (31) with parmbsc0 corrections to α/γ (32) and the χOL corrections to χ torsion angles (33) in TIP3P water (34). We refer to this combination as χ_{OL3}. These simulations were taken from our previous studies (19, 35). ii) $\chi_{\mathrm{OL3}}$with corrections to Van der Waals oxygen radii (15) and using the optimal 3-charge, 4 point (OPC) water model (16). We refer to this combination as χ_{OL3}-VdW-OPC. Parameters are available at http://github.com/srnas/ff. Molecular dynamics simulations were performed using the 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 stochastic velocity rescaling thermostat (38). All bonds were constrained with the LINCS algorithm (39), equations of motion were integrated with a time step of 2 fs. Tetranucleotides were simulated using temperature replica exchange (40) using 24 replicas in the temperature range 278 K-400 K for 1.0 μs per replica. All the analyses presented here were performed for the 300K replica and using 20000 frames. Averages and standard errors of the mean are calculated using four blocks of 5000 samples each.

NMR data. Experimental NOE and scalar couplings have previously been published (9, 10). We use a Gaussian-distributed experimental errors of 1.5Hz for scalar couplings (Fig. S14) and of 0.1Å for unobserved NOE. The error for NOE was estimated as $\min(r^{\mathrm{EXP}}_{\mathrm{max}} - r^{\mathrm{EXP}}, r^{\mathrm{EXP}} - r^{\mathrm{EXP}}_{\mathrm{min}})$. The number of experimental averages for each NMR parameter and for each tetranucleotide sequence is reported in Table 2. The complete list of experimental data is available as textfiles at https://github.com/sbottaro/tetranucleotides_data. NOE intensities from simulations are calculated as averages over the N samples $\mathrm{NOE}_{\mathrm{CALC}} = (\sum_i^N w_i r_i^{-6})$. ^{3}J scalar couplings are calculated using the Karplus relationships as described in Fig. S14 using the software baRNAba https://github.com/srnas/barnaba.

Table 2. Number of experimental averages.

	NOE	^3J Sugar	^3J Backbone	uNOE
AAAA	36	11	17	243
CCCC	27	11	15	245
GACC	20	12	17	284
UUUU	9	10	15	282

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 is observed, then the potential contact is classified as an unobserved NOE. If the spectral position of a potential cross-peak does not overlap any other observed cross-peak, then the minimum detectable cross-peak volume is assumed to be two times the standard deviation of spectral noise, V_{err}. Scalar coupling results in NOE cross-peaks that are split into multiplets of 2, 4, 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 $2MV_{\mathrm{err}}$. V_{err} and a scaling factor, c, obtained in the original work (9, 10) from NOESY spectra with 200 msec mixing time, are used to associate a distance,R, with the minimum detectable volume: $R = (c/2MV_{\mathrm{err}})^{1/6}$. The analysis of unobserved NOEs was carried out here with 800 msec NOESY spectra where cross-peaks are typically 2.5 to 3-fold greater than at 200 msec, so the minimum detectable NOE volume was reduced by a factor of 2.5 (after correcting for any difference in number of NMR scans). If the spectral position of a potential cross-peak partially overlaps one or more observed cross-peaks, then the minimum detectable volume of the potential cross-peak is 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 is medium or weak, respectively, then a potential cross-peak exhibiting no apparent intensity is classified as unobserved with a volume that corresponds to an internuclear distance of greater than 3.3 or 4.0 A. If the overlapping observed cross-peak is strong or the potential cross-peak is close to the diagonal, then the potential cross-peak is not classified as unobserved.

ACKNOWLEDGMENTS. The research is funded by a grant from The Velux Foundations (SB, KLL) and a Hallas-Møller Stipend from the Novo Nordisk Foundation (KLL). GB has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant agreement no. 306662, S-RNA-S. DHT was supported by NIH grant R01 GM22939. We thank Simon Olsson for helpful comments on the manuscript.

1. Jaynes ET (1978) Where do we stand on maximum entropy. *The maximum entropy formalism* pp. 15–118.
2. Pitera JW, Chodera JD (2012) On the use of experimental observations to bias simulated ensembles. *J Chem Theor Comput* 8(10):3445–3451.
3. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K (2014) Combining experiments and simulations using the maximum entropy principle. *PLoS Comp Biol* 10(2):e1003406.
4. Hummer G, Köfinger J (2015) Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys* 143(24):12B634_1.
5. Bonomi M, Heller GT, Camilloni C, Vendruscolo M (2017) Principles of protein structural ensemble determination. *Curr Opin Struct Biol* 42:106–116.
6. Cesari A, Gil-Ley A, Bussi G (2016) Combining simulations and solution experiments as a paradigm for RNA force field refinement. *J Chem Theor Comput* 12(12):6192–6200.
7. Borkar AN, et al. (2016) Structure of a low-population binding intermediate in protein-RNA recognition. *Proc Nat Acad Sci* 113(26):7171–7176.
8. Yildirim I, Stern HA, Tubbs JD, Kennedy SD, Turner DH (2011) Benchmarking AMBER force fields for RNA: comparisons to NMR spectra for single-stranded r(GACC) are improved by revised χ torsions. *J Phys Chem B* 115(29):9261–9270.
9. Tubbs JD, et al. (2013) The nuclear magnetic resonance of CCCC RNA reveals a right-handed helix, and revised parameters for AMBER force field torsions improve structural predictions from molecular dynamics. *Biochem* 52(6):996–1010.
10. Condon DE, et al. (2015) Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J Chem Theor Comput* 11(6):2729–2742.
11. Bergonzo C, et al. (2013) Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J Chem Theor Comput* 10(1):492–499.
12. Bergonzo C, Henriksen NM, Roe DR, Cheatham TE (2015) Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* 21(9):1578–1590.
13. Bottaro S, Di Palma F, Bussi G (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res* 42(21):13306–13314.
14. Bergonzo C, Cheatham III TE (2015) Improved force field parameters lead to a better description of RNA structure. *J Chem Theor Comput* 11(9):3969–3972.
15. Steinbrecher T, Latzer J, Case D (2012) Revised AMBER parameters for bioorganic phosphates. *J Chem Theor Comput* 8(11):4405–4412.
16. Izadi S, Anandakrishnan R, Onufriev AV (2014) Building water models: a different approach. *J Phys Chem Lett* 5(21):3863–3871.
17. Zagrovic B, Van Gunsteren WF (2006) Comparing atomistic simulation data with the NMR experiment: how much can NOEs actually tell us? *Proteins* 63(1):210–218.
18. Różycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19(1):109–116.
19. Gil-Ley A, Bottaro S, Bussi G (2016) Empirical corrections to the amber RNA force field with target metadynamics. *J Chem Theor Comput* 12(6):2790–2798.
20. Bottaro S, Banas P, Sponer J, Bussi G (2016) Free energy landscape of GAGA and UUCG RNA tetraloops. *J Phys Chem Lett* 7(20):4032–4038.
21. Bonomi M, Camilloni C, Cavalli A, Vendruscolo M (2016) Metainference: A bayesian inference method for heterogeneous systems. *Sci Adv* 2(1):e1501177.
22. Macke TJ, Case DA (1998) Modeling unusual nucleic acid structures. (ACS Publications).
23. Bottaro S, Lindorff-Larsen K (2017) Mapping the universe of RNA tetraloop folds. *Biophys J* 113(2):257–267.
24. Sripakdeevong P, et al. (2014) Structure determination of noncanonical RNA motifs guided by 1h NMR chemical shifts. *Nature Methods* 11(4):413–416.
25. Tropp J (1980) Dipolar relaxation and nuclear overhauser effects in nonrigid molecules: the effect of fluctuating internuclear distances. *J Chem Phys* 72(11):6035–6043.
26. White AD, Voth GA (2014) Efficient and minimal method to bias molecular simulations with experimental data. *J Chem Theor Comput* 10(8):3023–3030.
27. Olsson S, Strotz D, Vögeli B, Riek R, Cavalli A (2016) The dynamic basis for signal propagation in human pin1-WW. *Structure* 24(9):1464–1475.
28. Ferrarotti MJ, Bottaro S, Pérez-Villa A, Bussi G (2014) Accurate multiple time step in biased molecular simulations. *J Chem Theory Comput* 11(1):139–146.
29. Yang C, Lim M, Kim E, Pak Y (2017) Predicting RNA structures via a simple van der Waals correction to an all-atom force field. *J Chem Theory Comput* 13(2):395–399.
30. Sponer J, et al. (2018) RNA structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem Rev.*
31. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* 21(12):1049–1074.

32. Pérez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: Improving the description of α γ conformers. *Biophys J* 92(11):3817–3829.
33. Zgarbová M, et al. (2011) Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J Chem Theor Comput* 7(9):2886–2902.
34. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.
35. Bottaro S, Gil-Ley A, Bussi G (2016) RNA folding pathways in stop-motion. *Nucleic Acids Res* 44(12):5883–5891.
36. Pronk S, et al. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29(7):845–854.
37. Joung IS, Cheatham III TE (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* 112(30):9020–9041.
38. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126(1):014101.
39. Hess B, Bekker H, Berendsen HJ, Fraaije JG (1997) LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 18(12):1463–1472.
40. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314(1):141–151.