Simultaneous NMR characterisation of multiple minima in the free energy landscape of an RNA UUCG tetraloop†

Aditi N. Borkar, a Pramodh Vallurupalli, b Carlo Camilloni, a Lewis E. Kay b and Michele Vendruscolo * a

RNA molecules in solution tend to undergo structural fluctuations of relatively large amplitude and to populate a range of different conformations some of which with low populations. It is still very challenging, however, to characterise the structures of these low populated states and to understand their functional roles. In the present study, we address this problem by using NMR residual dipolar couplings (RDCs) as structural restraints in replica-averaged metadynamics (RAM) simulations. By applying this approach to a 14-mer RNA hairpin containing the prototypical UUCG tetraloop motif, we show that it is possible to construct the free energy landscape of this RNA molecule. This free energy landscapes reveals the surprisingly rich dynamics of the UUCG tetraloop and identifies the multiple substates that exist in equilibrium owing to thermal fluctuations. The approach that we present is general and can be applied to the study of the free energy landscapes of other RNA or RNA-protein systems.

Introduction

Experimental studies of the conformational properties of RNA are challenging as these molecules are structurally heterogeneous. 1–8 These conformational properties, however, are highly important, since the dynamics of RNA molecules are often associated with their cellular functions. 1–3 From a theoretical perspective, the free energy landscape framework enables one to represent the structural and dynamical properties of macromolecules in an effective and concise manner. 2,9–12 The conformational space arising from localised secondary structural rearrangements. For example, flipped bases in the ground state tend to become increasingly stacked in excited states, and base-paired residues in the ground state tend to lose stacking in higher energy conformers. 15,16 Such conformational changes lead to sequestering or exposure of certain residues in the low populated states that can promote or inhibit RNA function depending upon the location and direction of local motion of the residue. Thus, such studies point out that RNA is inherently prone to complex structural dynamics at several hierarchical tiers 4,15,16,18,21 and it is a challenging task to isolate and characterise the sparsely populated states that are relevant to cellular function.

In this study, we address this challenge by exploiting the information about the dynamics of RNA provided by NMR residual dipolar couplings (RDCs), showing that this information makes it possible to obtain low populated structures in a prototypical RNA system – the UUCG tetraloop motif (Fig. 1). UUCG tetraloops, which belong to the UNCG family of RNA tetraloops, are thermodynamically stable, 23 and are frequently observed to cap RNA hairpins 24,25 and to provide nucleation sites for RNA folding, 26 although they have not been implicated yet in RNA–RNA
Fig. 1  Schematic representation of the 14-nucleotide RNA hairpin containing the UUCG tetraloop that we studied in this work. (a) The nucleobases of the tetraloop are referred to as UL1, UL2, CL3 and GL4. (b and c) Signature hydrogen bond interactions and residue orientations in the UUCG tetraloop. (b) UL1 and GL4 are base-paired via (i) UL1(O2')..GL4(O6) and (ii) UL1(O2')..GL4(N1) and/or N2. (c) Additionally, base-backbone interactions appear via (i) CL3(N4)..<UL2(3phosphate) and (iv) GL4(N7)..<UL2(O2'). This network of hydrogen bonds between the tetraloop residues confers high thermodynamic stability to this structural motif. The GL4 base is in the syn conformation (i.e. with $\gamma < 90^\circ$, where the dihedral angle $\gamma$ defines the rotation around the ribose-nucleobase glycosidic bond). CL3 stacks under UL1, while UL2 remains unpaired, non-stacked and exposed to the solvent. Both UL2 and CL3 are in a C2'-endo conformation of the sugar pucker that facilitates the bending of the RNA backbone at the tetraloop.

or RNA-protein interactions. Since its initial report in 1988, the UUCG structure has been characterized in great detail using high-resolution X-ray crystallography, and NMR structures are available for this tetraloop in the context of different RNA hairpins, where all its structural features are characterised in detail. The network of hydrogen bonds between the tetraloop residues confers high thermodynamic stability to this structural motif (Fig. 1b and c), which has also been investigated in detail using molecular dynamics simulations, probing the structural basis for its exceptional thermostability. Although studies on dynamics of the tetraloop have been mainly focused on the nanosecond timescale using in particular 13C NMR relaxation experiments. The first is the ‘timescale problem’, which concerns the length of the simulations. Our results illustrate instead a fairly rugged free energy landscape for this RNA tetraloop where several substates exist in equilibrium, consistently with a recent study on a related tetraloop.

Results and discussion

Replica-averaged metadynamics simulations

There are two major challenges in characterizing the structure and dynamics of the UUCG tetraloop, as well as more generally those of RNA molecules, via in situ experiments. The first is the ‘timescale problem’, which concerns the length of the simulations. To obtain an accurate sampling of the heterogeneous conformational space of this type of molecule it is necessary to identify the major structures that are populated, together with their corresponding populations, which at equilibrium are given by the statistical weights defined by the Boltzmann distribution. This is a challenging task, since the molecular dynamics simulations should be long enough to sample this distribution, or in other words, to converge to equilibrium. As standard molecular dynamics simulations, even for a system as simple as a RNA tetraloop, require trajectories well above the millisecond timescale, the ‘timescale problem’ can be more readily addressed by adopting enhanced sampling methods, which enable an efficient exploration of the conformational space by circumventing the requirement of generating realistic trajectories, thus enabling longer timescales to be accessed.

In this work we used the metadynamics approach, where, in order to enhance the conformational sampling, molecular dynamics simulations are biased by a history-dependent potential constructed as a sum of Gaussian functions deposited along the trajectory for suitably chosen collective variables. This bias enables the system to readily overcome energy barriers by discouraging the return to regions of the conformational space that have already been visited.

The second major challenge is the ‘force field problem’, which concerns the accuracy of the force field used during the sampling. This problem is distinct from the first, as using a force field that does not reproduce accurately the interatomic interactions that determine the motion of RNA molecules will result in a precise, but not necessarily accurate, free energy landscape, even if the sampling has fully reached convergence. In this case, the statistical weights obtained from the simulations will reproduce closely those corresponding to the force field used, but not necessarily those of the actual system under study. Given the importance of this challenge, substantial work has been carried out over many years to improve the accuracy of the force fields for nucleic acids. Here, rather than aiming at generating a ‘transferable’ force field by modifying its parameters...
to enable the simulations of any RNA system, we implement a 'system-dependent' strategy in which experimental data for a specific system are incorporated in the simulations as structural restraints. Alternatively, the simulated trajectories could be 'filtered' a posteriori by requiring experimental measurements to match the values back-calculated by selected conformations. In this work, we use the recently proposed replica-averaged metadynamics (RAM) approach, which enables the generation of ensembles of conformations consistent with the maximum entropy principle. We choose this strategy because it results in a structural ensemble that is the most probable one, given the force field and the experimental data used as restraints. Thus, by combining the advantages of enhanced sampling and of the experimental restraints, the RAM approach enables us to address both the timescale and the force field problems.

**RAM simulations of the 14-nucleotide RNA hairpin**

In order to characterize the conformational fluctuations of the 14-nucleotide RNA hairpin containing the UUCG tetraloop, we measured RDCs for 53 CH bonds in Pf1 phage alignment (see ESI, Table S1 and Fig. 3a, set A). We used 39 (Fig. 3b, set A) of these RDCs as restraints in RAM simulations, which at
convergence (Fig. S1, ESI†) generated the RAM ensemble (see ESI†). The remaining 14 RDCs (Fig. 3b, set A") from set A were used to validate the results of the RAM simulations. The initial conformation for the RAM simulations (see ESI†) was taken from a high-resolution NMR structure of the RNA hairpin, which was determined using an independent set of 30 RDCs30 (Fig. 3a, set B, and Table S2, ESI†). We note that both the RAM and the 2KOC30 structures were generated using RDCs as restraints, with the difference that the RAM structures are aimed at characterizing the conformational fluctuations of the RNA hairpin, while the 2KOC structures provide a representation of its high-resolution average conformation. The 23 bonds in common between the A and B sets of RDCs (Fig. 3a, green) are correlated (Fig. S2, ESI†) since they were measured under similar experimental conditions except for the concentration of the PF1 phage used for alignment.

For validation, in addition to the set A" of 14 RDCs mentioned above, we selected the subset ‘A not B’ (Fig. 3a, orange) of 30 RDCs for the 2KOC ensemble, and the subset ‘B not A’ (Fig. 3a, blue), of 7 RDCs for the RAM ensemble. We thus back-calculated the RDCs in five sets (A’, A not B’, A", B and ‘B not A’) for the 2KOC, RAM and MD ensembles, and compared them using three different metrics (Table S3, ESI†): the RMSD (in Hz), the Q factor and the Pearson’s coefficient of correlation (R). As expected, a consistent feature from the comparison between the experimental and calculated RDCs (Table 1 and Table S3, ESI†) is that the use of restraints led to an improvement in the agreement for the restrained bonds in the different ensembles. These results are closely resembling those obtained by a similar approach in the case of proteins.45,67 Both set B RDCs back-calculated over the 2KOC structures and set A' RDCs back-calculated over the RAM ensemble resulted in low RMSD values (2.20 Hz and 1.17 Hz, respectively). However, while the value for set A' is within the average experimental error (1.36 Hz) for RDCs measured here, the value for set B is larger. For the bonds that were not used as restraints for the 2KOC structure determination (set ‘A not B’), the RMSD (4.68 Hz, Table 1) is comparable to those back-calculated over the MD ensemble (4.57 Hz, Table 1). Conversely, the set A’ back-calculated over the RAM shows a RMSD (3.12 Hz) approximately consistent with the experimental error (2.2 Hz) and lower than that back-calculated over the 2KOC (4.17 Hz) or the MD (4.13 Hz) ensembles. Thus, the RAM ensemble consists of structures in good agreement with the relative orientations of both the restrained and unrestrained bonds as monitored by the experimental RDC values. These results indicate that the RAM ensemble describes rather accurately the motions of the 14-nucleotide RNA hairpin.

![Fig. 3](image-url) Summary of the RDC sets used as restraints in the molecular dynamics simulations and in the validation of the UUCG structural ensembles described in this work. The bars denote the intersection (total number of common bonds) or negation (number of unique bonds) and the circles below them denote the sets from which these are derived. For example, 23 bonds are common between the RDC sets A and B (green bar in panel a). Intersections are shown in green, and unique bonds in sets A and B are shown in orange and blue, respectively. (a) The set A of 53 measured RDCs comprises a set A' of 39 RDCs, which are used as restraints to generate the RAM ensemble, and a set A'' of 14 RDCs, which are used for validation. A previously reported30 set B of 30 RDCs is also used for validation; from the set B, a subset ‘B not A’ of 7 RDCs is extracted by considering RDC measurements for bonds not included in the set A. (b) From the set A, a subset ‘A not B’ of 30 RDCs is extracted by considering RDC measurements for bonds not included in set B.

**Table 1.** Assessment of the quality of the 2KOC, RAM and MD ensembles. Root mean square distance (RMSD, in Hz) between experimental and calculated RDCs. The results are shown for the three conformational ensembles (rows) and five RDC sets (columns) analysed in this study (see Fig. 3). RMSD values in italics denote the quality check of the restrained bonds and those in bold denote the validation of the unrestrained bonds in the 2KOC and RAM ensembles. All RMSD values for RAM and MD ensembles are calculated as weighted-averages of the RDCs obtained by fitting a single alignment tensor to each substate in the ensembles. Experimental errors for the 53 measured RDCs span about a 1–5 Hz range, as the measurements were performed at natural 13C abundance. The majority of these RDCs had, however, an experimental error of about 1 Hz, and most of such RDCs were included in set A', which was used for generating the RAM ensemble. More precisely, the experimental errors for the RDCs sets are about 1.4 Hz (set A'), 1.8 Hz (set ‘A not B’), and 2.2 Hz (set A' ''). All RDCs in set B have been reported to have an experimental error of about 1 Hz.10

<table>
<thead>
<tr>
<th>Set A'</th>
<th>Set A''</th>
<th>Set ‘B not A’</th>
<th>Set B</th>
<th>Set ‘A not B’</th>
</tr>
</thead>
<tbody>
<tr>
<td>2KOC</td>
<td>3.58</td>
<td>4.17</td>
<td>1.13</td>
<td>2.20</td>
</tr>
<tr>
<td>RAM</td>
<td>1.17</td>
<td>3.12</td>
<td>1.17</td>
<td>2.46</td>
</tr>
<tr>
<td>MD</td>
<td>4.44</td>
<td>4.13</td>
<td>1.35</td>
<td>3.78</td>
</tr>
<tr>
<td>Exp. error</td>
<td>1.4</td>
<td>2.2</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
RAM ensemble validation with data not used as restraints

To assess the quality of the RAM ensemble, which was generated using RDCs as structural restraints, we used independent NMR measurements. In the present case, the use of NMR $S^2$ order parameters is problematic because the presence of multiple substates of the RNA tetraloop (see below) complicates the structural interpretation of the experimental measurements, as these parameters measure the amplitude of motion around a given axis, which here may be different in the different substates. Thus, we used other available NMR parameters, including the nuclear Overhauser enhancement (NOE)-derived distances\(^ {30}\) (Fig. S3, ESI†) and dihedral angles derived from J-couplings\(^ {30}\) (Fig. S4, ESI†). The dihedral angles in Fig. S4 (ESI†) were calculated as an average from the Cartesian coordinates of the conformations in the RAM and MD ensembles. In this way, the fluctuations between rotameric states were also averaged without taking into account rotameric population distributions. The experimental dihedral angles values used for the validation of these calculated averages were sourced from ref. 30, where these values were calculated from $^3$J homo- and heteronuclear coupling constants and cross correlation rates.

The RAM and MD ensembles had only four NOE violations $>1$ Å (Fig. S3, ESI†). Two NOE violations are between the sugar 1H5′ and the base H8 atoms in G2 and A4. As there can be spin diffusion between the 1H5′ and 2H5′ nuclei that are right next to each other, these NOE violations are not surprising. The other two violations arise in the C5:G10 and A4:U11 base pairs immediately adjoining the UUCG tetraloop. Also the ensemble-averaged dihedral angles calculated from the RAM and MD ensembles conform mostly to the values determined from the experimental J-couplings, except for a few notable cases where the dihedral angles exhibit a high standard deviation from their calculated ensemble-averaged values and thus seem to deviate significantly from the experimental restraints (Fig. S4, ESI†). These results are indicative of conformational dynamics at individual residues of the RNA hairpin. For example $\gamma$ and $\zeta$ dihedral angles exhibit highest conformational flexibility in RNA,\(^ {68}\) particularly in the non-helical regions and this is reflected in the $\zeta$ values for UL2 and GL4 in the simulations. These angles show bimodal populations at gauche− (major) and gauche+ (and trans too for UL2) values and thus seem to, on average, deviate away from the gauche− restraints. In the recent NMR 2KOC structure,\(^ {30}\) although the $\beta$ and $\epsilon$ dihedral angles for the tetraloop residues fit better to a conformational distribution,\(^ {30}\) a higher number of NOE restraints has resulted in a higher precision of the loop structure and thus it is likely that the 2KOC structure underestimates the conformational dynamics of the tetraloop, in particular for UL2.

Free energy landscape of the UUCG tetraloop

In order to characterise the conformational ensembles of the UUCG tetraloop, we constructed the free energy landscape of the RAM ensemble as a function of its sketch-map CVs (see ESI†) and compared it with the corresponding free energy landscape of the MD ensemble. The advantage of using the sketch-map CVs is that the construction of the free energy landscape is not biased towards the CVs used in the RAM simulations; instead it is a two-dimensional projection of the multi-dimensional parameter hyperspace used to classify the RAM structures. In this study, we wanted to characterise the UUCG dynamics and thus we used the six backbone ($\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$ and $\zeta$) and the glycosidic ($\chi$) torsion angles of the six residues (CL – 1, UL1, UL2, CL3, GL4 and GL + 1) that include the UUCG tetraloop residues and its closing base pair (CL – 1 and GL + 1). The sketch-map CVs are thus a two-dimensional projection of the 42-dimensional eUUCGt torsion angle hyperspace.

The free energy landscape of the RAM ensemble (Fig. 4a) is more extended than that of the MD ensemble (Fig. 4b). For example, apart from the native conformations (R1 in Fig. 4a and U1 in Fig. 4b), we found other native-like conformations (R4 in Fig. 4b and U4, U5 and U6 in Fig. 4b) that only differ in the relative orientations of the UL2 $\gamma$ angle, as well as other non-canonical conformations (R2, R3, R5 and R6 in Fig. 4a and U2 and U3 in Fig. 4b) involving significantly larger rearrangements of the tetraloop residues. A comparison between the 2KOC structure of the UUCG tetraloop and the native and native-like conformations obtained in the RAM and MD ensembles is shown in Fig. S6 (ESI†). Although in Fig. 4 representative conformations might look similar, they have distinct sets of the 42 torsion angles used to calculate the sketch-map (Fig. S5, ESI†). Thus, different combinations of the torsion angles in RNA can result in apparently similar arrangement of the individual residues in the molecule.

Validation of the non-canonical UUCG conformations

Since multiple substates are present in both the RAM and MD ensembles, it is important to consider whether the structures of the substates in the MD ensemble are as accurate as those in the RAM ensemble. To test this possibility, we back-calculated (Table 2) the sets $A'$ and $A''$ of RDCs (Fig. 3) individually on the R1–R6 substates in the RAM ensemble (Fig. 4a), and on the U1–U6 substates in the MD ensemble (Fig. 4b). An important and consistent feature that we found is that the RMSD of the weighted-averages of the RDCs back-calculated over all the substates in the RAM and MD ensembles is better than that of any individual substate in the ensembles. This result illustrates the importance of taking into account conformational averaging for systems that populate different substates, rather than an individual free energy minimum. However, only for the RAM ensemble this average value is approximately consistent with the experimental error for the restrained (set $A'$, 1.4 Hz) and unrestrained (set $A''$, 2.2 Hz) bonds. This analysis suggests that the conformational heterogeneity present in the RAM ensemble represents better the extent of the structural fluctuations occurring in the UUCG tetraloop in solution.

Non-canonical UUCG conformations in structural databases

Conformational ensembles derived from structural databases may provide a representative sample of the structural fluctuations of nucleic acids under native conditions.\(^ {69}\) As observed previously in the case of protein dynamics, this analysis is based on the fluctuation-dissipation theorem, according to which the equilibrium structural fluctuations are equivalent to the changes caused by
small perturbations. One can consider each structure in the conformational ensemble derived from structural databases as subject to a slightly different perturbation, such as a bound ligand, a mutation, or the effect of crystal packing, which favours a particular minimum on the native state energy surface. If a sufficiently large number of conformations are collected they will reflect the statistical weights of the various minima in the free energy landscape of the unperturbed system.

In case of the UUCG structural database ensemble, a wide range of conformations occurring in the context of diverse full-length parent RNA molecules are available. In many of these cases, the UUCG is a component of the biological sequence of the RNA studied, and in others it is added to cap the truncated portions of some larger systems. We thus used these conformations to better understand the possible biological function of non-canonical UUCG states. In this analysis, we found several examples of non-canonical conformations, in ligand bound forms (PDB IDs 1EKZ, 1RAW, 1TLR and 3AMI) and ribosome structures arrested in translocation (PDB IDs 1FKA, 2GY9, 2GYB and 3IZF). These results are consistent with the view that RNA conformational transitions occur through complex, often multilayer RNA dynamics consisting of internal motions and externally induced rearrangements.

Conclusions

In this work we have addressed the problem of characterizing accurately the extent of conformational heterogeneity present in RNA in solution. By using molecular dynamics simulations with replica-averaged RDC restraints, we have described the conformational fluctuations of the UUCG tetraloop, detecting the presence of multiple interchanging minima. We have thus gained an understanding about why this motif exhibits high thermodynamic stability and yet is present in databases in alternative structures. More generally, our results illustrate the features of the free energy landscapes of RNA where several low populated states may exist in the vicinity of the most populated conformation.

Notes and references


29 E. Ennifar, et al., The crystal structure of UUGC tetraloop, *J. Mol. Biol.*, 2000, **304**, 35–42.


