Published online 18 April 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 12* ***5883–5891***
doi: 10.1093/nar/gkw239

RNA folding pathways in stop motion

Sandro Bottaro*, Alejandro Gil-Ley and Giovanni Bussi*

Scuola Internazionale Superiore di Studi Avanzati, International School for Advanced Studies, 265, Via Bonomea
I-34136 Trieste, Italy

Received February 2, 2016; Revised March 29, 2016; Accepted March 29, 2016

ABSTRACT

We introduce a method for predicting RNA folding pathways, with an application to the most important RNA tetraloops. The method is based on the idea that ensembles of three-dimensional fragments extracted from high-resolution crystal structures are heterogeneous enough to describe metastable as well as intermediate states. These ensembles are first validated by performing a quantitative comparison against available solution nuclear magnetic resonance (NMR) data of a set of RNA tetranucleotides. Notably, the agreement is better with respect to the one obtained by comparing NMR with extensive all-atom molecular dynamics simulations. We then propose a procedure based on diffusion maps and Markov models that makes it possible to obtain reaction pathways and their relative probabilities from fragment ensembles. This approach is applied to study the helix-to-loop folding pathway of all the tetraloops from the GNRA and UNCG families. The results give detailed insights into the folding mechanism that are compatible with available experimental data and clarify the role of intermediate states observed in previous simulation studies. The method is computationally inexpensive and can be used to study arbitrary conformational transitions.

INTRODUCTION

Despite its simple four-letters alphabet, RNA exhibits an amazing complexity, which is conferred by its ability to engage and interconvert between a variety of specific interactions with itself as well as with proteins and ions (1). A delicate balance between many different factors, such as hydrogen bonding, stacking interactions, electrostatics and backbone/sugar flexibility, determines structure and dynamics of RNA. These interactions are explicitly described in atomistic molecular dynamics (MDs) simulations, that represent an important computational tool for the investigation of RNA dynamics. Since the first MD simulation on tRNA (2), the accuracy of atomistic force fields has steadily improved, to the point that it is nowadays possible to obtain stable trajectories on the microsecond time scale for A-form double helices and tetraloops (3,4). MD simulations have also proven useful to aid the interpretation of experimental data for structured RNAs (5,6) and protein–RNA complexes (7–9). However, models used for nucleic acids are still significantly less accurate compared to those used for proteins (10). Recent simulations on systems amenable to converged sampling showed that none of the current atomistic force fields correctly reproduces the behavior of single stranded RNA tetranucleotides in solution (11,12). As a consequence, understanding the folding dynamics of small motifs such as RNA tetraloops is extremely challenging using MD, both because of the force-field limitations and of the high computational cost.

From a different perspective, structural bioinformatics approaches to study RNA exploit the empirical relationship between sequence and three-dimensional (3D) structure. These methods typically proceed by first extracting local conformations from a database of experimental structures, which are then assembled together to form complete models. This idea lead to the development of successful fragment assembly algorithms for RNA (13–16) and proteins (17). Fragment assembly methods often rely on the working hypothesis that frequencies of appearance in solved structures can be considered as an approximation to the underlying Boltzmann distribution. While this latter statement is in general not true, the statistics obtained from protein structures in the protein data bank (PDB) were shown to reproduce distributions as measured by nuclear magnetic resonance (NMR) spectroscopy (18) as well as quantum mechanical calculations (19,20). Additionally, the successful applications of statistical potentials to different RNA structure prediction problems suggest this approximation to be in practice acceptable (21).

By pushing these assumptions further, one could imagine the possibility to use fragment libraries to predict not only the most stable conformations but also intermediate and excited states (22), so as to provide a description of reactive pathways. For example, the conformational variations observed in crystal structures of RNA triloops have been linked to their internal dynamics (23), and a simple isomerization process occurring in the backbone of a small

*To whom correspondence should be addressed. Tel: +39 040 3787 407; Fax: +39 040 3787 528; Email: bussi@sissa.it
Correspondence may also be addressed to Sandro Bottaro. Tel: +39 040 3787 301; Fax: +39 040 3787 528; Email: sbottaro@sissa.it

peptide was recently rationalized by analyzing high-energy structures trapped in the PDB (24).

A further methodological step is however required to reconstruct the dynamics of systems undergoing large conformational changes, possibly involving multiple pathways, when only an ensemble of structures at equilibrium is given. Of particular interest in this context is the concept of diffusion map (25), which describes the long timescale dynamics of a complex system using a transition matrix directly calculated from the data. This important theoretical tool was applied to characterize the dynamics of proteins systems (26). Interestingly, the slow eigenmodes of the transition matrix have been shown to be consistent with an analysis based on transition networks of long MDs simulations (27). This result suggests that the reactive pathways of macromolecular systems can be obtained from equilibrium ensembles alone.

In this paper, we show that the structural ensemble of RNA fragments extracted from the PDB exhibits remarkably good agreement with available NMR experimental data for five tetranucleotides. We then consider the fragment ensembles of the two most important families of RNA tetraloops: GNRA and UNCG (R = A/G and N =any). By using a formalism related to diffusion maps, we build a random walk on a graph in which each node is an experimental 3D fragment and edges are weighted with an appropriate measure of similarity (28). The properties of the resulting random walks are then analyzed using the standard machinery of Markov state modeling (29) to obtain detailed folding trajectories.

We call the method for obtaining folding pathways from experimental fragments *stop-motion modeling* (SMM), in analogy with the animation technique that produces the impression of motion through juxtaposition of static pictures. SMM is a very general tool for obtaining transition pathways, it is computationally inexpensive, and it is therefore ideally suited to complement MD simulations and to aid the interpretation of NMR spectroscopy data and kinetic experiments.

MATERIALS AND METHODS

Stop motion modeling

We here describe how to set up the SMM procedure.

(i) *Pairwise distances*. We first calculate all pairwise distances between all 3D structures within an ensemble as $d_{ij} = \mathcal{E}\mathrm{RMSD}(\mathbf{x}_i, \mathbf{x}_j)$. Here $\mathbf{x}_i$ and $\mathbf{x}_j$ are the coordinates of structures *i* and *j* and $\mathcal{E}$RMSD is an RNA-specific metric based on the relative orientation of nucleobases only (28). This measure has the important property of being highly related to the temporal distance: this means that when two structures are close in $\mathcal{E}$RMSD distance, then they are also kinetically close. In a previous work (28), we have shown that this property is satisfied to a larger extent by $\mathcal{E}$RMSD compared to standard RMSD after optimal alignment (30) as well as distance RMSD measures. Additionally, we have proven $\mathcal{E}$RMSD to be accurate in recognizing known RNA motifs within the structural database. $\mathcal{E}$RMSD is also highly correlated with interaction network fidelity (31), and thus distinguishes structures with a different pattern of base–base interaction.

(ii) *Transition matrix*. We build a graph on an ensemble of structures by constructing the adjacency matrix *K* with Gaussian kernel $K_{ij} = \exp(-d_{ij}^2/2\sigma^2)$. Here, we set $\sigma = 0.2$, which is $\approx 1/3$ of the typical $\mathcal{E}$RMSD threshold used to consider two structures significantly similar. This ensures that the transition probability among structures that exhibit a different pattern of base–base interaction is vanishingly small. Following a procedure similar to the diffusion maps approach (25), and common in graph theory (32), we construct the Markov matrix *T* dividing by the diagonal degree matrix *D*, defined as $D_{ii} = \sum_j K_{ij}$. To obtain a *T* matrix that is simultaneously normalized and symmetric we introduce here an iterative normalization procedure $T_{(t+1)} = D_{(t)}^{-1/2} T_{(t)} D_{(t)}^{-1/2}$. Here the matrix product is implicit, the subscript $_{(t)}$ indicates the iteration, and $T_{(0)} = K$. At convergence, this procedure yields a matrix *T* that can be interpreted as a transition probability matrix. Here T_{ij} is the probability of observing a direct transition from state *i* to state *j*. *T* is normalized ($\sum_j T_{ij} = 1$) and has an uniform equilibrium distribution over the ensemble of fragments ($\sum_i T_{ij} = 1$). This procedure is closely related to the most common version of graph Laplacian normalization. However, we notice that in the usual procedure the transition matrix is made non symmetric by normalization, and thus has an equilibrium distribution that is not necessarily uniform. The advantage of our formulation is that the averages computed from the random walk are by construction identical the to ensemble averages obtained from the original set of structures.

(iii) *Transition pathways between states.* Dynamical properties of the system are calculated from the transition matrix *T* as described in Ref. (29), and briefly reported in Supporting Information 1 for clarity. The flux calculations are performed using pyEMMA (33). To facilitate the analysis of the fluxes, nodes are lumped together using a standard spectral clustering procedure (34) on the transition matrix *T*. Notice that the lumping is only used to compute the aggregate fluxes, and does not influence in any way the underlying Markov model. We set the number of clusters to 25 and 45 for UNCG and GNRA tetraloops, respectively. We empirically verified that the fluxes are robust with respect to the number of clusters. The flux between distant structures (compared to the Gaussian width σ) depends on the number of transition pathways connecting them. If the overall transition requires crossing intermediate states with very low population, corresponding to high free-energy barriers, the resulting flux will be very small. In the limit of missing intermediates, the graph becomes disconnected and the resulting flux is zero.

(iv) *Low-dimensional embedding.* For the sake of visualization, folding pathways and clusters are projected on a low-dimensional space. To this aim we use the diffusion map technique, where the top eigenvectors of the matrix *T* are interpreted as coordinates that provide a low-dimensional embedding in which the local structure (small distances) is preserved (25,26). We empirically found that to make the visualization clearer it is convenient to use a value of σ 3-4 times larger compared to the one used to calculate the fluxes. This choice only affects the two-dimensional projection, and not the calculated pathways and clusters.

Molecular dynamics simulations

MDs simulations were performed using the GROMACS software package (35). RNA was modeled using the Amber99 force field with parmbsc0 and χ_{OL3} corrections and was solvated in explicit water and ions (3,36–39). Parameters are available at http://github.com/srnas/ff. Temperature replica exchange MD was used to accelerate sampling (40). For each system 24 replicas in the temperature range 300–400K were simulated for 2.2 μs per replica. More details are available in SI2.

RESULTS

We extract all the fragments with a given 4-nucleotide sequence from high-resolution crystal structures in the PDB. We consider all RNA-containing structures with resolution 3.5Å or better available in the PDB database as of 19 August 2015, for a total of 1882 structures. A complete list is provided in Supplementary Information 3. This procedure yields a collection of structures (fragments ensemble) composed by 2–15 thousands fragments, depending on the sequence.

Comparison against NMR data

Recent NMR spectroscopy studies on RNA tetranucleotides showed AAAA, CCCC, CAAU and GACC to be mostly in A-form in solution, while data for UUUU were compatible with more disordered, partially stacked conformations (11,41). In all cases, no evidence of intramolecular hydrogen bonding was found.

In this section we compare available NMR data with ideal A-form helices, with ensembles of fragments from the PDB, and with MD simulations. As a term of reference, we first set out to compare available experimental NMR data with the prediction obtained from ideal A-form helices. Figure 1 shows the agreement of nuclear overhauser effect (NOE) data with the values predicted from A-form helices in terms of (i) root mean square deviation (RMSD), (ii) percentage of NOE distance violations and (iii) number of non-local false positives, i.e. distances between protons in non-consecutive nucleotides predicted to have average NOE distance ≤5 Å but not visible in the experiment. These definitions are consistent with those used in Ref. (11). The deviation between predicted and experimental data is in general low (Figure 1A). However, since flexible tetraloops can adopt non-A-form structures, a large fraction of predicted distances falls outside the experimental range (Figure 1B). Furthermore, systematic discrepancies between NMR spectroscopy and X-ray crystallography have been discussed (42) and contribute to the fraction of NOE violations. In all cases, no false positive is observed. Compatibly with its more disordered behavior in solution, the predicted NOE distances for the A-form UUUU is poorer compared to the other tetranucleotides. In Figure 1 we show the agreement between experimental and predicted ${}^3J$ scalar couplings, considering data reporting on sugar pucker (panel D) and on backbone conformation (panel E). As observed for NOE, experimental scalar couplings are compatible with A-form helices, with the exception of UUUU. In the latter case, the experimental data suggest high sugar mobility, with significant deviation from the common C3′-endo sugar pucker conformation.

The accord of the fragment ensembles with experimental data is very similar to what observed for the A-form helix (Figure 1). This can be rationalized at least for the first four tetranucleotides by considering that Watson–Crick double helices are the dominating conformations in the crys-

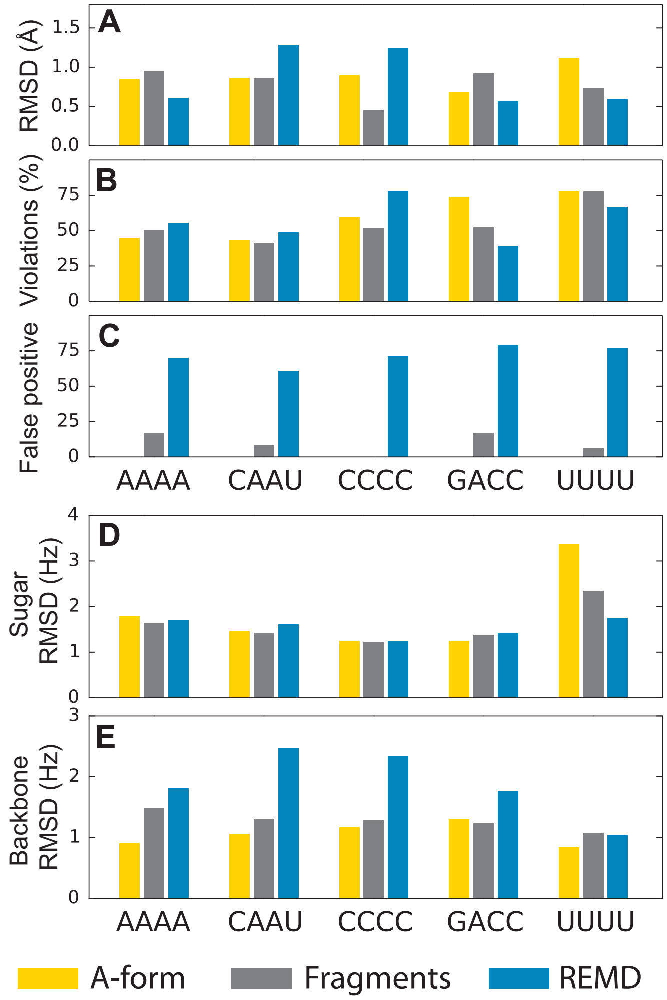

Figure 1. Comparison between calculated and experimental NMR data. Calculations were performed on a single ideal A-form helix (A-form), on the ensemble of fragments extracted from the PDB (Fragments), and on replica exchange molecular dynamics simulations (REMD). (**A**) RMSD between calculated and predicted average NOE distances. (**B**) Percentage of predicted NOE average distance outside the experimental range. (**C**) Number of false positives, i.e. predicted distances below 5 Å not observed in experimental data. (**D**) RMSD between experimental and predicted ${}^3J_{1'2'}$, ${}^3J_{2'3'}$, ${}^3J_{3'4'}$ scalar couplings, reporting on sugar pucker geometry. (**E**) RMSD between experimental and predicted ${}^3J_{5'/5''P}$, ${}^3J_{4'5'/5''}$, ${}^3J_{3'P}$ scalar couplings, reporting on backbone geometry.

tal structures deposited in the PDB database. In the case of the UUUU tetranucleotide, the conformational ensemble obtained from the fragments improves the agreement with NMR data when compared with the ideal A-form helix. This highlights the importance of using a diverse ensemble of structures to describe a flexible molecule in solution.

Figure 1 also reports the agreement between experimental data and temperature replica exchange molecular dynamics (REMD) (40) at atomistic resolution. In terms of RMSD, NOE distances predicted by REMD simulations are comparable with those obtained from the fragments ensemble. However, a high number of non-local NOE false positives can be observed for all the systems (Figure 1C) . As reported in recent MD studies, AMBER force fields over-stabilize intercalated structures with stacking between bases 1–3 and 1–4 (11,12). These structures, which are also observed in our simulations, cause the appearance of spurious contacts that are not compatible with experimental NOE distances (i.e. false positives). The wrong pattern of stacking interactions observed in REMD simulations affects the backbone conformation as well, resulting in a poor agreement with backbone ^{3}J scalar coupling data (Figure 1E). With respect to sugar pucker, instead, it is worth noting that for UUUU both C3′-endo and C2′-endo conformations are significantly populated in REMD simulations, in accord with scalar coupling data (Figure 1D). Scatter plots of experimental versus predicted NOEs and ^{3}J scalar couplings are shown in Supplementary Information 4 and 5.

Taken together, these results show that the fragments ensembles are overall in accord with NMR data. For tetranucleotides which in solution adopt the A-helix form, the agreement is comparable with what obtained from a single structure. On the contrary, noticeable artifacts are observed in MD simulations, in agreement with previously reported results where all the state-of-the-art force fields were tested (11,12). The most striking discrepancy between simulations reported here and NMR is caused by the presence of intercalated structures. This indicates that results obtained sampling the fragment ensemble could be potentially more accurate than expensive MD simulations in reproducing biomolecular dynamics, at a fraction of the computational cost.

Stop-motion modeling: mimicking dynamics using static snapshots

The results described above show that the conformational ensemble of fragments on five different systems is in agreement with NMR data. Building upon this result, we use the fragments ensemble to study the dynamical properties of arbitrary RNA sequences. First, we assume that the fragments ensemble for a given sequence represents the equilibrium distribution at some temperature. We then construct a Markov matrix on all the fragments within the ensemble. Following a procedure closely related to the one used in the construction of diffusion maps (25), we calculate transition probabilities as Gaussian function of their distance. In the present context, distances are measured using the $\mathcal{E}$RMSD (28), which is based on the relative position and orientation of nucleobases only. The resulting Markov matrix contains kinetic information, from which it is possible to either generate reactive trajectories with a stochastic procedure or to analyze the associated fluxes as it is customarily done for Markov state models (29,43). We call this procedure SMM. In the next section we present the results of the SMM on different tetraloops, while we refer the reader to the 'Materials and Methods' section for an in-depth, technical description of the algorithm.

Folding pathways of RNA tetraloops in stop motion

We use the SMM approach described above to study the helix-to-loop transition of UNCG and GNRA tetraloop families (see Table 1). When stabilized by additional Watson–Crick base pairs, these sequences are known to adopt specific stem-loop structures (44). The corresponding fragments ensembles are indeed typically composed by (i) fully stacked structures in A-form conformations, (ii) folded tetraloops and (iii) other conformations that are distant from both loop and A-form.

Here, we assume that the folding mechanism is determined by the tetraloop only, and we do not model the full stem-loop sequence. As a consequence, our investigation only provides insight on the folding mechanism of the analyzed tetraloop sequences, without considering possible effects of the stem sequence.

UNCG Tetraloop. The UNCG is one of the most abundant and well-characterized families of tetraloops (45). NMR and X-ray structures of these tetraloops revealed that their high stability is conferred by a peculiar trans-Watson–Crick/sugar-edge (tWS) base pair (46) between G4 and U1, together with extensive U2-G4 stacking and a U2-C3 base–phosphate interactions (47,48). In Figure 2, we show the main folding pathways obtained from SMM connecting the A-form helix (cluster 1) to the UNCG tetraloop (cluster 7). For visualization purpose, the folding pathways are projected on the first two diffusion coordinates, that describe the slowest directions of propagation of the Markov chain (25).

Additionally, we project the folding pathways on the U1-G4 distance and on the value of the χ torsion angle in G4, reporting respectively on U1/G4 base-pair formation and on the *anti*/*syn* transition of the G4 glycosidic bond (Figure 2B). For UUCG we found three dominant folding pathways. In the first one (36% of the total flux) we observe an initial elongation phase, during which U1/U2 unstack (cluster 2), followed by the loss of U2/C3 stacking interaction (cluster 3). The loop then undergoes a major rearrangement, in which G4 flips into *syn* conformation (cluster 4), U1 and G4 approach (clusters 5–6), until the formation of the tetraloop, featuring the characteristic G4-U1 tWS base pair (cluster 7). The second pathway (23% of the flux) is qualitatively similar to the first one, as most of the intermediates are in common. In the third pathway (17% of the flux), instead, unstacking first occurs at the 3′ end. A stochastic trajectory representing the folding process is shown in Supplementary Movie 1.

UACG tetraloop folding proceeds through a similar mechanism, although G4 flipping can also occur after loop compaction, as shown in Supplementary Information 6. The SMM analysis could not be performed on the UCCG

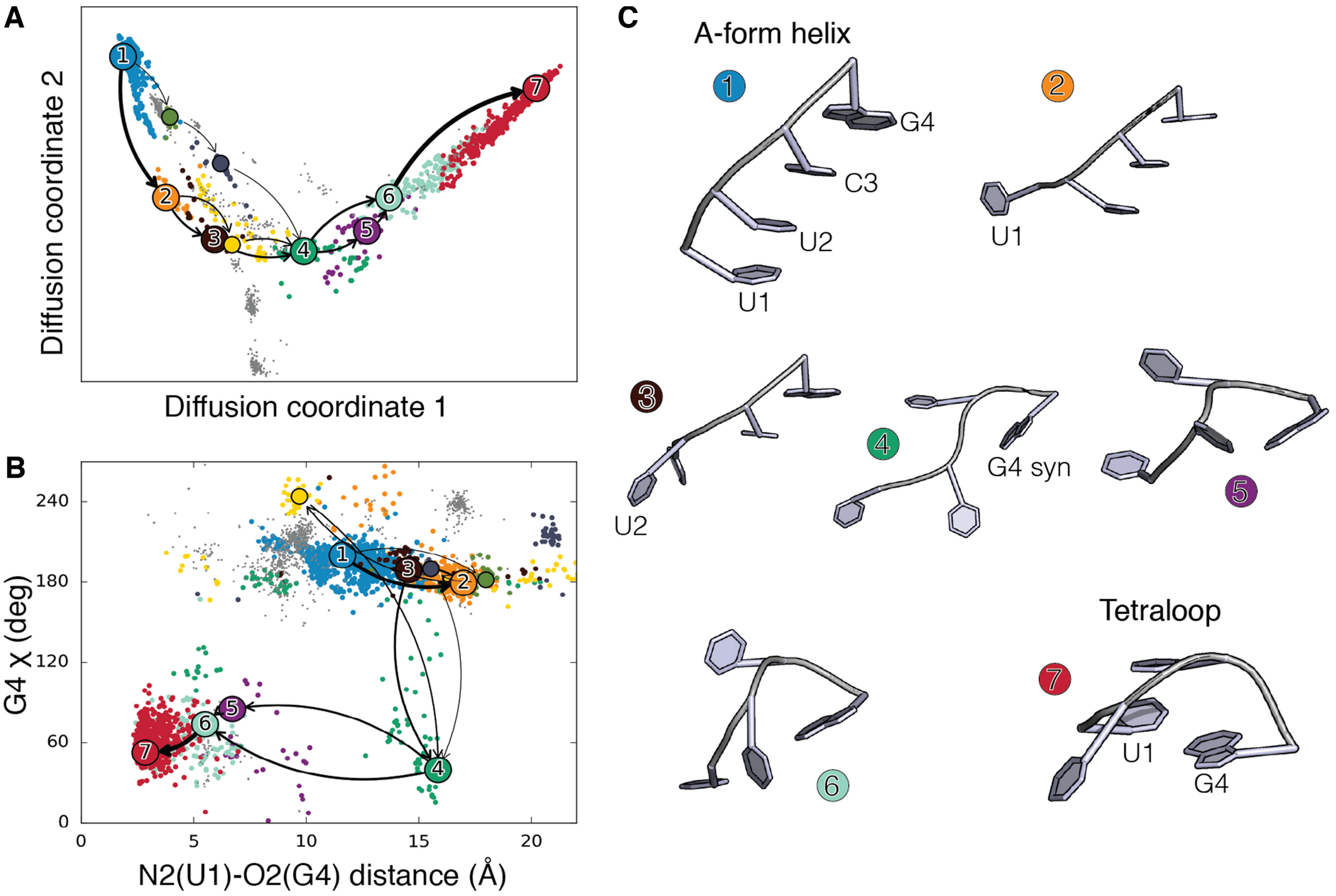

Figure 2. (**A**) UUCG Helix to loop folding pathways projected on the first two diffusion coordinates. Clusters corresponding to A-form helix and to the loop are colored in blue and red, respectively. On-pathways are shown in colors, while gray points correspond to clusters not contributing to the folding pathway. Arrows show the dominant folding pathways (≈75% of the total folding flux). Line width is proportional to the flux. (**B**) Helix to loop folding pathway projected on the distance between atom O2 in base U1 and atom N1 in base G4 versus the glycosidic torsion angle in G4. The distance reports on the formation of a signature interaction of the tetraloop. (**C**) Centers of the clusters in the main folding pathway, numbered and colored as in panels A and B.

Table 1. Number of fragments within each ensemble

Sequence	# of structures	Loop (%)	A-form (%)
UACG	3674	9	37
UCCG	6673	3	67
UGCG	5872	0	61
UUCG	3969	15	17
GAAA	12 908	16	8
GAGA	7596	10	27
GCAA	7835	14	23
GCGA	10 946	7	17
GGAA	12 395	3	24
GGGA	16 764	0.7	46
GUAA	8100	6	15
GUGA	11 432	10	10

The fraction of loops and A-form helices is calculated considering all structures with $\mathcal{E}$ RMSD < 0.6 from the consensus loop/ideal helix.

and UGCG sequences because none or very few of the fragments analyzed adopt the tetraloop conformation (Table 1).

It is worthwhile observing that the remaining ≈25% of the reactive flux visits not only the clusters discussed above, but also other fragments (shown as gray dots in Figure 2). It is however possible to calculate the contribution of pathways visiting specific conformations of interest. In a previous MD simulation study it was identified a metastable UUCG tetraloop in which G4 is in *anti* conformation (49). This structure was discussed in terms of its similarity to loop 32–37 in PDB ID: 3AM1 (50). In our analysis, this specific structure from the PDB is assigned to one of the least populated clusters (Supplementary Information 7). The pathways visiting this cluster only contribute to a very small fraction of the reactive flux (≈0.2%), suggesting that this could be an off-pathway intermediate.

GNRA tetraloop. Contrary to UNCG, GNRA tetraloops often mediate RNA tertiary interactions (51). GNRA tetraloops are characterized by a trans sugar/Hoogsteen (tSH) G1-A4 non-canonical base-pair, and by N2-R3-A4 stacking (52). We show in Figure 3 the folding pathways obtained from SMM on the GAAA tetraloop. Starting from a fully stacked A-form helix (cluster 1), the main folding pathway consists in two steps, namely G1 unstacking at 5′ end (cluster 2) followed by a rotation around the α dihedral angle in A2 (53), leading to the formation of the tSH G4-

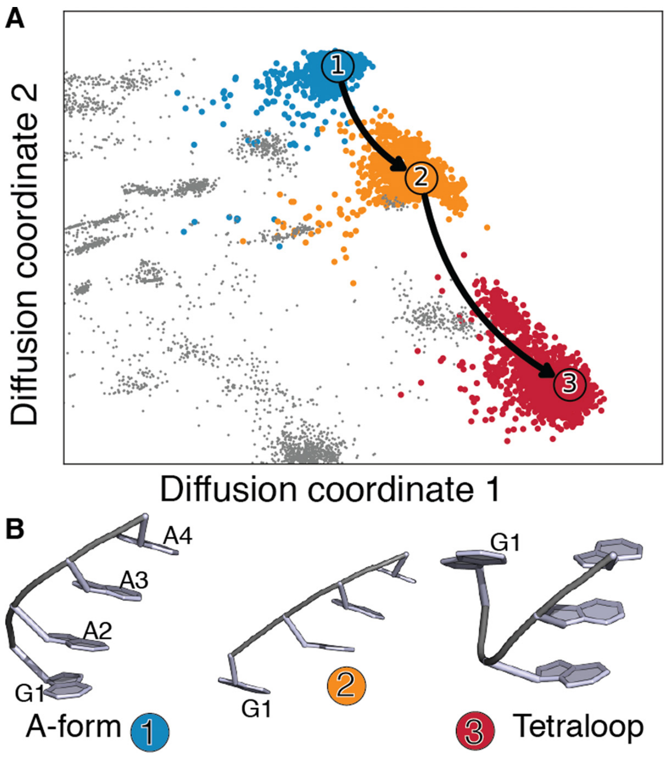

Figure 3. (A) GAAA A-form helix to loop folding pathways projected on the first two diffusion coordinates. The first folding pathway, contributing for more than 90% of the total flux, is indicated as a solid line. Only the part of the diffusion map containing the clusters that contribute significantly to the folding pathways is shown. See Supplementary Information 8 for the full two-dimensional projection. (B) Centers of the clusters in the main folding pathway, numbered and colored as in panel A.

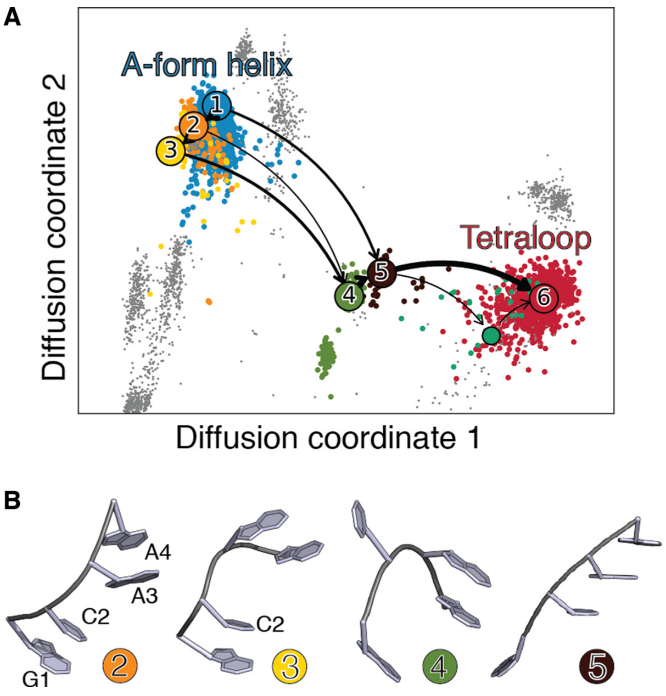

Figure 4. (A) GCAA helix to loop folding pathways projected on the first two diffusion coordinates. Clusters corresponding to A-form helix and to the loop are colored in blue and red, respectively. On-pathways are shown in colors, gray points correspond to fragments not contributing to the folding pathway. The first three folding pathways, contributing for ≈80% of the total flux, are indicated with black arrows. Line width is proportional to the flux. (B) Centers of the clusters in the main folding pathway, numbered and colored as indicated in panel A. The three-dimensional structures of the A-form helix (cluster 1) and tetraloop (cluster 6) are not shown, as they are equivalent to clusters 1 and 3 in Figure 3.

A1 base pair (cluster 3). For GAAA, as well as for GAGA tetraloops (see Supplementary Information 8 and 9) this pathway contributes for 90% of the flux.

When considering the third most common GNRA tetraloop sequence (GCAA) a folding mechanism similar to the one described above is observed, with the additional presence of stacking/unstacking dynamics between C2 and A3, as shown in Figure 4, clusters 1-3. More generally, we systematically observe significant unstacking dynamics in all GYRA tetraloops (Y = C or U) which is absent in GAAA and GAGA tetraloops (see Supplementary Information 9). This is expected, as purine-pyrimidine stacking is less strong compared to purine–purine stacking. Note that the high resolution structures used here do not contain a significant percentage of GGGA and GGAA sequences forming the consensus GNRA tetraloop (Table 1), thus making it difficult to perform the SMM on these two sequences.

Diffusion maps identify kinetically distant motifs. The Euclidean distance on the diffusion map is related to the rate of connectivity of the points in the Markov chain (25). Therefore, the diffusion map can be used to identify structures which are kinetically far from each other, irrespectively of their relevance in the helix-to-loop transition. As an example, we report the presence of structures in the GAAA fragment ensemble featuring the typical signature interactions of the UUCG tetraloop (see Supplementary Information 8). This similarity has not been reported before and suggests that some tetraloops with GNRA sequence could have a functional role similar to the one of UNCG tetraloops.

DISCUSSION

In this paper, we present a method to build RNA folding pathways based on the analysis of the ensemble of fragments in the available structural database. The different conformations and their frequencies are dictated by many factors: the available experimental structures, the crystallization conditions, the crystal packing, and the non-local interactions with proteins, ions, and other RNA bases. All these differences act as perturbations that stabilize intermediate states and alternate minima.

Firstly, we show that the fragments ensembles quantitatively agree with available NMR spectroscopy data for five selected RNA tetranucleotides. This result is compatible with a previous study showing that the structural ensembles of proteins in the PDB provide a representative sampling of the heterogeneity of their native states as probed by various NMR measurements (18). The agreement is a non trivial result since the PDB has an intrinsic database bias, in which A-form helices are likely over-represented with respect to other structures. It is important to mention that this bias could be alleviated or removed by using filtering procedures based on available experimental data (54,55). In principle, one could even extrapolate the structural ensembles to different temperatures or ionic conditions, provided that enough overlap between distributions exists. We also performed extensive, fully atomistic MD simulations showing instead artifacts not compatible with NMR data for the same tetranucleotides. This confirms the results obtained in recent simulation studies (11,12).

Secondly, we extend the scope of fragments ensemble from thermodynamics to dynamics. More precisely, we introduce a procedure, called SMM, for analyzing structural ensembles, so as to produce reaction pathways for generic conformational transitions in RNA. This procedure finds its theoretical underpinnings in the framework of transition-path theory (43) and Markov-state models (29), widely used in the protein simulation community (56). Here, however, we introduce two significant modifications. First, we use experimental crystal fragments instead of MDs to construct the Markov states. Second, we calculate transition probabilities from structural distances instead of estimating rates from many, short MD simulations. This latest idea is borrowed from the framework of diffusion maps, where a transition matrix is obtained from structural distances (26).

We employ SMM to study the helix-to-loop transition of UNCG and GNRA tetraloop families, leading to a detailed description of their folding pathways. This study would be very difficult or not possible using atomistic MD simulations as well as other structural bioinformatics approaches (13,57), due to the known issues in properly modeling RNA tetraloops. Note that the stability of the full tetraloop structures is known to be dependent on the length and sequence of the stem. Here we combined fragments with a fixed sequence in the loop without explicitly considering the nucleotides of the stem, as our procedure relies on the fact that the heterogeneity of sequences observed in the PDB acts as a perturbation apt to stabilize the most accessible intermediate states. As an example, by exclusively considering fragments with sequence cUUCGg, the ensemble would consist almost entirely of folded tetraloops. The inclusion in the ensemble of fragments with sequence cUUCGc has the net effect of destabilizing the loop, making it possible to observe extended conformations.

For the UUCG tetraloop we observe a folding mechanism characterized by a progressive unstacking followed by a concerted movement involving *anti* to *syn* flip of the glycosidic bond and loop compaction. T-jump experiments suggested a four-states sequential folding model characterized by unstacked (S), unfolded (U), frayed (E) and native (N) state (58,59). We hypothesize the unfolded (yet stacked) state U to correspond to clusters 2–3 in Figure 2, and the frayed state to correspond to the compact clusters 5–6. In T-jump experiments the unfolded and unstacked state S is populated only at high temperatures. Consistently, this state is not observed in the fragment ensemble. Our analysis also shows the G4 *anti*→*syn* flip to occur concertedly with stem formation, with no direct evidence suggesting the tetraloop folding to occur from a misfolded, compact structure. It is important to observe that structures where the stem is formed and G4 is in *anti* were observed in previous MD simulation studies (49,60). Although similar structures were also present in the PDB and thus included in our fragment ensembles, the pathways visiting these conformers contribute to a small extent to the reactive flux, suggesting these structures to be off-pathway intermediates.

Tetraloops of the GNRA family show a simpler folding mechanism, in which G1 unstacks from N2 and then forms a base pair with A4. Fluorescence decay experiments (61) as well as NMR measurements (62) strongly suggest that GNRA tetraloops undergo significant conformational dynamics, in which N2 and R3 can interconvert between different stacking arrangements. In particular, C2 in GCAA tetraloop was found to be much more dynamic with respect to A2 in GAGA and GAAA. This is completely consistent with our analysis that shows significant stacking dynamics in GCAA (Figure 4), and more generally in sequences where N2 is a pyrimidine.

The results presented in this paper show fragment ensembles to provide a quick and realistic manner to generate equilibrium distributions for short RNA sequences compatible with solution experiments. The introduced SMM procedure makes it possible to obtain reaction pathways from these ensembles at a small computational cost. This allows us to provide a detailed description of the folding mechanism for the most common RNA tetraloops that is compatible with available experimental data and clarifies the role of intermediate states observed in previous MD studies. In this work we focused on systems for which the statistics of fragments extracted from the PDB is sufficient to generate paths between relevant metastable conformations. The SMM procedure can be straightforwardly applied to larger systems provided that meaningful ensembles are available. We envision the possibility of using SMM to analyze ensembles generated by MD or fragment-assembly techniques (13,14) for the characterization of conformational dynamics in larger molecules.

AVAILABILITY

Stop motion modeling has been implemented in the baRNAba package freely available at https://github.com/srnas/barnaba.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

Pascal Auffinger, Pavel Banas, Alan Chen, Richard Andre Cunha, Kathleen Hall and Eric Westhof are acknowledged for reading the manuscript and providing useful comments.

FUNDING

European Research Council under the European Union's Seventh Framework Programme [FP/2007-2013]/ERC [306662, S-RNA-S].

Conflict of interest statement. None declared.

REFERENCES

1. Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

2. Harvey,S.C., Prabhakaran,M., Mao,B. and McCammon,J.A. (1984) Phenylalanine transfer RNA: molecular dynamics simulation. *Science*, **223**, 1189–1191.
3. Banáš,P., Hollas,D., Zgarbová,M., Jurečka,P., Orozco,M., Cheatham,T.E. III, Šponer,J. and Otyepka,M. (2010) Performance of molecular mechanics force fields for RNA simulations: stability of UUCG and GNRA hairpins. *J. Chem. Theory Comput.*, **6**, 3836–3849.
4. Giambasu,G.M., York,D.M. and Case,D.A. (2015) Structural fidelity and NMR relaxation analysis in a prototype RNA hairpin. *RNA*, **21**, 963–974.
5. Musiani,F., Rossetti,G., Capece,L., Gerger,T.M., Micheletti,C., Varani,G. and Carloni,P. (2014) Molecular dynamics simulations identify time scale of conformational changes responsible for conformational selection in molecular recognition of HIV-1 transactivation responsive RNA. *J. Am. Chem. Soc.*, **136**, 15631–15637.
6. Pinamonti,G., Bottaro,S., Micheletti,C. and Bussi,G. (2015) Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments. *Nucleic Acids Res.*, **43**, 7260–7269.
7. Whitford,P.C., Blanchard,S.C., Cate,J.H. and Sanbonmatsu,K.Y. (2013) Connecting the kinetics and energy landscape of tRNA translocation on the ribosome. *PLoS Comput. Biol.*, **9**, e1003003.
8. Pérez-Villa,A., Darvas,M. and Bussi,G. (2015) ATP dependent NS3 helicase interaction with RNA: insights from molecular simulations. *Nucleic Acids Res.*, **43**, 8725–8734.
9. Krepl,M., Havrila,M., Stadlbauer,P., Banas,P., Otyepka,M., Pasulka,J., Stefl,R. and Sponer,J. (2015) Can we execute stable microsecond-scale atomistic simulations of protein–RNA complexes? *J. Chem. Theory Comput.*, **11**, 1220–1243.
10. Laing,C. and Schlick,T. (2011) Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol.*, **21**, 306–318.
11. Condon,D.E., Kennedy,S.D., Mort,B.C., Kierzek,R., Yildirim,I. and Turner,D.H. (2015) Stacking in rna: NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput.*, **11**, 2729–2742.
12. Bergonzo,C., Henriksen,N.M., Roe,D.R. and Cheatham,T.E. (2015) Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA*, **21**, 1578–1590.
13. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 14664–14669.
14. Parisien,M. and Major,F. (2008) The MC-fold and MC-sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
15. Major,F., Turcotte,M., Gautheret,D., Lapalme,G., Fillion,E. and Cedergren,R. (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, **253**, 1255–1260.
16. Gautheret,D., Major,F. and Cedergren,R. (1993) Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol.*, **229**, 1049–1064.
17. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
18. Best,R.B., Lindorff-Larsen,K., DePristo,M.A. and Vendruscolo,M. (2006) Relation between native ensembles and experimental structures of proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10901–10906.
19. Morozov,A.V., Kortemme,T., Tsemekhman,K. and Baker,D. (2004) Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 6946–6951.
20. Butterfoss,G.L. and Hermans,J. (2003) Boltzmann-type distribution of side-chain conformation in proteins. *Protein Sci.*, **12**, 2719–2731.
21. Miao,Z., Adamiak,R.W., Blanchet,M.-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cheng,C., Chojnowski,G., Chou,F.-C., Cordero,P. *et al.* (2015) RNA-puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA*, **21**, 1066–1084.
22. Lee,J., Dethoff,E.A. and Al-Hashimi,H.M. (2014) Invisible RNA state dynamically couples distant motifs. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 9485–9490.
23. Lisi,V. and Major,F. (2007) A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence structure relationships. *RNA*, **13**, 1537–1545.
24. Brereton,A.E. and Karplus,P.A. (2015) Native proteins trap high-energy transit conformations. *Science Adv.*, **1**, e1501188.
25. Coifman,R.R., Lafon,S., Lee,A.B., Maggioni,M., Nadler,B., Warner,F. and Zucker,S.W. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7426–7431.
26. Rohrdanz,M.A., Zheng,W., Maggioni,M. and Clementi,C. (2011) Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, **134**, 124116.
27. Zheng,W., Qi,B., Rohrdanz,M.A., Caflisch,A., Dinner,A.R. and Clementi,C. (2011) Delineation of folding pathways of a β-sheet miniprotein. *J. Phys. Chem. B*, **115**, 13065–13074.
28. Bottaro,S., Di Palma,F. and Bussi,G. (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.*, **42**, 13306.
29. Noé,F., Schütte,C., Vanden-Eijnden,E., Reich,L. and Weikl,T.R. (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 19011–19016.
30. Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**, 922–923.
31. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
32. Chung,F.R. (1997) Spectral graph theory. *CBMS Regional Conference Series in Mathematics*. Vol. **92**, p. 212.
33. Scherer,M.K., Trendelkamp-Schroer,B., Paul,F., Pérez-Hernández,G., Hoffmann,M., Plattner,N., Wehmeyer,C., Prinz,J.-H. and Noé,F. (2015) PyEMMA 2: a Software package for estimation, validation, and analysis of Markov models. *J. Chem. Theory Comput.*, **11**, 5525–5542.
34. Ng,A.Y., Jordan,M.I. and Weiss,Y. (2002) On spectral clustering: analysis and an algorithm. *Adv. Neural. Inf. Process. Syst.*, **2**, 849–856.
35. Pronk,S., Páll,S., Schulz,R., Larsson,P., Bjelkmar,P., Apostolov,R., Shirts,M.R., Smith,J.C., Kasson,P.M., van derSpoel,D. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845–854.
36. Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
37. Pérez,A., Marchán,I., Svozil,D., Sponer,J., Cheatham,T.E. III, Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: Improving the description of αγ conformers. *Biophys. J.*, **92**, 3817–3829.
38. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
39. Joung,I.S. and Cheatham,T.E. III (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, **112**, 9020–9041.
40. Sugita,Y. and Okamoto,Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**, 141–151.
41. Tubbs,J.D., Condon,D.E., Kennedy,S.D., Hauser,M., Bevilacqua,P.C. and Turner,D.H. (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. *Biochemistry*, **52**, 996–1010.
42. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
43. Dellago,C., Bolhuis,P.G. and Chandler,D. (1998) Efficient transition path sampling: application to lennard-jones cluster rearrangements. *J. Chem. Phys.*, **108**, 9236–9245.
44. Hall,K.B. (2015) Mighty tiny. *RNA*, **21**, 630–631.
45. Tuerk,C., Gauss,P., Thermes,C., Groebe,D.R., Gayle,M., Guild,N., Stormo,G., d'AubentonCarafa,Y., Uhlenbeck,O.C. and Tinoco,I. (1988) Cuucgg hairpins: extraordinarily stable RNA secondary

structures associated with various biochemical processes. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 1364–1368.

46. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
47. Cheong,C., Varani,G. and Tinoco,I. (1990) Solution structure of an unusually stable RNA hairpin, 5GGAC (UUCG) GUCC. *Nature*, **346**, 680–682.
48. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
49. Kuhrova,P., Banas,P., Best,R.B., Sponer,J. and Otyepka,M. (2013) Computer folding of RNA tetraloops? are we there yet? *J. Chem. Theory Comput.*, **9**, 2115–2125.
50. Sherrer,R.L., Araiso,Y., Aldag,C., Ishitani,R., Ho,J.M., Söll,D. and Nureki,O. (2011) C-terminal domain of archaeal o-phosphoseryl-tRNA kinase displays large-scale motion to bind the 7-bp d-stem of archaeal tRNAsec. *Nucleic Acids Res.*, **39**, 1034–1041.
51. Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group i catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
52. Heus,H.A. and Pardi,A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science*, **253**, 191–194.
53. Westhof,E., Romby,P., Romaniuk,P.J., Ebel,J.-p., Ehresmann,C. and Ehresmann,B. (1989) Computer modeling from solution data of spinach chloroplast and of xenopus laevis somatic and oocyte 5 s rrnas. *J. Mol. Biol.*, **207**, 417–431.
54. Salmon,L., Bascom,G., Andricioaei,I. and Al-Hashimi,H.M. (2013) A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. *J. Am. Chem. Soc.*, **135**, 5457–5466.
55. Emani,P.S., Bardaro Jr,M.F., Huang,W., Aragon,S., Varani,G. and Drobny,G.P. (2014) Elucidating molecular motion through structural and dynamic filters of energy-minimized conformer ensembles. *J. Phys. Chem. B*, **118**, 1726–1742.
56. Chodera,J.D., Singhal,N., Pande,V.S., Dill,K.A. and Swope,W.C. (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, **126**, 155101.
57. Frellsen,J., Moltke,I., Thiim,M., Mardia,K.V., Ferkinghoff-Borg,J. and Hamelryck,T (2009) A probabilistic model of RNA conformational space. *PLoS Comput. Biol.*, **5**, e1000406.
58. Ma,H., Proctor,D.J., Kierzek,E., Kierzek,R., Bevilacqua,P.C. and Gruebele,M. (2006) Exploring the energy landscape of a small RNA hairpin. *J. Am. Chem. Soc.*, **128**, 1523–1530.
59. Sarkar,K., Meister,K., Sethi,A. and Gruebele,M. (2009) Fast folding of an RNA tetraloop on a rugged energy landscape detected by a stacking-sensitive probe. *Biophys. J.*, **97**, 1418–1427.
60. Chen,A.A. and García,A.E. (2013) High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 16820–16825.
61. Zhao,L. and Xia,T. (2007) Direct revelation of multiple conformations in RNA by femtosecond dynamics. *J. Am. Chem. Soc.*, **129**, 4118–4119.
62. Jucker,F.M., Heus,H.A., Yip,P.F., Moors,E.H. and Pardi,A. (1996) A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.*, **264**, 968–980.