Unexpected Dynamics in the UUCG RNA Tetraloop

Sandro Bottaro[1,2*], **Parker J. Nichols**[3], **Beat Vögeli**[3], **Michele Parrinello**[1], **Kresten Lindorff-Larsen**[2*]

[1]Atomistic Simulations Laboratory, Istituto Italiano di Tecnologia, Genova, Italy; [2]Structural Biology and NMR Laboratory, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [3]Department of Biochemistry and Molecular Genetics, University of Colorado Denver, Anschutz Medical Campus, Aurora, Colorado, USA.

***For correspondence:**
sandro.bottaro@iit.it (SB);
lindorff@bio.ku.dk (KLL)

Abstract Many RNA molecules are dynamic, but characterizing their motions by experiments is difficult, often requiring application of complex NMR experiments. Computational methods such as molecular dynamics simulations, on the other hand, still suffer from difficulties in sampling and remaining force field errors. Here, we provide an atomic-level description of structure and dynamics of the 14-mer UUCG RNA stem-loop by combining molecular dynamics simulations with exact nuclear Overhauser enhancement data. The integration of experiments and simulation via a Bayesian/Maximum entropy approach enables us to discover and characterize a new state of this molecule, which we show samples two distinct states. The most stable conformation corresponds to the native, consensus three-dimensional structure. The second, minor state has a population of 11%, and is characterized by the absence of the peculiar non-Watson-Crick base pair between U and G in the loop region. By using machine learning techniques, we identify key contacts in the NOESY spectrum that are compatible with the presence of the low-populated state. Together, our results demonstrate the validity of our integrative approach to determine the structure and thermodynamics of conformational changes in RNA molecules.

INTRODUCTION

RNA loops are structural elements that cap A-form double helices, and as such are fundamental structural units in RNA molecules. The great majority of known RNA loops contain four nucleotides [1], and these so-called tetraloops are one of the most common and well-studied RNA three-dimensional motifs [2]. The great majority of known RNA tetraloops have the sequence GNRA or UNCG, where N

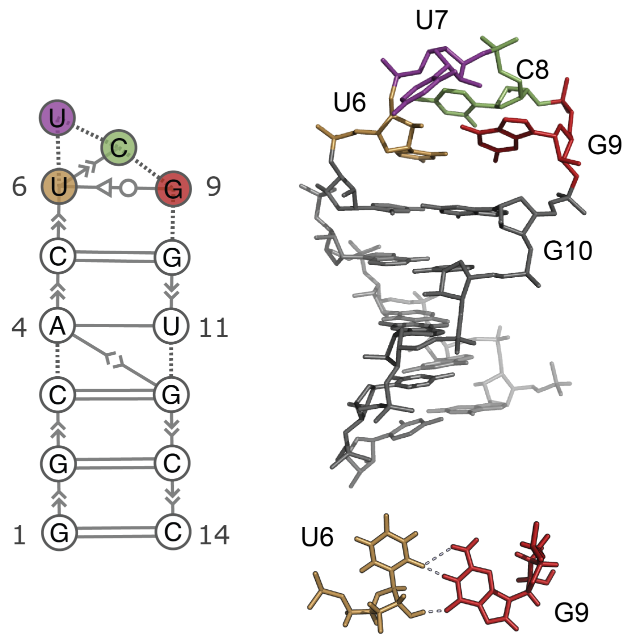

Figure 1. Consensus secondary structure (left) and three dimensional structure (right) of the UUCG tetraloop [6]. The stem is formed by 5 consecutive Watson-Crick base-pairs capped by the loop U6-U7-C8-G9. One of the most distinctive feature of this structure is the trans-Sugar-Watson interaction between U6 and G9 (bottom). Extended secondary structure annotation follows the Leontis-Westhof nomenclature [8]

is any nucleotide and R is guanine or adenine. Their small size, together with their biological relevance, has made these systems primary targets for nuclear magnetic resonance (NMR) spectroscopy, X-ray-crystallography, and atomistic molecular dynamics (MD) simulation studies [3, 4, 2].

The UUCG tetraloop has been long known to be highly stable, and both crystallographic and NMR studies suggest that this tetraloop adopts a well-defined three dimensional structure [5, 6] (Fig. 1). Experimentally, the UUCG tetraloop is used to stabilize the secondary structure of larger RNA molecules without interacting with other RNAs or proteins [7].

Despite its stability, the UUCG tetraloop is not rigid. In particular, three recent studies by independent groups indicate the presence of alternative loop conformations [9, 10, 11]. Earlier NMR studies [6, 12] also suggested the presence of loop dynamics, without providing a detailed structural interpretation of the data. More generally, the atomic-detailed characterization of RNA structure and dynamics requires specialized techniques and substantial experimental effort, including NMR measurements of nuclear Overhauser effects (NOE), scalar couplings, chemical shifts, residual dipolar couplings, cross-correlated relaxation rates as well as a wide range of relaxation-dispersion type NMR experiments [13, 14].

While NOEs are typically used to determine RNA and protein structures, they also contain dynamic information. Because ensemble-averaged NOEs are highly sensitive to the underlying distance fluctuations, they may contain contributions even from minor populations. Normally, such information is difficult to extract because standard NOE measurements are relatively inaccurate. It has, however, been demonstrated that a substantial part of the information content inherent to these probes can be obtained from exact NOE measurements (eNOEs) [11]. As opposed to conventional NOEs, eNOEs can be converted into tight upper and lower distance limit restraints [15, 16, 17].

Previous computational studies of the UUCG tetraloops focused either on the dynamics around the near-native state [18] or on the difficulty in separating force-field inaccuracies from insufficient sampling [19, 20]. In a previous study we reported converged free-energy landscape for RNA 8-mer and 6-mer loops, and we have shown that native-like states are not the global free-energy minimum using the current AMBER RNA force-field [21]. This problem has been addressed in a new parameterization of the AMBER force-field, that improves the description of the UUCG 14-mer and other RNA systems [22]. Nevertheless, it remains difficult to assess the accuracy of these simulations, because experiments alone do not provide an atomic-detailed description of structure and dynamics that serve as a benchmark.

Here, we use extensive atomistic MD simulations to map the conformational landscape of the UUCG tetraloop using enhanced sampling techniques and a recent force-field parameterization. To further improve the description of this system, we perform an a posteriori refinement of the MD simulation using eNOE data via a Bayesian/maximum entropy procedure [23, 24]. By construction, the refined ensemble shows better agreement with eNOE relative to the original MD simulation. We validate the eNOE-refined ensemble against independent NMR measurements, and we find an agreement that is on average comparable with NMR structures of the UUCG tetraloop deposited in the Protein Data Bank (PDB).

Our experimentally-refined ensemble reveals the presence of two conformational states. The dominant, major state (here called state A) is the consensus UUCG structure shown in Fig. 1. The second, previously unreported lowly-populated state (state B) is characterized by the absence of the signature U6-G9 non-Watson-Crick base pair, with the G9 base exposed into solution. The salient features of state B are identified using a technique adapted from the field of machine learning called harmonic linear discriminant analysis (HLDA) [25]. Among all possible proton-proton distances, we identify specific contacts between C8 and G10 that are present in state B but not in state A. We inspect the NOESY spectrum for such contacts in order to provide independent evidence for the presence of the low-populated state.

The paper is organized as follows: we first compare the predictions obtained from MD simulation against different experimental datasets. We then discuss the effect of the refinement procedure, showing how it improves the agreement with experiments and how it affects the population of different conformations. We proceed by identifying the relevant degrees of freedom and contacts that characterize the two states. Finally, we identify peaks in the NOESY spectrum corresponding to contacts that are present in state B but not in state A. We accompany this paper with the commented code, in form of Jupyter notebooks, to reproduce step-by-step the complete analysis, including all figures and supplementary results presented in the manuscript.

Results

MD simulations and comparison with experimental data

We simulate the RNA 14-mer with sequence GGCACUUCGGUGCC starting from a completely extended conformation. Studying the folding free-energy landscape of this system is computationally expensive: for this reason previous attempts required μs-long simulations in combination with tempering protocols [22, 26, 27].

Here, we combine two enhanced sampling techniques: solute tempering in the REST2 formulation [28] and well-tempered metadynamics [29]. We used a nucleic-acid specific metric, called eRMSD, [30] as a collective variable for enhanced sampling. The MD simulation setup and convergence analysis are presented in supporting information 1 (SI1).

Before describing the conformational ensemble provided by MD, we compare the computational prediction with available NMR spectroscopy data. More precisely, we consider the following experimental datasets:

- **Dataset A**. Exact eNOEs [11], consisting in 62 bidirectional exact NOE, 177 unidirectional eNOE and 77 generic normalized eNOE (gn-eNOE). This dataset alone was used to determine the structure of the UUCG tetraloop with PDB accession codes 6BY4 and 6BY5. In addition to the original dataset, we added 1 new eNOE and 6 new gn-eNOEs, as described in SI2.
- **Dataset B**. 97 ^{3}J scalar couplings, 31 RDCs and 250 NOE distances. This data, among other NMR measurements, was used to calculate the consensus UUCG tetraloop structure (PDB 2KOC [6]).
- **Dataset C**. 38 (RDC1) plus 13 (RDC2) residual dipolar couplings. These RDCs have been used in conjunction with MD simulations to obtain a dynamic ensemble of the UUCG tetraloop. [9].
- **Dataset D**. 91 solvent paramagnetic resonance enhancement (sPRE) measurements [10].

The grey bars in Fig. 2 show the agreement between simulation and the different experimental datasets. The agreement with NOE and ^{3}J scalar couplings is expressed using the reduced χ^2 statistics, defined as the average square difference between the experimental measurement (F^{exp}) and the back-calculated ensemble average ($< F(\mathbf{x}) >$) normalized by the experimental error σ:

$$\chi^2 = \frac{1}{m}\sum_i^m \frac{(< F(\mathbf{x}) >_i - F_i^{EXP})^2}{\sigma_i^2} \tag{1}$$

Hence, the lower the χ^2, the better the agreement. As a rule of thumb, $\chi^2 < 1$ can be considered small, as the difference between experiment and prediction is within experimental error. For RDC and sPRE we calculate the the Spearman correlation coefficient (ρ), that approaches the value of 1 when experimental measurement and computational prediction are perfectly correlated. See SI2 for additional details on this comparison.

As a reference, we report in Fig. 2 the agreement calculated on the PDB ensembles 6BY5 [11] and 2KOC [6]. For bidirectional eNOE and gn-eNOE, the agreement of the MD with experiment is considerably poorer than the one calculated on 6BY5. We recall that this latter ensemble was determined by fitting dataset A, we thus expect χ^2 to be small in this case. On datasets B, C, and D, all different ensembles behave similarly. When considering other statistics (e.g. root mean square error, Pearson correlation, number of violations), the same conclusions apply. Note that χ^2 for ^{3}J couplings is large in all cases. This discrepancy may arise both from the imperfect ensembles as well as from the limitation of the function used to calculate the experimental quantity from the atomic positions (i.e. the forward model). As an example, the parameters in the Karplus equation for HCOP couplings critically depend on a single experimental data point measured in 1969 [31].

Bayesian/Maximum entropy refinement of the MD ensemble

As described above, our MD simulation provide a conformational ensemble consisting of a rich and diverse set of conformations, that, however, do not match all experimental data perfectly, especially when considering dataset A. On the other hand, the 6BY5 ensemble matches the eNOE data remarkably well, but may underestimate the dynamics of the tetraloop.

In order to improve the description provided by the MD simulation, we calculate a refined conformational ensemble by a posteriori including experimental information into simulations. In brief, the refinement is obtained by assigning a new weight to each MD snapshot, in such a way that the averages calculated with these new weights match a set of input (or “training”) experimental data within a given error. Among all the possible solutions to this underdetermined problem, we use the one that maximize the Shannon cross-entropy [32, 33].

Here, we refine the simulation by using dataset A as a training set, while datasets B–D serve for cross-validation (see also SI3). By construction, the refinement procedure improves the agreement on the training data (dataset A). We choose the free hyper-parameter of the algorithm as the one that maximize the agreement on the validation datasets.

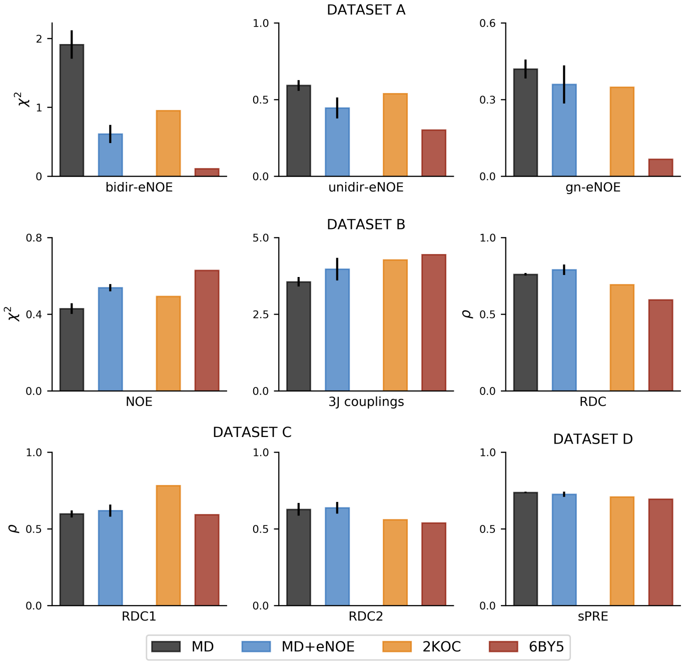

Figure 2. Comparison between experiment and conformational ensembles. We consider four ensembles (MD, MD+eNOE, 2KOC and 6BY5) on nine different experimental datasets. Agreement is expressed using χ^2 (NOE and ^{3}J scalar couplings) and by the Spearman correlation coefficient ρ (RDC and sPRE). Error bars show the standard error estimated using four blocks.

Taken together, our results show that the refined ensemble (MD+eNOE) fits all available experimental data to a degree that is comparable to the one calculated from PDB structures 2KOC and 6BY5 (Fig. 2).

Free energy landscape

In this section we analyze in detail the MD+eNOE ensemble, and discuss the differences with respect to the original simulation and previously determined structures. We consider the free energy surface projected along the distance from the consensus structure (PDB 2KOC). Distances are measured using the eRMSD, a nucleic-acid specific metric that takes into account both position and orientations between nucleobases [30]. The free energy surface projected onto the distance from the fully-formed stem (residues 1-5 and 10-14) in Fig. 3**a** shows a single global minimum around eRMSD=0.5. This indicates that in the global free-energy minimum all five Watson-Crick base-pairs are formed. As a rule-of-thumb, two

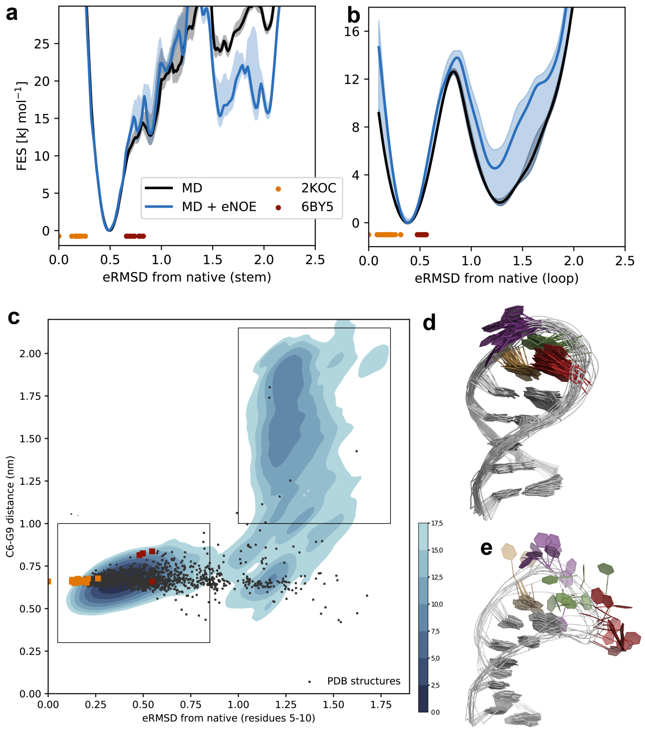

Figure 3. a) Free energy surface projected on to the eRMSD from native stem (residues 1-5 and 10-14). The eRMSD from native of 2KOC and 6BY5 are indicated as dots. **b**) Free energy surfaces projected onto the loop eRMSD (residues 5-10). Shades show the standard error estimated using four blocks. **c**) Two-dimensional free energy surface of the experimentally-refined MD simulations projected onto the eRMSD from native loop and onto the distance between the center of the six-membered rings in C6 and G9. Isolines are shown every 2.5 kJ/mol. The rectangles show the regions defining state A and state B. **d**) Representative state A conformations. **e**) Representative state B conformations.

structures with eRMSD ≤ 0.7 are typically very similar one to another, and share the same base-pair and stacking patterns [30, 34].

When considering the loop region only (Fig. 3**b**) there exist two distinct minima. The global minimum on the left (state A) corresponds to the consensus loop structure. Both 2KOC and 6BY5 structures lie in the vicinity of this minimum. The other minimum is a different loop conformation (state B) in which this non-canonical base pair is not present.

The picture emerging from the combination of MD simulations and eNOE is summarized in Fig. 3c, showing the free energy landscape projected onto the distance from native and onto the C6-G9 distance. The global free energy minimum is the native state A, with all the Watson-Crick base pairs in the stem formed together with the signature trans-sugar-Watson base pair between U6 and G9 (Fig. 3d). In state A, U7 is free to fluctuate into the solvent. In state B (Fig. 3e) all Watson-Crick base-pairs are formed, but the loop presents significant differences with respect to state A: the U6-G9 interaction is lost, and G9 is flipped out (Fig. 3e). From the regions defined in Fig. 3c we estimate a population of $84 \pm 7\%$ for state A and $11 \pm 6\%$ for state B, corresponding to a free energy difference of -5.7 ± 2.9 kJ/mol.

On top of the free-energy surface, in Fig. 3c we plot the two tetraloop structures 2KOC and 6BY5. Both ensembles fall within our definition of state A. Note also that the original experimental study described the presence of two sub-states in 6BY5, that can be distinguished along the y projection. In addition, we extract from the PDB all stem-loop structures with sequence NNUUCGNN as described previously [2]. These structures, when projected on the surface in Fig. 3, are spread in different regions of the free-energy landscape. Experimentally solved tetraloops are subject to a variety of perturbations, including crystal packing, different buffer conditions or tertiary interactions. It has been shown in the case of proteins and nucleic acids that these perturbations are compatible with the equilibrium fluctuations [35, 36], and Fig.3c is consistent with this picture. Note that a handful of PDB structures with sequence UUCG fall into the state B region. While it would be tempting to use this fact to support the existence of state B, we noticed that these hits all belong to solvent-exposed regions in cryo-electron microscopy structures.

Describing state B using harmonic linear discriminant analysis.

Having discovered this new B-state, we proceed to analyse its structural features and seek for experimental validation. While the main global minimum is known and structurally well-defined, it is not trivial from a simple visual inspection to identify which are the main structural features distinguishing the two loop conformations. Here, we address this question by using the harmonic linear discriminant analysis (HLDA), a variant of the linear discriminant analysis (LDA) [37]. LDA is routinely used in the field of data science and machine learning to find a linear combination of descriptors that best separates two or more classes. This idea has been applied for analysing complex transitions in biomolecular simulations [38, 39], and HLDA has successfully been used as biased collective variables to enhance sampling [25, 40, 41].

Here, we are interested in finding the most relevant descriptors (degrees of freedom) that discriminate the two states. To this end, we perform HLDA considering as descriptors a cosine function of the dihedral angles $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \chi$ in the 14-mer (see Methods section).

We show in Fig. 4 the coefficients of the non-zero eigenvector. The larger in magnitude the coefficient, the more important the corresponding descriptor in the linear combination. The largest coefficients are localized in nucleotides C8 and G9, both belonging to the loop region (Fig. 1). Indeed, the distribution of the descriptor with the highest coefficient (ζ in C8) has two distinct peaks. This angle is in the gauche+ (g+) conformation in the native state, and we find the alternative loop conformation to adopt the gauche- (g-) rotameric state. The χ angle in G9 is not among the highest-ranked descriptor because it is in syn conformation both in state A and in state B.

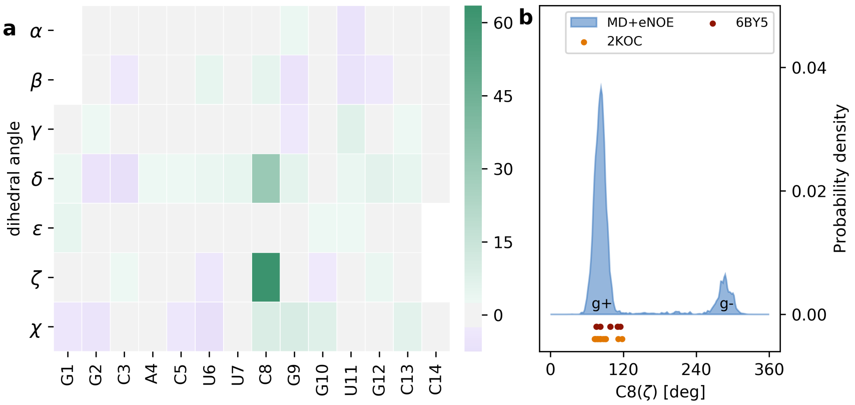

Figure 4. a) Eigenvector coefficients from HLDA using state A and B as classes and a cosine function of the torsion angles as descriptors. The larger the magnitude of the coefficient, the more relevant the angle in describing the separation between the two states. ζ in C8 is the degree of freedom with the largest coefficient. **b**) Probability distribution of the C8(ζ) calculated on the MD+eNOE ensemble, together with the values from the PDB structures 2KOC and 6BY5.

HLDA also makes it possible to address a different question: which distances that are short in state B but not in state A – and vice-versa – would be measurable by eNOEs? To this end, we consider all the H-H inter-nucleotide distances in the 14-mer whose calculated NOE-derived distance is smaller than 6Å. We obtain in this way 801 H-H distances, that we use as descriptors in HLDA. Again, the eigenvector coefficients allow us to rank the most important distances that are different in the two states. Among the highest-ranked coefficients we find several contacts between C8 and G10 that are shorter in state B compared to state A. Because NOE-derived distances are highly sensitive to distance fluctuations, in particular when measured via eNOE protocols, such B-specific contacts should be able to provide further evidence for the structure and population of the B state.

By inspecting the NOE spectrum for the presence of C8-G10 contacts, we identify several NOE that were not part of the original dataset [11], but are here included the training set used for ensemble refinement. In Fig. 5**c** we show selected NOE-derived distances, together with the predicted values from MD+eNOE and PDB ensembles. The first three NOEs are used as lower-distance bounds estimated from the spectral noise. Note that MD+eNOE average is at the limit of the boundary for C8 H5-G10 H1' and C8 H2'-G10 H8, suggesting the presence of the B state to be overestimated in our refined ensemble. The contact C8-H4' to G10-H8 is less informative, as the corresponding eNOE matches the experimental value both in A and B states. Note that the presence of the B-state is compatible with short contacts between G9 and U6, that are satisfied even if the GU base-pairs is not formed at all times (Fig. 5**c,d**). The NOE spectrum shows a peak corresponding to the C8 H1’ to G10 H8, that overlaps with G9 H1’ to G10 H8 (Fig. 5**b**). The combined signal is compatible with the distances sampled in the MD+eNOE, but incompatible with 2KOC. Note that this new eNOE is also satisfied in the 2-state ensemble 6BY5.

An additional argument supporting the presence of the B-state is provided by sPRE data. In the original paper [10], the authors measured unusually large calculated sPRE in G9-H1 and U6-H3, corresponding to a larger than expected solvent accessibility of these atoms, and observed that these values could not be explained from available PDB structures. In our MD+eNOE ensemble we observe a large G9-H1 sPRE, in agreement with experiments (see also SI4). At variance with experimental evidence, we do not predict large sPRE for U6-H3. Different reasons may contribute to this discrepancy: the lack of U6 dynamics in simulations, inaccuracies in the empirical model employed to calculate sPRE from structures, or solvent-exchange effects [42]. Conversely, on-resonance 13C R1ρ relaxation dispersion experiments on a UUCG tetraloop with a different stem sequence showed no significant exchange contributions, indicating the absence of motions with substantial chemical shift variation in the μ-ms timescale [43].

Conclusions

Based on our extensive MD simulations and integrating them with exact NOE data, we report the free energy landscape of a prototype stem-loop RNA 14-mer known as the UUCG tetraloop. The main finding of the present study is the previously unreported presence of a non-native free-energy minimum with an estimated population of 11% ± 6%. The low-populated state differs from the known structure only in the loop region, and it is characterized by the absence of the tSW base-pair between C6 and G9, with the latter nucleotide partially exposed into solution. This result has been obtained by using atomistic MD simulations and eNOE, without the need of additional data.

The free-energy surfaces and estimated population provided here are based on the available experimental data, on the employed model, and the extent of our sampling. Therefore, they are subject to inaccuracies. However, both simulations and eNOE data are consistent with the presence of the B state as described in this paper. This interpretation is qualitatively consistent with several NMR studies, that also suggested the presence of dynamics in G9 [12, 6, 10]. Note also that G9-exposed structures were reported in previous MD simulations [26, 20, 44], suggesting our finding to be robust with respect to the choice of the force-field and water model.

In order to further test our findings, it could be useful to perform dedicated

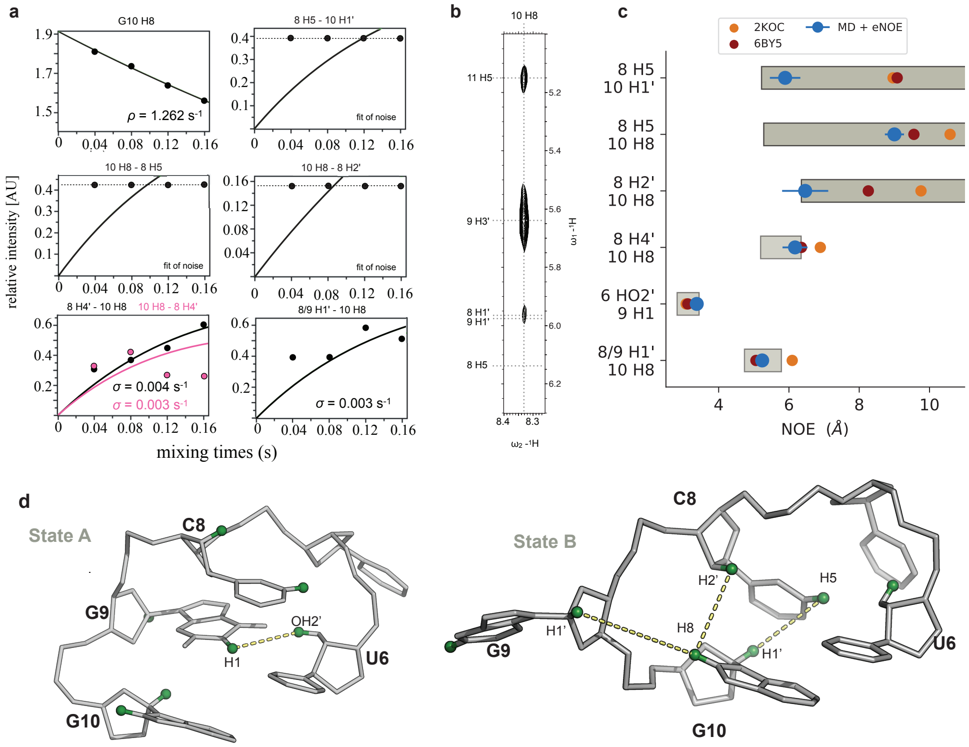

Figure 5. a) NOESY diagonal decay and cross peak buildup curves are shown for spin pairs with significantly shorter distances in the B state than in the A state. There is no visible cross peak in the spectrum corresponding to some of the proton pairs. In these cases the horizontal broken line showing the spectral noise puts an upper limit on the peak intensities. The NOESY series was recorded as described in [11]. **b**) Strip of the NOESY spectrum at maximum mixing time showing buildup cross peaks caused by the magnetization transfers to 10 H8. **c**) Calculated and experimental NOE for selected proton-pairs. The average from the MD+eNOE simulation ensemble is shown in blue, and the experimental measure is shown in gray. Red and orange shows the eNOE calculated from the 6BY5 and 2KOC ensembles. For 8 H5 - 10 H8, 8 H5 - 10 H1', 8 H2'-10 H8, the bar shows the allowed range as derived from the spectral noise. **d**) Comparison between A and B state in the loop region. Short-range contacts between C8 and G10 are possible when G9 is bulged out.

experiments probing long-timescale dynamics such as R1ρ [43] or chemical exchange saturation transfer experiments.

In this work we have used eNOEs to reweight a posteriori the ensemble generated via enhanced sampling MD simulations. This refinement procedure is a post-processing approach [23, 24] that is in principle less powerful compared to on-the-fly methods that samples directly from the target probability distribution [45]. Refinement, however, is computationally cheap, as such one can easily experiment by trying different combinations of training/cross-validation sets, and to include new data when they become available. Here we have taken advantage of this property, and we used the refined ensemble to make predictions and to suggest new experiments.

In order to identify the experimental measurement to probe the existence of state B, we resorted to a variant of the linear discriminant analysis, a method adapted from the field of machine learning. HLDA provides a concise way to interpret differences between biomolecular conformations that cannot be easily summarized in terms of a small number of collective variables [46, 47].

During the course of this study we have attempted to refine the simulation by matching RDC data (datasets B and C), but this resulted in a decreased agreement with other datasets. We have observed a similar behaviour when using sPRE (dataset D) for refinement. Instead, enforcing the agreement with eNOE (dataset A) marginally affects the agreement with other datasets (Fig. 2 and SI2). Different reasons can contribute to this behaviour. First, we do not expect all experimental data to be perfectly compatible one with the other, because measurements were conducted in similar, but not identical conditions. Second, the forward models might not be accurate for arbitrary molecular conformation. For example, if the forward model can accurately predict the RDC given the native structure, but fails on unfolded/misfolded conformations, we obtain artefacts that cannot be easily accounted for in our refinement procedure. Note that this problem is typically less relevant when using experimental RDC, sPRE or chemical shift data for scoring structures [48, 43, 10].

Finally, we note that the approach taken here is general and it is applicable to other RNA or protein systems [49, 50]. Previous characterization of slow, larger motions in RNA molecules have mostly relied on relaxation-dispersion, chemical exchange saturation transfer or related NMR experiments that probe chemical shift differences between different conformational states. We hope that the integration of MD simulations and eNOE measurements provides further opportunities for characterizing the free energy landscapes of RNA molecules.

Acknowledgements

The research is funded by a grant from The Velux Foundations (S.B. and K.L.-L.), the Lundbeck Foundation BRAINSTRUC initiative (K.L.-L.), a start-up package from the University of Colorado (B.V.), and by NSF grant 1917254 for Infrastructure Innovation for Biological Research (B.V.). The authors thank Giovanni Bussi for helpful comments. We also thank Tobias Madl, Harald Schwalbe, and coworkers for useful discussions and for sharing their data.

Methods

Integrating MD simulation and experimental data

We combine the MD simulation with experimental data using a maximum entropy/Bayesian procedure [33, 51, 23]. In our previous work, we have described this reweighting procedure as Bayesian/MaxEnt (BME) [52, 24] . In BME we use the experimental data to modify a posteriori the simulation so that the new conformational ensemble has the following properties: (i) the calculated averages are close to the experimental values taking uncertainty into account and (ii) it maximizes the relative Shannon entropy with respect to the original simulation ensemble. The modification comes in the form of a new set of weights w_j^*, one for each simulation frame.

It can be shown that this problem can be cast as a minimization problem, in which one seeks the minimum of the function Γ with respect to the set of Lagrange multipliers $\bar{\lambda} = \lambda_1 \cdots \lambda_m$, with m being the number of experimental constraints.

$$\Gamma(\bar{\lambda}) = \log(Z(\bar{\lambda})) + \sum_i^m \lambda_i F_i^{\text{exp}} + \frac{\theta}{2} \sum_i^m \lambda_i^2 \sigma_i^2 \qquad (2)$$

Here, σ_i are the uncertainties on the experimental measurements F_i^{exp} and include experimental errors and inaccuracies introduced by the calculation of the experimental quantity from the atomic positions ($F(\mathbf{x})$). θ is a free parameter, while the partition function Z is defined as

$$Z(\bar{\lambda}) = \sum_{j=1}^n w_j^0 \exp[-\sum_i^m \lambda_i F_i(\mathbf{x}_j)] \qquad (3)$$

The sum over the index j runs over the n frames in the simulation, and w_j^0 are the original weights. $w^0 = 1/n$ when using plain MD simulations or enhanced sampling techniques that sample directly from the target distribution (e.g. parallel tempering). In this paper we use WT-METAD, and the original weights w^0 are estimated using the final bias potential [53]. The minimization of Eq. 2 yields a set of Lagrange multipliers $\bar{\lambda}^*$ that are used to calculate the optimal weights

$$w_j^* = \frac{1}{Z(\bar{\lambda}^*)} w_j^0 \exp[-\sum_i^m \lambda_i^* F_i(\mathbf{x}_j)] \qquad (4)$$

In the context of the UUCG tetraloop, we use the dataset A described in the previous section to refine the simulation ensemble, and cross-validate the results against datasets B, C, and D. Details on the comparison between simulations and experiments, on the BME procedure and on the choice of the regularization parameter θ can be found in SI 2,3, and 4.

Harmonic linear discriminant analysis (HLDA)

In HLDA, the goal is to find the projection $\mathbf{W}$ that maximize the degree of separation between M classes in the N dimensional space of the descriptors [41]. The separation is measured by the ratio

$$\mathcal{J}(\mathbf{W}) = \frac{\mathbf{W}^T \mathbf{S_b} \mathbf{W}}{\mathbf{W}^T \mathbf{S_w} \mathbf{W}} \qquad (5)$$

Where the between classes $\mathbf{S_b}$ and within class $\mathbf{S_w}$ scatter matrices are defined as

$$\mathbf{S_w} = \left[\sum_i^M \Sigma_i^{-1}\right]^{-1} \qquad \mathbf{S_b} = \sum_i^M (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \tag{6}$$

Here, μ_i, Σ_i are the mean and covariance of the i^{th} class and $\bar{\mu}$ is the overall average. The maximization of $\mathcal{J}(\mathbf{W})$ under the constraint $\mathbf{W}^T\mathbf{S_w}\mathbf{W} = 1$ can be cast to an eigenvalue problem of the form

$$\mathbf{S_b}\mathbf{W} = \lambda\mathbf{S_w}\mathbf{W} \tag{7}$$

Note that there are $M-1$ non-zero eigenvalues, and in the simplest case of $M = 2$, the eigenvector corresponding to the only non-zero eigenvalue is given by

$$\mathbf{W}^* = \mathbf{S_w}^{-1}(\mu_A - \mu_B) \tag{8}$$

The magnitude and sign of the components $\mathbf{W}^* = \{W_1^* \cdots W_N^*\}$ carry information on the importance of the different descriptors. The larger the absolute value of the coefficient, the more relevant is the corresponding descriptor in discriminating the states.

References

[1] Wolters, J. The nature of preferred hairpin structures in 16s-like rrna variable regions. *Nucleic acids research* **20**, 1843–1850 (1992).

[2] Bottaro, S. & Lindorff-Larsen, K. Mapping the universe of rna tetraloop folds. *Biophys. J.* **113**, 257–267 (2017).

[3] Cheong, C., Varani, G. & Tinoco Jr, I. Solution structure of an unusually stable rna hairpin, 5GGAC (UUCG) GUCC. *Nature* **346**, 680 (1990).

[4] Woese, C., Winker, S. & Gutell, R. Architecture of ribosomal rna: constraints on the sequence of" tetra-loops". *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8467–8471 (1990).

[5] Ennifar, E. *et al.* The crystal structure of UUCG tetraloop1. *J. Mol. Biol.* **304**, 35–42 (2000).

[6] Nozinovic, S., Fürtig, B., Jonker, H. R., Richter, C. & Schwalbe, H. High-resolution nmr structure of an rna model system: the 14-mer cuucgg tetraloop hairpin rna. *Nucleic Acids Res.* **38**, 683–694 (2010).

[7] Hall, K. B. Mighty tiny. *RNA* **21**, 630–631 (2015).

[8] Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of rna base pairs. *Rna* **7**, 499–512 (2001).

[9] Borkar, A. N., Vallurupalli, P., Camilloni, C., Kay, L. E. & Vendruscolo, M. Simultaneous nmr characterisation of multiple minima in the free energy landscape of an rna uucg tetraloop. *Phys. Chem. Chem. Phys.* **19**, 2797–2804 (2017).

[10] Hartlmüller, C. *et al.* Rna structure refinement using nmr solvent accessibility data. *Sci. Rep.* **7**, 5393 (2017).

[11] Nichols, P. J. *et al.* High-resolution small rna structures from exact nuclear overhauser enhancement measurements without additional restraints. *Communications Biology* **1**, 61 (2018).

[12] Duchardt, E. & Schwalbe, H. Residue specific ribose and nucleobase dynamics of the cuucgg rna tetraloop motif by mnmr 13 c relaxation. *J. Biomol. NMR* **32**, 295–308 (2005).

[13] Salmon, L., Yang, S. & Al-Hashimi, H. M. Advances in the determination of nucleic acid conformational ensembles. *Annu. Rev. Phys. Chem.* **65**, 293–316 (2014).

[14] Marušič, M., Schlagnitweit, J. & Petzold, K. Rna dynamics by nmr. *ChemBioChem* (2019).

[15] Vögeli, B. The nuclear overhauser effect from a quantitative perspective. *Prog. Nucl. Mag. Res. Sp.* **78**, 1–46 (2014).

[16] Nichols, P. *et al.* The exact nuclear overhauser enhancement: recent advances. *Molecules* **22**, 1176 (2017).

[17] Nichols, P. J. *et al.* Extending the applicability of exact nuclear overhauser enhancements to large proteins and rna. *ChemBioChem* **19**, 1695–1701 (2018).

[18] Giambaşu, G. M., York, D. M. & Case, D. A. Structural fidelity and nmr relaxation analysis in a prototype rna hairpin. *RNA* **21**, 963–974 (2015).

[19] Banás, P. *et al.* Performance of molecular mechanics force fields for rna simulations: stability of uucg and gnra hairpins. *J. Chem. Theory Comput.* **6**, 3836–3849 (2010).

[20] Bergonzo, C., Henriksen, N. M., Roe, D. R. & Cheatham, T. E. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common rna force fields. *RNA* **21**, 1578–1590 (2015).

[21] Bottaro, S., Banas, P., Sponer, J. & Bussi, G. Free energy landscape of gaga and uucg rna tetraloops. *J. Phys. Chem. Lett.* **7**, 4032–4038 (2016).

[22] Tan, D., Piana, S., Dirks, R. M. & Shaw, D. E. Rna force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.* 201713027 (2018).

[23] Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 12B634_1 (2015).

[24] Bottaro, S., Bengtsen, T. & Lindorff-Larsen, K. Integrating molecular simulation and experimental data: A bayesian/maximum entropy reweighting approach. *bioRxiv* 457952 (2018).

[25] Piccini, G., Mendels, D. & Parrinello, M. Metadynamics with discriminants: A tool for understanding chemistry. *J. Chem Theory Comput.* **14**, 5040–5044 (2018).

[26] Kuhrova, P., Banas, P., Best, R. B., Sponer, J. & Otyepka, M. Computer folding of rna tetraloops? are we there yet? *J. Chem. Theory Comput.* **9**, 2115–2125 (2013).

[27] Chen, A. A. & García, A. E. High-resolution reversible folding of hyperstable rna tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16820–16825 (2013).

[28] Wang, L., Friesner, R. A. & Berne, B. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). *J. Phys. Chem. B* **115**, 9431–9438 (2011).

[29] Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020603 (2008).

[30] Bottaro, S., Di Palma, F. & Bussi, G. The role of nucleobase interactions in rna structure and dynamics. *Nucl. Acids Res.* **42**, 13306–13314 (2014).

[31] Bentrude, W. G. & Hargis, J. H. Conformations of 6-membered-ring phosphorus heterocycles: the 5-t-butyl-2-oxo-1, 3, 2-dioxaphosphorinans. *J. Chem. Soc. D* 1113b–1114 (1969).

[32] Pitera, J. W. & Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *Journal of chemical theory and computation* **8**, 3445–3451 (2012).

[33] Boomsma, W., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Combining experiments and simulations using the maximum entropy principle. *PLoS Comput. Biol.* **10**, e1003406 (2014).

[34] Bottaro, S. *et al.* Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA* **25**, 219–231 (2019).

[35] Best, R. B., Lindorff-Larsen, K., DePristo, M. A. & Vendruscolo, M. Relation between native ensembles and experimental structures of proteins. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10901–10906 (2006).

[36] Bottaro, S., Gil-Ley, A. & Bussi, G. Rna folding pathways in stop motion. *Nucleic acids research* **44**, 5883–5891 (2016).

[37] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188 (1936).

[38] Sakuraba, S. & Kono, H. Spotting the difference in molecular dynamics simulations of biomolecules. *J. Chem. Phys.* **145**, 074116 (2016).

[39] Uyar, A., Karamyan, V. & Dickson, A. Long-range changes in neurolysin dynamics upon inhibitor binding. *J. Chem Theory Comput.* **14**, 444–452 (2017).

[40] Mendels, D., Piccini, G., Brotzakis, Z. F., Yang, Y. I. & Parrinello, M. Folding a small protein using harmonic linear discriminant analysis. *J. Chem. Phys.* **149**, 194113 (2018).

[41] Mendels, D., Piccini, G. & Parrinello, M. Collective variables from local fluctuations. *J Phys. Chem. Lett.* **9**, 2776–2781 (2018).

[42] Gong, Z., Schwieters, C. D. & Tang, C. Theory and practice of using solvent paramagnetic relaxation enhancement to characterize protein conformational dynamics. *Methods* **148**, 48–56 (2018).

[43] Salmon, L. *et al.* Modulating rna alignment using directional dynamic kinks: application in determining an atomic-resolution ensemble for a hairpin using nmr residual dipolar couplings. *Journal of the American Chemical Society* **137**, 12954–12965 (2015).

[44] Cesari, A. *et al.* Fitting corrections to an rna force field using experimental data. *Journal of chemical theory and computation* (2019).

[45] Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).

[46] Brandt, S., Sittel, F., Ernst, M. & Stock, G. Machine learning of biomolecular reaction coordinates. *The journal of physical chemistry letters* **9**, 2144–2150 (2018).

[47] Fleetwood, O., Kasimova, M. A., Westerlund, A. M. & Delemotte, L. Extracting molecular insights from conformational ensembles using machine learning. *BioRxiv* 695254 (2019).

[48] Sripakdeevong, P. *et al.* Structure determination of noncanonical rna motifs guided by 1 h nmr chemical shifts. *Nature Methods* **11**, 413 (2014).

[49] Escobedo, A. *et al.* Side chain to main chain hydrogen bonds stabilize polyglutamine helices in transcription factors. *Nat. Comm.* **10** (2019).

[50] Crehuet, R., Jorro, P. J. B., Lindorff-Larsen, K. & Salvatella, X. Bayesian-maximum-entropy reweighting of idps ensembles based on nmr chemical shifts. *BioRxiv* 689083 (2019).

[51] Beauchamp, K. A., Pande, V. S. & Das, R. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.* **106**, 1381–1390 (2014).

[52] Bottaro, S., Bussi, G., Kennedy, S. D., Turner, D. H. & Lindorff-Larsen, K. Conformational ensembles of rna oligonucleotides from integrating nmr and molecular simulations. *Sci. Adv.* **4**, eaar8521 (2018).

[53] Branduardi, D., Bussi, G. & Parrinello, M. Metadynamics with adaptive gaussians. *Journal of chemical theory and computation* **8**, 2247–2254 (2012).