The Journal of Chemical Physics

The mechanism of RNA base fraying: Molecular dynamics simulations analyzed with core-set Markov state models

Cite as: J. Chem. Phys. **150**, 154123 (2019); https://doi.org/10.1063/1.5083227
Submitted: 29 November 2018 . Accepted: 28 March 2019 . Published Online: 19 April 2019

Giovanni Pinamonti, Fabian Paul, Frank Noé, Alex Rodriguez, and Giovanni Bussi

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

The mechanism of RNA base fraying: Molecular dynamics simulations analyzed with core-set Markov state models

Cite as: J. Chem. Phys. 150, 154123 (2019); doi: 10.1063/1.5083227
Submitted: 29 November 2018 • Accepted: 28 March 2019 •
Published Online: 19 April 2019

Giovanni Pinamonti,[1,a)] Fabian Paul,[2] Frank Noé,[1] Alex Rodriguez,[3] and Giovanni Bussi[4,b)]

AFFILIATIONS

[1]Department for Mathematics and Computer Science, Freie Universität, Berlin, Germany
[2]Department of Biochemistry and Molecular Biology, Gordon Center for Integrative Science, The University of Chicago, Chicago, Illinois 60637, USA
[3]ICTP, International Centre for Theoretical Physics, Trieste, Italy
[4]Scuola Internazionale Superiore di Studi Avanzati, via Bonomea 265, Trieste, Italy

Note: This article is part of the Special Topic "Markov Models of Molecular Kinetics" in J. Chem. Phys.
[a)]**Electronic mail:** giovanni.pinamonti@fu-berlin.de
[b)]**Electronic mail:** bussi@sissa.it

ABSTRACT

The process of RNA base fraying (i.e., the transient opening of the termini of a helix) is involved in many aspects of RNA dynamics. We here use molecular dynamics simulations and Markov state models to characterize the kinetics of RNA fraying and its sequence and direction dependence. In particular, we first introduce a method for determining biomolecular dynamics employing core-set Markov state models constructed using an advanced clustering technique. The method is validated on previously reported simulations. We then use the method to analyze extensive trajectories for four different RNA model duplexes. Results obtained using D. E. Shaw research and AMBER force fields are compared and discussed in detail and show a non-trivial interplay between the stability of intermediate states and the overall fraying kinetics.

I. INTRODUCTION

Ribonucleic acid (RNA) plays a fundamental role in the biology of the cell.[1] RNA molecules fold in intricate structures, that undergo complex rearrangements,[2] to fulfill a number of biological functions, such as gene regulation, splicing, catalysis, and protein synthesis. It is thus key to get a more precise understanding of the mechanisms involved in RNA folding and conformational transitions. Current experimental techniques are limited to ensemble measurements or to low spatiotemporal resolution. For this reason, computational tools are fundamental for the study of biomolecular systems, including ribonucleic acids. Molecular dynamics (MD) simulations using empirical force fields, propelled by numerous theoretical and technical improvements,[3–7] have enabled scientists to accurately study the thermodynamics and kinetics of proteins[8–11] and nucleic acids.[12–15] In particular, the framework of Markov state models (MSMs)[16–20] makes it possible to perform a systematic analyses of the metastable states and kinetics of biomolecular systems. In principle, these computational tools could be used as a highly accurate "computational microscope,"[21] allowing the quantitative description of the individual steps leading to, e.g., the rupture and formation of a double helix. In practice, the results that can be obtained for RNA molecules are still limited by several factors, the most important of which being the accuracy of the force fields employed.[14,22–24]

The process of "base fraying," that is, the breaking of base pairing and stacking interactions at the termini of a RNA (or DNA) double helix, is an apparently simple yet far from trivial process. Frayed states are intermediate in the RNA zipping and unzipping processes, have been proposed to be important in the interaction of RNA with proteins (see, e.g., Refs. 25–28), and might be

relevant in strand invasion[29] and, in general, in secondary structure rearrangements required for the riboswitch function.[30] The characterization of fraying kinetics by experimental techniques is, however, difficult due to their short lifetimes.[31,32] Base fraying in RNA has been characterized by means of computer simulations in several studies.[27,33,34] Colizzi and Bussi[27] characterized the thermodynamics of the process, suggesting that dangling bases at the 3′-end are more stable than those at the 5′-end and thus might be important intermediates in duplex unzipping. This finding is in agreement with the higher stabilization provided to duplexes by 3′-end dangling bases when compared to 5′-end dangling bases,[35] although this latter quantity also depends on the energy of stacking in single stranded RNAs. The stabilities computed by Colizzi and Bussi were, however, largely overestimated by the adopted unidirectional pulling and were thus only usable to rank the fraying propensity of different sequences. Zgarbová *et al.*[33] performed a detailed characterization of the non-canonical structures observed with current force fields at the termini of DNA and RNA duplexes, without however, aiming at obtaining quantitative populations. Both these studies did not explicitly analyze the kinetics of the process. Finally, Xu *et al.*[34] presented a partial kinetic model reproducing the opening of the base on the 5′ terminus of a RNA duplex. However, a comparison of the kinetics of the two ends and a quantitative analysis of its sequence dependence are still missing. A number of papers addressed fraying in DNA (see, e.g., Refs. 36–38 and the already mentioned Ref. 33) or in an RNA:DNA hybrid in complex with a protein.[28]

We here employ extensive MD simulations using 4 different sequences with the goal of quantitatively characterizing the fraying kinetics, through the use of MSMs. Specifically, we extend the core-based MSM framework[39,40] with density-based clustering.[41,42] The procedure is validated on the kinetics of short oligonucleotides first and then applied to base fraying in RNA duplexes. We chose different sequences in order to assess the effect of the position of purines/pyrimidines and the influence of the neighboring base pair. Two state-of-the-art force fields are compared, specifically (i) the one recently published by the D. E. Shaw research laboratory[43] (in the following, DESRES) and (ii) the latest refinement of the AMBER force field,[44] which is the default AMBER force field for RNA systems. Both force fields are based on previous versions of the AMBER force field.[45,46] We studied the sequence dependence of stability, fraying rate, and the different pathways which the process can follow. Interestingly, the results display a non-trivial interplay between the stability of intermediate structures and the kinetics of the process.

II. METHODS

A. Molecular dynamics simulations

We simulated the dynamics of short helices composed of a 3 base-pair GC stem plus a GC terminal pair. In particular, the following four permutations were used as model constructs: $^{5'-\mathrm{ACGC}}_{3'-\mathrm{UGCG}}$, $^{5'-\mathrm{AGCG}}_{3'-\mathrm{UCGC}}$, $^{5'-\mathrm{UCGC}}_{3'-\mathrm{AGCG}}$, and $^{5'-\mathrm{UGCG}}_{3'-\mathrm{ACGC}}$. In the rest of the paper, we will refer to these constructs using only the sequence of the strand fraying at its 5′-end, respectively, ACGC, AGCG, UCGC, UGCG.

Initial structures were obtained using the Make-NA server (http://structure.usc.edu/make-na/). RNA duplexes were solvated in explicit water, adding Na^+ counterions to neutralize the RNA charge, plus additional NaCl to reach the nominal concentration of 0.1M. The system was inserted in a truncated dodecahedral box with periodic boundary conditions and box size 5.17 nm. RNA was described using either the DESRES force field[43,45,46] with the TIP4P-D water model[47] or the AMBER force field[44–46] with the TIP3P water model.[48] Ions in DESRES simulations were described using the CHARMM parameters[49] as recommended in Ref. 43, whereas in AMBER simulations, they were described using AMBER-adapted parameters for Na^+[50] and Cl^-.[51] Based on previous results, we do not expect RNA dynamics to be highly affected by the ion parameters at this concentration.[52,53] The equations of motion were integrated with a 2 fs time step. All bond lengths were constrained using the LINCS algorithm.[54] Long-range electrostatics was treated using particle-mesh-Ewald summations.[55] Trajectories were generated in the isothermal-isobaric ensemble using stochastic velocity rescaling[56] and the Parrinello-Rahman barostat.[57] All simulations were performed using GROMACS (version 4.6.7, calculations using the AMBER force field, and 5.1.2, calculations using the DESRES force field). Force field parameters can be found at https://github.com/srnas/ff. Since we decided to focus the study on the fraying of the A-U terminal pair, we restrained the distances between the heavy atoms involved in the hydrogen bonds corresponding to the G-C pairs, using harmonic potentials. Additional details of the simulations are given in the supplementary material.

For each system, we ran 32 independent simulations, each approximately 1.0–1.5 μs long. After an initial energy minimization using a steepest descent algorithm, 32 independent simulations were initialized with random seeds and simulated for 100 ps at T = 400 K, then equilibrated for additional 100 ps at T = 300 K. The final configurations were used as starting points for the production runs. The simulations using the DESRES and AMBER force fields were performed starting from exactly the same conformations. Frames were stored for later analysis every 100 ps. The minimum distance observed between solute atoms from periodic images was 1.55 nm (DESRES simulations). Stacking interactions were analyzed by using both the stacking score[58] and the so-called G-vectors introduced in Ref. 59. We analyzed the trajectories using Barnaba[60] and MDTraj.[61]

B. Core Markov state model combined with density-based clustering

MSMs have been successfully applied to the study of many biomolecular systems (see, e.g., Refs. 11 and 62–66). The idea underlying an MSM is to reduce the complexity of a simulation by partitioning the phase space into discrete microstates via a clustering algorithm. The transition probabilities between these microstates can be then computed counting the transitions observed in the MD trajectories.

A possible approach to compute these probabilities is the so-called "transition-based-assignment" or "coring," first proposed in Ref. 39 and further analyzed in Ref. 40. The idea is to define a collections of "core sets," i.e., metastable regions of the phase space, which are not required to be in contact among each other. A transition between states A and B is counted only when a trajectory goes from the core region of A ($\mathscr{C}_A$) to the core region of B ($\mathscr{C}_B$)

without passing through any other core region. Then, the system will be considered in state B until it goes back to $\mathscr{C}_A$ or reaches a third core region, independently of how many times it exits and re-enters in $\mathscr{C}_B$ before reaching a new state.

The fundamental step of this approach is to start with a good definition of metastable core sets. This requirement is usually in contrast with the fact that, when studying the dynamics of a complex biomolecule, no prior knowledge of the free-energy landscape of the system is available. Therefore, in order to successfully apply this method, it is necessary to extract this information from the simulation data, preprocessing the trajectories in order to identify different states and define realistic core regions. A smart way to do this is to make use of a density-based clustering algorithm to separate the MD data set into a collection of clusters and identify the core regions of these clusters as the regions with higher density.[67]

To construct core-based MSMs we proceed as follows. We started by describing the system using the same set of coordinates that we employed in a previous study:[62] (i) G-vectors (4D vectors connecting the nucleobases ring centers, as described in Ref. 59), and (ii) the sine and cosine of backbone dihedrals, sugar ring torsional angles, and glycosidic torsional angles. The dimensionality of this input was then reduced using time-lagged independent component analysis (TICA)[68] with a lag time of 5 ns, and data were projected on the slowest TICs using a kinetic map projection.[69] Subsequently, we used the pointwise-adaptive k-free energy estimator (PAk) algorithm[70] combined with the TWO-NN algorithm[71] to estimate the pointwise density in TICA space, which was then used to cluster the data using density peak clustering.[41,42] We defined the core of each cluster as the set of all points i for which $\rho_i/\rho_{MAX} > e^{-1}$, where ρ_{MAX} is the maximum density in the cluster. According with Rodriguez *et al.*,[70] this corresponds approximately to a maximum of 1 $k_B T$ free-energy difference between the configurations included in the core set and those belonging to the transition areas. Finally, the MD trajectories were discretized by assigning each frame to the last core set visited, and the resulting discrete trajectories were used to estimate a reversible MSM.[72,73] More details on the procedure are given in the supplementary material.

This procedure leads to robust and reliable MSMs. In order to validate this procedure, we compared the results with those of a standard MSM in which the phase space was discretized using k-means clustering on TICA projected space, and a transition was counted every time a trajectory jumped from one microstate to the other. Results of this validation are given in Sec. III A.

Afterwards, a lag time τ = 100 ps was then used to construct a core-based MSM that approximates the dynamics of the discretized system. The quality of the Markovian approximation was tested by looking at the convergence of the implied time scales predicted by the MSM for increasing values of τ as described in Ref. 17.

The MSM construction and analysis was performed using the software PyEMMA 2.2.[74] Density peak clustering was performed using the code available at https://github.com/alexdepremia/Advanced-Density-Peaks.

C. Classification of states

In order to obtain an easy-to-interpret representation of the fraying kinetics, we classified the microstates obtained with the MSM procedure in different groups. The classification was performed using a number of structural determinants, including root-mean-square deviation (RMSD) from native conformation[75] and stacking score.[58]

For each system, microstates were grouped into the following states:

- closed (C): canonical double helix, with both terminal bases in their native conformations, stacking on the adjacent G or C base, and forming pairing interactions between each other;
- open (O): frayed structures, with broken pairing between the two terminal bases, which are both unstacked and freely moving;
- 5′-open ($5P$): the base at the 5′ terminal base is not forming any stacking and pairing interactions, while the base in 3′ is still in its native conformation;
- 3′-open ($3P$): same as $5P$, but inverting 5′ and 3';
- misfolded (M): the base on the 3′ terminus is rotated by 180° and stacking "upside down" on its adjacent base; and
- undefined (U): all conformations not falling into the previous categories, including, among others, microstates where the base at the 5′ terminus is rotated upside down, or configurations in which one of the two terminal bases stacks on the top of a base in the opposite strand.

Technical details of this classification are reported in the supplementary material.

III. RESULTS

A. Validation of the core-based MSM

As a first step, we performed a validation of the introduced MSM procedure using core-sets obtained with the PAk algorithm and DP clustering. In particular, we here analyzed trajectories reported in a previous paper[62] for RNA adenine di- and tri-nucleotides. Details of this analysis are provided in Sec. II B and in the supplementary material. Figure 1(a) reports a comparison between the results obtained with a core-set MSM and those obtained with a standard MSM approach, as described in Sec. II. Specifically, the time scales as a function of the lag time τ are shown, and we can see that the core-based MSM lead to time scales fully compatible with the standard approach. Strikingly, the time scales are basically independent of the chosen lag time, showing that this procedure is extremely robust and allows the selection of a relatively short lag time for the MSM construction. We also notice (See Table SI 1) that the number of clusters resulting from the DP clustering is consistently smaller than the number of microstates that are required for a good discretization using k-means.

B. Energetic and kinetic analysis

After being validated, the method is used to analyze large scale simulations of 4 short duplexes, consisting in 32 simulations, with a total simulation time of 35–54 μs for each sequence (see Table SI 1 for details). Two different force fields were employed. We here report results using the DESRES force field,[43] whereas results using the standard AMBER force field are presented in Sec. III C.

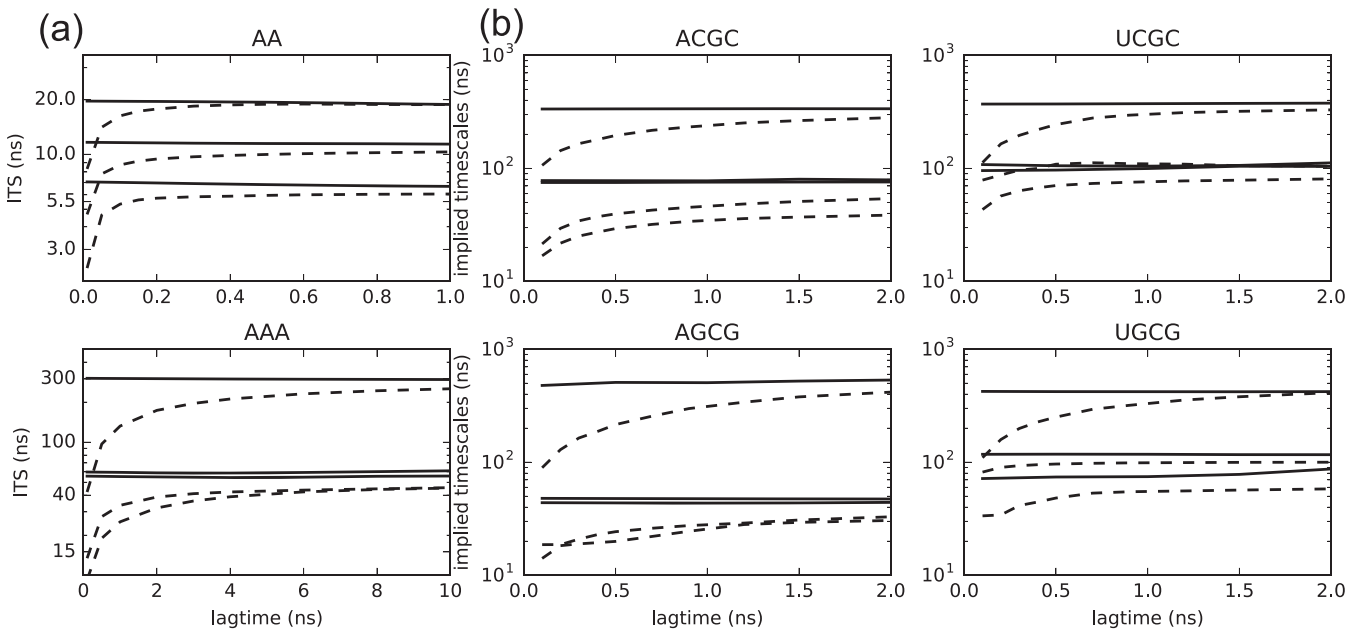

FIG. 1. Implied time scales of the MSM for different systems, as a function of the lag time. The three slowest time scales obtained with the core-based MSM (continuous lines) are compared with a standard MSM with k-mean clustering (dashed lines). Panel (a) shows the validation on the adenine di- and trinucleotide simulations taken from Ref. 62. Panel (b) reports the results for the 4 RNA duplexes studied in this work.

From the equilibrium population of the microstates obtained from the MSM, we computed the free-energy difference between the *O* and *C* states. Table I reports the computed difference in free energy between the closed (*C*) state and the open (*O*) one. The native structure is the most stable one, as expected. The stability of the closed structure can be compared with thermodynamic experiments[35] where the stabilization of a duplex due to the presence of an additional base pair is measured. The negative of this number provides a lower bound for the stacking energy. Indeed, by assuming that unstacked nucleobases do not form any interaction, the contribution of an additional base pair to the stability of a duplex (ΔF_{stab}) can be approximated as

$$\Delta F_{\text{stab}} \approx \Delta F^{ss_1}_{S\to U} + \Delta F^{ss_2}_{S\to U} - \Delta F_{C\to O}. \quad (1)$$

Here, $\Delta F^{ss_1}_{S\to U}$ and $\Delta F^{ss_2}_{S\to U}$ are the free-energy changes related to breaking a terminal stacking interaction in the two individual single strands. Since single strands are expected to display a significant amount of stacking[76] these terms are expected to be positive so that $-\Delta F_{\text{stab}} < \Delta F_{C\to O}$. The ranking of the four investigated systems is thus qualitatively consistent.

We also computed the stability of the intermediate states where only one of the two nucleobases is open and the other is stacked (3′-open, 3*P*, and 5′-open, 5*P*, if the stacked nucleobase is at the 5′ or 3′ end of the helix, respectively). The relative ΔF, with respect to state *C*, of the 3*P* and 5*P* intermediate states are reported in Table SI 3–6. As expected, adenine (purine) terminal bases form stacking interactions that are stronger when compared with uracils (pyrimidines). Moreover, we find that the 5′-open states are on average more stable than the 3′-open states, although the difference is modulated by the sequence. In particular, the two stabilities are roughly comparable when the purine (A) is located at the 5′-end, whereas the 5′-open state is significantly more stable when the purine is located at the 3′-end. These results are qualitatively consistent with previous findings.[27]

Finally, states *U* and *M*, where one or both nucleotides are not in their native structure nor unstacked, appear with non-negligible population. Their stabilities are significantly smaller than that of the native closed structure. We are not aware of solution experiments

TABLE I. Thermodynamic and kinetic properties obtained from the MSMs of the 4 RNA duplexes, specifically: Free-energy difference between *O* and *C* states ($\Delta F_{C\to O}$) negative experimental stabilization of each duplex by the terminal base pair computed with nearest neighbors parameters ($-\Delta F_{\text{stab}}$);[35] MFPTs from *C* to *O*, and vice versa; slowest implied time scale (t_1) obtained from the MSM. Energies are expressed in kcal/mol. Notice that the negative experimental stabilization is by construction expected to be smaller than the stacking energy (see the text for discussion). Uncertainties were estimated by using a Jackknife procedure.

	$\Delta F_{C\to O}$	$-\Delta F_{\text{stab}}$	MFPT (μs)		
Seq.	MSM	Experiment	$C\to O$	$O\to C$	t_1 (ns)
ACGC	3.7 ± 0.05	2.0	10 ± 1	0.28 ± 0.01	336 ± 12
AGCG	3.1 ± 0.15	1.8	10 ± 4	0.47 ± 0.03	479 ± 31
UCGC	4.6 ± 0.02	2.1	52 ± 6	0.36 ± 0.05	371 ± 57
UGCG	2.9 ± 0.07	1.8	10 ± 1	0.29 ± 0.01	424 ± 21

that rule out these structures as possible alternatives. Focusing on state *M*, which consists of a clearly defined ensemble of conformations, we tried to search structures similar to these ones[59] within the whole structural database using Barnaba.[60] Although fragments extracted from the database are expected to be highly biased due to their structural context and to the variety of experimental conditions under which they were obtained, they were shown to agree with solution experiments to a significant extent both in proteins[77] and nucleic acids.[78] A significant number of fragments with virtually identical base pairing can be found (see Table SI 6), suggesting that these misfolded structures are plausible metastable states.

We then used the obtained MSMs to characterize the kinetics of fraying. Interestingly, the slowest process always corresponds to the unstacking and rotation of the nucleobase at the 3′-end [see Fig. 2(a)], and thus represents the interconversion to the misfolded structure mentioned above. The time scales of this process for the four different systems are also reported in Table I. We then computed the fraying kinetics for the four systems using transition-path-theory (TPT),[79] in the formulation of MSMs.[80] The mean-first-passage time (MFPT) associated with the fraying transition for the four systems is reported in Table I. This number is inversely proportional to the fraying rate. Interestingly, the MFPT for ACGC, AGCG, and UGCG are very similar (10 μs) to each other, while UGCG exhibits a much smaller fraying rate. The MFPT for the inverse process, i.e., terminal pairing, is also reported in Table I. We can see that this quantity shows a small dependence on the sequence and is correlated with the time scale of the slowest process, i.e., the rotation of the 3′ nucleobase, showing that the state *M* acts as a kinetic trap during terminal pairing.

We further investigated the mechanism of fraying, focusing on the first opening base. Using TPT, we obtained the flux of fraying trajectories going from *C* to *O* through either 3*P* or 5*P*. Results are reported in Fig. 2(b). The most likely path for fraying is, for the four investigated systems, the one through the 3*P* intermediate. In other words, based on these results one would expect the nucleobase at the 3′-end to most likely break its stacking interaction with the adjacent base before the one at the 5′-end. Interestingly, the most probable intermediate in the transition between *C* and *O* is always 3*P*, even in sequences UCGC and UGCG where it is the "least" stable one according to the free-energy analysis reported above. This can be rationalized by the individual rates reported in Table SI 3–6. Estimated rates for transitions from 5*P* to *O* are either null or very small. This is a consequence of the fact that no or few transitions are observed in the MD simulation along this pathway.

C. Comparison with AMBER force-field

We also analyzed an identical set of simulations performed using the latest AMBER force field.[44] Results are reported in Fig. SI 2 and Table SI 7. These simulations resulted in a larger number of non-canonical structures when compared with those obtained in the simulations performed with the DESRES force field. In particular, we observed a non-negligible population of so-called ladder-like structures[44] (See Fig. SI 2B). Similar structures were also observed in previous studies[22,81] and might be a consequence of both the short length of the duplex simulated here and the presence of restraints on the base-pair distances in the duplex.

In addition, our simulations show a large population of misfolded structures. See the supplementary material for more details about these structures. In particular, for all the constructs except UCGC the stability of the misfolded structures was larger than that of the native structure. Whereas these conformations are plausible metastable states (see also Table SI 6), their high populations make the results much more difficult to interpret.

In general, the closed state is less stable with respect to the open state. This can be seen both from the ΔF, which are smaller in general, and from the shorter MFPTs (See Fig. SI 2A). Regarding the predicted fraying mechanics, most of the fraying pathways are going through the misfolded (*M*), the undefined (*U*), or the ladder-like (*L*)

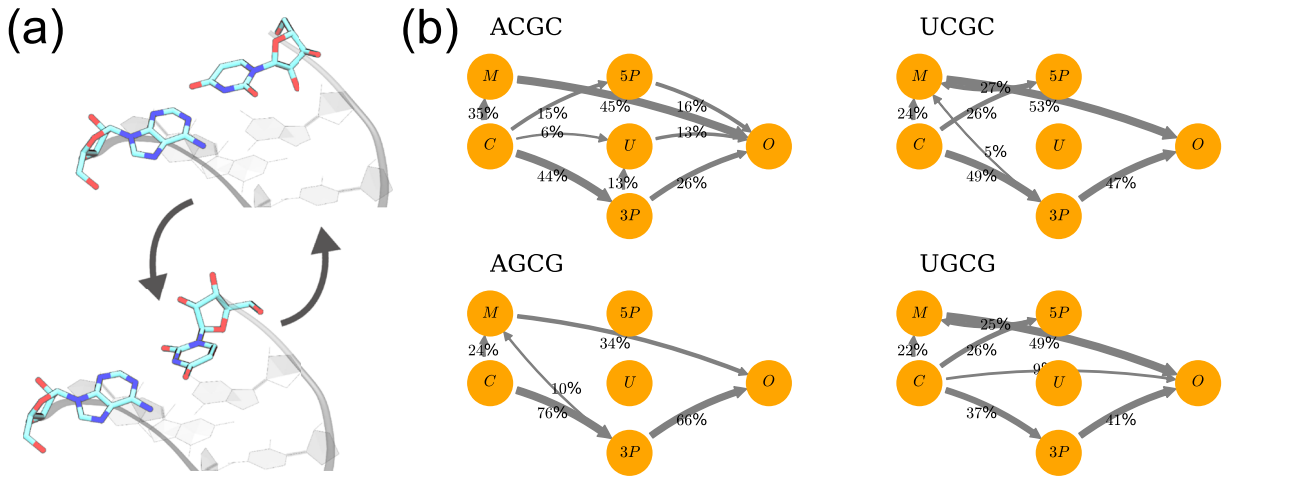

FIG. 2. Results of the MSMs of the four sequences, based on DESRES simulations. Panel (a): Slowest process in the MSMs of the 4 duplexes. Two structures from the simulations of ACGC are shown, as representatives for all sequences. Panel (b): Flux of fraying trajectories computed from the four MSM by means of TPT.

states. This makes it difficult to reach a definite conclusion regarding the mechanism. However, one can observe a general tendency for a mechanism where the 5′ base opens before the 3′ one, in contrast with what predicted from the simulations with the DESRES force field.

IV. DISCUSSION AND CONCLUSIONS

In this work, we developed a robust recipe for the construction of core-based MSMs and applied it to the characterization of fraying kinetics in RNA. When compared with standard MSMs the core-based method enables to obtain MSMs with a limited number of microstates, making the following analysis both clearer and more practical. At the same time, the implied time scales are robustly estimated, even for very short values of the lag time. One of the advantages of a short lag time is a statistically robust estimation of the MFPT, which is in general a challenging task for an MSM with a large lag time, due to the effect of recrossing events.

We applied the introduced core-based MSM to study the thermodynamics and kinetics of base fraying and analyzed the pathway followed during the process. We first focused on the free-energy difference between the native helical conformation and the frayed state using the DESRES force field.[43,45,46] This difference can be compared with the stabilization of a duplex resulting from the addition of an individual base pair, as obtained from optical melting experiments.[35] The ranking of the four analyzed sequences is qualitatively consistent with the experiments. This result is by itself not obvious, given that we are comparing systems with the same numbers of GC and AU pairs. In other words, the force field is capable to qualitatively capture the difference between placing a purine or a pyrimidine on each of the two strands, and the interplay between the hydrogen bonds formed in the first and in the second base pair of a helix. The comparison is, however, only qualitative since the experimental free energies report the difference in the stability of two duplexes with a different number of pairs. By closing the thermodynamic cycle, the experimental free energies should correspond to the difference between the stacking energy in a duplex and the stacking energy in two separated single strands that are not included here. Previous studies performing the full thermodynamic cycle using an older version of the AMBER force field[82] and the version used here[83] report agreement with experimental free energies. In addition, optical melting experiments are not sensitive to the precise structure and only report the overall stability of a bimolecular complex that might originate from the combination of different structures. It must be also observed that, although the DESRES force field is the only force field to date that was shown to be able to predict the folded structure of RNA tetraloops including their signature interactions,[43] its capability to reproduce experimentally observed non-canonical interactions has been recently questioned.[84]

We then focused on the kinetics of the fraying process, estimating the fraying rate and the weight of different pathways. The fraying rates of three of the four sequences are all around 10^5 s^{-1}. The fourth sequence, UCGC, displays a 5-times slower fraying. This effect can be attributed to the larger stability of the UC and AG stacking interactions, also observed in the thermodynamic parameters. The reported rates are in qualitative agreement with those measured using imino-proton exchanges.[31] We observe that rates for terminal pairing show a slight dependence with respect to the sequence. The ratio between the lowest and highest rate is ≈1.7, corresponding to a contribution of ≈0.3 kcal/mol to the sequence-dependence of the stability of individual base pairs. In comparison, the ratio between the highest and the lowest unpairing rate is ≈5, corresponding to a contribution of ≈1 kcal/mol. This agrees with the common notion that the stability differences depend more on the off rates than on the on rates.[85]

Interestingly, the fraying path with the largest flux always corresponds to a 3′-open intermediate, which is typically the least stable among the two intermediates. The opening of the 5′ end likely results in a reclosure of the pair because the 3′ dangling end stays stacked. The opening of the 3′ end can result in complete opening because the 5′ dangle is less stable. Therefore, a 3′ end opening, although less frequent, is more likely to proceed to complete opening. The 5′-open path has been proposed as the most likely one based on the frequency of 3′-dangling bases in crystal structures[86] and on the relative stability of the two intermediates as computed by molecular simulations.[27] It is, however, important to underline that the results reported here might be affected by the choice of the force field. In particular, we are not aware of any validation of the kinetics reported so far for the DESRES RNA force field.

Finally, we report a comparison with simulations performed with the standard AMBER force field.[44–46] Results are significantly different in the stability of the canonical duplexes, in the estimated fraying rates, and in the predicted pathways. In particular, the AMBER force field generates a larger population of non-canonical, misfolded structures. Whereas the observed ladder-like structures[44] might be an artifact related to the short length of the simulated helices, the non-canonical interactions at the terminal bases have been already reported elsewhere in the context of longer duplexes[33] and are probably intrinsically stabilized by this force field. The relative population of these structures might change using recent corrections that aimed at providing a better balance between important hydrogen bond interactions.[84] Overall, the presence of these structures makes the interpretation of the fraying process more difficult.

Interestingly, whereas the relative stability of the different intermediates exhibits a similar trend between different force-fields, the pathways of fraying are substantially different. Indeed, it is known that the impact of force fields on kinetics can be large even when the thermodynamic properties are similar.[87] In this particular case, the different kinetics might be a consequence of different energetic barriers (e.g., in the backbone torsional angles) that are difficult to validate experimentally. The possibility to have different force fields resulting in the same native structure but predicting different pathways have been discussed for instance in the case of protein folding simulations.[88]

In conclusion, we constructed a core-based MSM with the goal of reproducing the kinetic properties of the terminal base pair of an RNA double helix. The introduced method makes it possible to obtain a robust estimation of rates and MFPT between folded and frayed structures to identify metastable states and their relative stabilities, as well as the unzipping pathways followed by the system. Although the obtained rates are in qualitative agreement with experimental data, the appearance of non-canonical structures and/or their excessive suppression make the preferential unzipping pathway dependent on the chosen force field.

SUPPLEMENTARY MATERIAL

Supplementary material contains details of MD and MSM construction, definition of states, supplementary results of kinetic analysis, results of structural database search, and results of AMBER force field simulations.

ACKNOWLEDGMENTS

G.P. and G.B. received support by the European Research Council, Starting Grant No. 306662. F.N. and F.P. acknowledge funding from European Research Commission (Nos. ERC StG 307494 "pcCell" and ERC CoG 772230 "ScaleCell") and Deutsche Forschungsgemeinschaft (Nos. SFB1114/C03 and SFB1114/A03). F.P. acknowledges funding from the Yen Post-Doctoral Fellowship in Interdisciplinary Research and from the National Cancer Institute of the National Institutes of Health (NIH) through Grant No. CAO93577. We thank Alessandro Laio, Maria d'Errico, and Elena Facco, for their help in the clustering analysis. We are grateful to Sandro Bottaro for many helpful advices and for interesting scientific discussions. We also thank the anonymous reviewers whose comments have greatly improved this manuscript.

REFERENCES

[1] K. V. Morris and J. S. Mattick, "The rise of regulatory RNA," Nat. Rev. Genet. **15**, 423 (2014).
[2] H. M. Al-Hashimi and N. G. Walter, "RNA dynamics: It is about time," Curr. Opin. Struct. Biol. **18**, 321–329 (2008).
[3] G. R. Bowman, V. S. Pande, and F. Noé, *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer Science & Business Media, 2013), Vol. 797.
[4] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L.-S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y.-H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. Ben Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young, "Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, *SC'14* (IEEE Press, Piscataway, NJ, USA, 2014), pp. 41–53.
[5] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," Annu. Rev. Phys. Chem. **67**, 159–184 (2016).
[6] V. Mlýnský and G. Bussi, "Exploring RNA structure and dynamics through enhanced sampling simulations," Curr. Opin. Struct. Biol. **49**, 63–71 (2018).
[7] C. Camilloni and F. Pietrucci, "Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems," Adv. Phys.: X **3**, 1477531 (2018).
[8] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, "Long-timescale molecular dynamics simulations of protein structure and function," Curr. Opin. Struct. Biol. **19**, 120–127 (2009).
[9] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande, "To milliseconds and beyond: Challenges in the simulation of protein folding," Curr. Opin. Struct. Biol. **23**, 58–65 (2013).
[10] J. Preto and C. Clementi, "Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics," Phys. Chem. Chem. Phys. **16**, 19181–19191 (2014).
[11] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling," Nat. Chem. **9**, 1005–1011 (2017).
[12] S. Vangaveti, S. V. Ranganathan, and A. A. Chen, "Advances in RNA molecular dynamics: A simulator's guide to RNA force fields," Wiley Interdiscip. Rev.: RNA **8**, e1396 (2017).
[13] L. G. Smith, J. Zhao, D. H. Mathews, and D. H. Turner, "Physics-based all-atom modeling of RNA energetics and structure," Wiley Interdiscip. Rev.: RNA **8**, e1422 (2017).
[14] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, and M. Otyepka, "RNA structural dynamics as captured by molecular simulations: A comprehensive overview," Chem. Rev. **118**, 4177–4338 (2018).
[15] P. D. Dans, D. Gallego, A. Balaceanu, L. Darré, H. Gómez, and M. Orozco, "Modeling, simulations, and bioinformatics at the service of rna structure," Chem **5**, 51–73 (2019).
[16] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, "A direct approach to conformational dynamics based on hybrid Monte Carlo," J. Comput. Phys. **151**, 146–168 (1999).
[17] W. C. Swope, J. W. Pitera, and F. Suits, "Describing protein folding kinetics by molecular dynamics simulations. 1. Theory," J. Phys. Chem. B **108**, 6571–6581 (2004).
[18] F. Noé, I. Horenko, C. Schütte, and J. C. Smith, "Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states," J. Chem. Phys. **126**, 155102 (2007).
[19] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, "Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics," J. Chem. Phys. **126**, 155101 (2007).
[20] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," J. Chem. Phys. **134**, 174105 (2011).
[21] E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, "Discovery through the computational microscope," Structure **17**, 1295–1306 (2009).
[22] C. Bergonzo, N. M. Henriksen, D. R. Roe, and T. E. Cheatham, "Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields," RNA **21**, 1578–1590 (2015).
[23] P. Kuhrova, R. B. Best, S. Bottaro, G. Bussi, J. Sponer, M. Otyepka, and P. Banáš, "Computer folding of RNA tetraloops: Identification of key force field deficiencies," J. Chem. Theory Comput. **12**, 4534–4548 (2016).
[24] S. Bottaro, P. Banáš, J. Sponer, and G. Bussi, "Free energy landscape of gaga and uucg RNA tetraloops," J. Phys. Chem. Lett. **7**, 4032–4038 (2016).
[25] M. Betterton and F. Jülicher, "Opening of nucleic-acid double strands by helicases: Active versus passive opening," Phys. Rev. E **71**, 011904 (2005).
[26] J. F. Sydow, F. Brueckner, A. C. Cheung, G. E. Damsma, S. Dengl, E. Lehmann, D. Vassylyev, and P. Cramer, "Structural basis of transcription: Mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA," Mol. Cell **34**, 710–721 (2009).
[27] F. Colizzi and G. Bussi, "RNA unwinding from reweighted pulling simulations," J. Am. Chem. Soc. **134**, 5173–5179 (2012).
[28] L.-T. Da, F. Pardo-Avila, L. Xu, D.-A. Silva, L. Zhang, X. Gao, D. Wang, and X. Huang, "Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue," Nat. Commun. **7**, 11244 (2016).
[29] W. Huang, J. Kim, S. Jha, and F. Aboul-Ela, "The impact of a ligand binding on strand migration in the sam-I riboswitch," PLoS Comput. Biol. **9**, e1003069 (2013).
[30] A. Serganov and E. Nudler, "A decade of riboswitches," Cell **152**, 17–24 (2013).
[31] K. Snoussi and J.-L. Leroy, "Imino proton exchange and base-pair kinetics in RNA duplexes," Biochemistry **40**, 8898–8904 (2001).
[32] J. D. Liu, L. Zhao, and T. Xia, "The dynamic structural basis of differential enhancement of conformational stability by 5′-and 3′-dangling ends in RNA," Biochemistry **47**, 5962–5975 (2008).
[33] M. Zgarbová, M. Otyepka, J. Šponer, F. Lankas, and P. Jurečka, "Base pair fraying in molecular dynamics simulations of DNA and RNA," J. Chem. Theory Comput. **10**, 3177–3189 (2014).

[34]X. Xu, T. Yu, and S.-J. Chen, "Understanding the kinetic mechanism of RNA single base pair formation," Proc. Natl. Acad. Sci. U. S. A. **113**, 116–121 (2016).
[35]T. Xia, J. SantaLucia, Jr, M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs," Biochemistry **37**, 14719–14735 (1998).
[36]K.-Y. Wong and B. M. Pettitt, "The pathway of oligomeric DNA melting investigated by molecular dynamics simulations," Biophys. J. **95**, 5618–5626 (2008).
[37]A. Perez and M. Orozco, "Real-time atomistic description of DNA unfolding," Angew. Chem., Int. Ed. **49**, 4805–4808 (2010).
[38]M. F. Hagan, A. R. Dinner, D. Chandler, and A. K. Chakraborty, "Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA," Proc. Natl. Acad. Sci. U. S. A. **100**, 13922–13927 (2003).
[39]N.-V. Buchete and G. Hummer, "Coarse master equations for peptide folding dynamics," J. Phys. Chem. B **112**, 6057–6069 (2008).
[40]C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, "Markov state models based on milestoning," J. Chem. Phys. **134**, 204105 (2011).
[41]A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science **344**, 1492–1496 (2014).
[42]M. d'Errico, E. Facco, A. Laio, and A. Rodriguez, "Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering," e-print arXiv:1802.10549 (2018).
[43]D. Tan, S. Piana, R. M. Dirks, and D. E. Shaw, "RNA force field with accuracy comparable to state-of-the-art protein force fields," Proc. Natl. Acad. Sci. U. S. A. **115**, E1346–E1355 (2018).
[44]P. Banáš, D. Hollas, M. Zgarbová, P. Jurečka, M. Orozco, T. E. Cheatham III, J. Šponer, and M. Otyepka, "Performance of molecular mechanics force fields for RNA simulations: Stability of UUCG and GNRA hairpins," J. Chem. Theory Comput. **6**, 3836–3849 (2010).
[45]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," J. Am. Chem. Soc. **117**, 5179–5197 (1995).
[46]A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham III, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: Improving the description of α γ conformers," Biophys. J. **92**, 3817–3829 (2007).
[47]S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw, "Water dispersion interactions strongly influence simulated structural properties of disordered protein states," J. Phys. Chem. B **119**, 5113–5123 (2015).
[48]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," J. Chem. Phys. **79**, 926–935 (1983).
[49]A. D. MacKerell, Jr, D. Bashford, M. Bellott, R. L. Dunbrack, Jr, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha *et al.*, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," J. Phys. Chem. B **102**, 3586–3616 (1998).
[50]J. Aaqvist, "Ion-water interaction potentials derived from free energy perturbation simulations," J. Phys. Chem. **94**, 8021–8024 (1990).
[51]L. X. Dang, "Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: A molecular dynamics study," J. Am. Chem. Soc. **117**, 6954–6960 (1995).
[52]I. Beššeová, P. Banáš, P. Kührová, P. Kosinova, M. Otyepka, and J. Sponer, "Simulations of A-RNA duplexes. the effect of sequence, solute force field, water model, and salt concentration," J. Phys. Chem. B **116**, 9899–9916 (2012).
[53]G. Pinamonti, S. Bottaro, C. Micheletti, and G. Bussi, "Elastic network models for RNA: A comparative assessment with molecular dynamics and SHAPE experiments," Nucleic Acids Res. **43**, 7260–7269 (2015).
[54]B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).
[55]T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089–10092 (1993).
[56]G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).
[57]M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182–7190 (1981).
[58]D. E. Condon, S. D. Kennedy, B. C. Mort, R. Kierzek, I. Yildirim, and D. H. Turner, "Stacking in RNA: NMR of four tetramers benchmark molecular dynamics," J. Chem. Theory Comput. **11**, 2729–2742 (2015).
[59]S. Bottaro, F. Di Palma, and G. Bussi, "The role of nucleobase interactions in RNA structure and dynamics," Nucleic Acids Res. **42**, 13306 (2014).
[60]S. Bottaro, G. Bussi, G. Pinamonti, S. Reisser, W. Boomsma, and K. Lindorff-Larsen, "Barnaba: Software for analysis of nucleic acid structures and trajectories," RNA **25**, 219–231 (2019).
[61]R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," Biophys. J. **109**, 1528–1532 (2015).
[62]G. Pinamonti, J. Zhao, D. E. Condon, F. Paul, F. Noé, D. H. Turner, and G. Bussi, "Predicting the kinetics of RNA oligonucleotides using Markov state models," J. Chem. Theory Comput. **13**, 926–934 (2017).
[63]X. Huang, G. R. Bowman, S. Bacallado, and V. S. Pande, "Rapid equilibrium sampling initiated from nonequilibrium data," Proc. Natl. Acad. Sci. U. S. A. **106**, 19765–19769 (2009).
[64]D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of src kinase reveals intermediate states as targets for drug design," Nat. Commun. **5**, 3397 (2014).
[65]K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman, and V. S. Pande, "Cloud-based simulations on google exacycle reveal ligand modulation of GPCR activation pathways," Nat. Chem. **6**, 15 (2014).
[66]S. K. Sadiq, F. Noé, and G. De Fabritiis, "Kinetic characterization of the critical step in HIV-1 protease maturation," Proc. Natl. Acad. Sci. U. S. A. **109**, 20449–20454 (2012).
[67]O. Lemke and B. G. Keller, "Density-based cluster algorithms for the identification of core sets," J. Chem. Phys. **145**, 164104 (2016).
[68]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys. **139**, 015102 (2013).
[69]F. Noé and C. Clementi, "Kinetic distance and kinetic maps from molecular dynamics simulation," J. Chem. Theory Comput. **11**, 5002–5011 (2015).
[70]A. Rodriguez, M. d'Errico, E. Facco, and A. Laio, "Computing the free energy without collective variables," J. Chem. Theory Comput. **14**, 1206–1215 (2018).
[71]E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information," Sci. Rep. **7**, 12140 (2017).
[72]G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande, "Progress and challenges in the automated construction of Markov state models for full protein systems," J. Chem. Phys. **131**, 124101 (2009).
[73]B. Trendelkamp-Schroer, H. Wu, F. Paul, and F. Noé, "Estimation and uncertainty of reversible Markov models," J. Chem. Phys. **143**, 174101 (2015).
[74]M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Perez-Hernandez, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models," J. Chem. Theory Comput. **11**, 5525–5542 (2015).
[75]W. Kabsch, "A solution for the best rotation to relate two sets of vectors," Acta Crystallogr. Sect. A **32**, 922–923 (1976).
[76]S. Bottaro, G. Bussi, S. D. Kennedy, D. H. Turner, and K. Lindorff-Larsen, "Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations," Sci. Adv. **4**, eaar8521 (2018).
[77]R. B. Best, K. Lindorff-Larsen, M. A. DePristo, and M. Vendruscolo, "Relation between native ensembles and experimental structures of proteins," Proc. Natl. Acad. Sci. U. S. A. **103**, 10901–10906 (2006).
[78]S. Bottaro, A. Gil-Ley, and G. Bussi, "RNA folding pathways in stop motion," Nucleic Acids Res. **44**, 5883–5891 (2016).
[79]E. Weinan and E. Vanden-Eijnden, "Towards a theory of transition paths," J. Stat. Phys. **123**, 503 (2006).

[80]F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci. U. S. A. **106**, 19011–19016 (2009).
[81]R. A. Cunha and G. Bussi, "Unravelling Mg^{2+}-RNA binding with atomistic molecular dynamics," RNA **23**, 628–638 (2017).
[82]A. Spasic, J. Serafini, and D. H. Mathews, "The amber ff99 force field predicts relative free energy changes for rna helix formation," J. Chem. Theory Comput. **8**, 2497–2505 (2012).
[83]S. Sakuraba, K. Asai, and T. Kameda, "Predicting RNA duplex dimerization free-energy changes upon mutations using molecular dynamics simulations," J. Phys. Chem. Lett. **6**, 4348–4351 (2015).
[84]P. Kuhrova, V. Mlynsky, M. Zgarbova, M. Krepl, G. Bussi, R. B. Best, M. Otyepka, J. Sponer, and P. Banas, "Improving the performance of the amber RNA force field by tuning the hydrogen-bonding interactions," J. Chem. Theory Comput. (published online, 2019).
[85]V. Bloomfield, D. Crothers, and I. Tinoco, *Nucleic Acids: Structures, Properties, and Functions* (University Science Books, 2000).
[86]S. Mohan, C. Hsiao, H. VanDeusen, R. Gallagher, E. Krohn, B. Kalahar, R. M. Wartell, and L. D. Williams, "Mechanism of RNA double helix-propagation at atomic resolution," J. Phys. Chem. B **113**, 2614–2623 (2009).
[87]F. Vitalini, A. S. Mey, F. Noé, and B. G. Keller, "Dynamic properties of force fields," J. Chem. Phys. **142**, 084101 (2015).
[88]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," Biophys. J. **100**, L47–L49 (2011).