Coarse-grained modelling applied to RNA morphologies and flexibility

Author: Marc Burillo Garcia* and Advisor: Ignacio Pagonabarraga Mora
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Modesto Orozco López
Institute for Research in Biomedicine, Baldiri Reixac 10, 08028 Barcelona, Spain.

Abstract: Traditional methods in Molecular Dynamics are encountering their computational boundaries to simulate long-time periods and large macromolecules, such as RNA. We develop an alternative approach by using: i) an ultra-simplified coarse-grained (CG) representation of the molecule and ii) a flexible non-linear potential intended to represent the accessible conformational space of RNA. This model can reproduce the flexibility observed in force-field simulations and different morphologies can be modelled. These results underscore the necessity of non-linear dynamics and statistical physics in CG to capture the dynamic behaviour of Hydrogen Bonds in RNA.

I. INTRODUCTION

Ribonucleic acids (RNAs) are biomolecules essential for life, exhibiting different roles: from intermediates in the expression of genetic information to structural macromolecules or even catalysts[1, 2]. Contrary to deoxyribonucleic acid (DNA), which appears as a complementary double helix, RNA is found in the cell as a single strand, which folds adopting a myriad of conformations which can interchange in a longer time scale than the μs[2]. Representing such structural diversity from physical methods has been a central objective of theoretical biophysics for decades. Most of the developed strategies are based on classical Hamiltonians (force-fields; FF) which are used in the context of atomistic molecular dynamics (MD)[2]. Despite their impressive success, atomistic MDs have many intrinsic shortcomings derived from the inaccuracy of FF and the limited length of the trajectory that can be reached (typically below the μs)[2]. Coarse-grained (CG) MD is a methodology based on simplifying the system into a small set of *beads*, which are generally point-like particles, in geometrical points of interest. CG methods are based on Langevin Dynamics and aim to capture the essential degrees of freedom of the system. These methods have been able to simulate longer time intervals and bigger molecules than FF[2].
In this TFG, a 2-bead CG model is explored. This model is designed with a particular emphasis on devising novel potentials to capture the experimental behaviour of RNA. This TFG has generalized a CG model describing single-strand morphologies to simulate more complex structures. Firstly, we describe the fundamental physics underlying the original model, with specific attention to the statistical potential for torsional angles (*torsionals*) that distinguishes this model. Afterwards, a novel algorithm to form locally dynamic Hydrogen Bonds (HB) between nucleotides is described regardless of the potential chosen to model HB. Subsequently, the potentials used to incorporate HB and Stacking interactions are introduced, which are sophisticated variations of the Morse Potential that allow exploring diverse conformations. Finally, the behaviour of the model is portrayed in the analysis of pre-miR-31 which presents a complex morphology.

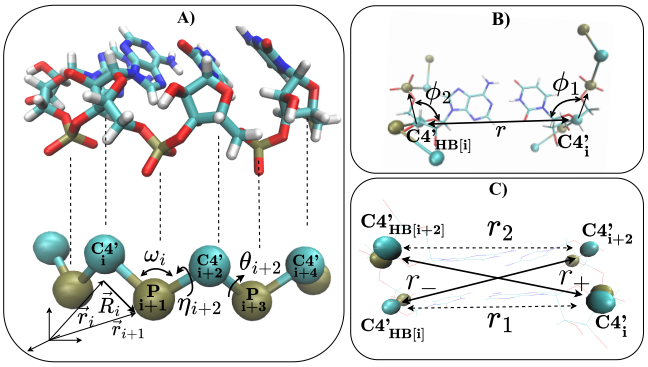

FIG. 1: A)At the top, four nucleotides are shown with all the atoms and below the CG representation of them. In addition, the main variables of the model are illustrated. B)Hydrogen Bond interaction is represented. C)Observables involved in the 'Cross-Stacking' interaction are depicted.

II. METHODS

This CG model simplifies the RNA molecule to 2 beads per residue located on $C4'$ and P atoms of the backbone (see Fig.1), dividing the average mass of a residue equally into both beads. These two positions allow us to identify easily the sugar bases and nitrogen groups. In addition, this mass distribution facilitates the study of RNA foldings. Langevin dynamics is the formalism adopted to incorporate Brownian motion and friction caused by the solvent. Computationally, the Velocity Verlet algorithm is used for integrating the motion equations, which is a well-established method in MD [3]. The prior interaction

*Email address: mburilga7@alumnes.ub.edu

potentials of this model before this TFG were:

$$V = V_{Bonded} + V_{Non-Bonded} \quad (1)$$
$$V_{Bonded} = V_{Bonds} + V_{Angles} + V_{S.T} \quad (2)$$
$$V_{Non-Bonded} = V_{DH} + V_{LJ} \quad (3)$$

Harmonic potentials have been proved adequate for the interaction between 2 and 3 successive atoms in different models [2]. For two atoms: $V_{Bonds} = \frac{1}{2}k_B(|\vec{R}_i| - R_0)^2$, where $\vec{R}_i = \vec{r}_{i+1} - \vec{r}_i$. For three contigous atoms: $V_{Angles} = \frac{1}{2}k_A(\omega_i - \omega_0)^2$ refers to the angle ω_i formed by $\vec{R}_i$ and $\vec{R}_{i+1}$. Both R_0 and ω_0 are the equilibrium distance/angle for each potential. All variables are illustrated in Fig. 1.

This CG model originally intended to reproduce the phase space of two consecutive torsionals (η, θ) characterized through the analysis of 10000 experimental structures [4], the most recent and largest inspection of these variables. The torsionals (η, θ) are defined with the torsional associate to $(C4'_i, P_{i+1}, C4'_{i+2}, P_{i+3})$ and $(P_{i+1}, C4'_{i+2}, P_{i+3}, C4'_{i+4})$, respectively (see Fig.1). A single torsional ϕ is calculated given a set of 4 beads $(i, i+1, i+2, i+3)$ as $\cos\phi = \hat{n}_{i,i+1,i+2} \cdot \hat{n}_{i+1,i+2,i+3}$ and counter-clockwise around the bond $(i+1, i+2)$, where each normal vector corresponds to the geometrical plane formed by 3 beads. This analysis [4] revealed that the two torsionals angles should be used to characterize properly the structure and they found clusters of experimental structures around specific regions in (η, θ) phase space.

The solution to the model for torsionals reproducing high-density conformations of experimental structures in [4] is a Statistical Torsionals potential $V_{S.T}$. This is constructed with the prior inference of a statistical model from a set of experimental observations. Among the different distributions for torsionals statistical potential in literature, the Bivariate von Mises Sine (BvMS) model has been priorly used in the modelization of proteins [5] and it is optimal for numerical simulations (see additional details in Appendix IV B).

Thus, the statistical model was $([0, 2\pi] \times [0, 2\pi]$, $P(\eta, \theta) = \sum_{n=0}^{Nc} \pi_n \mathcal{P}_n)$, where Nc is the number of clusters distinguished [4], π_n the weight of a particular BvMS modelling a cluster in the phase space, $\mathcal{P}_n$ is a BvMS that models the n cluster, which infers a specific set of parameters for each n. These sets and π_n are inferred through the Expectation-Maximization Machine-Learning algorithm[5]. This determines $P(\eta, \theta)$. Hence, the torsionals potential can be deduced using statistical physics. In the formalism of the canonical ensemble (simulations are at constant $T = 298$K), the probability of a configuration is directly related to its potential energy:

$$P(\eta, \theta) = \frac{1}{Z} e^{-V_{S.T}(\eta,\theta)/k_B T} \quad (4)$$

where Z is the partition function. In particular, $V_{S.T}(\eta, \theta) = -k_B T \log(\sum \pi_i \mathcal{P}_n(\eta, \theta))$.

Concerning $V_{Non-Bonded}$, the first term V_{DH} refers to Debye-Hückel which takes into consideration the interaction with ions present in the solvent and the shielding effect. The second term, V_{LJ} is Lennard-Jones potential that induces a divergence that avoids beads not related by V_{Bonded} colliding. These two potentials do not interact between beads involved in the same (η_i, θ_i) to minimize interactions with $V_{S.T}$ to reproduce their phase space better.

This set of potentials (Eq.1-3) model properly single strand structures exploring the torsionals phase space analysed by [4]. In the following section, a Pairing Algorithm for HB is described which incorporates dynamic HBs into the model. The next sections provide details on how the interaction potential has been generalized to incorporate HB and Stacking forces, which will expand Eq.3 and enable the simulation of more complex morphologies.

A. HB Pairing

HBs have a crucial role in determining secondary structure (2D representation), which influences the 3D morphology, consequently, the Non-Bonded contribution of this CG Model should account for the HB interactions. These bonds are not rigidly formed between static pairs in the time-life of RNA, instead, they are formed/broken depending on the local properties of RNA (mainly the dipole-dipole intermolecular forces of HBs). Due to the aim of reproducing this observed flexibility and exploring different morphologies, a dynamical solution to HB is desired.

Initially, an extensive review of methods of HB pair formation suggested Needleman-Wunsch [6] and Smith-Waterman algorithms [7], which are algorithms of sequence alignment between different nucleotides/proteins. However, when creating or breaking bonds, local observables beyond the Nitrogen group are fundamental; therefore, we cannot use these optimization algorithms.

Another possible approach could be to incorporate interactions between all Nitrogen bases, each interaction depending on local parameters. However, this solution is expensive computationally and it is not optimal for CG modelling.

The solution we propose is a Dynamic Local Pairing Algorithm (DLPA) that allows the addition of the minimum 2-body interactions and describes the pairing through local parameters. In this section, the algorithm will be explained for any given potential V_{HB} used to model HB, on the condition that it has only one minimum (only one equilibrium position). The next section will describe the HB potential chosen for this model.

The core of DLPA is again the formalism of the canonical ensemble: $P(HB_{ij}) = C \exp(-V_{HB}(HB_{ij})/k_B T)$. This probability and the local variables of $V_{HB}(HB_{ij})$, given an HB between the beads i and j, determine the pairing and, therefore the changes in secondary structure. However, HBs are canonically established in 2 pos-

sible pairs: AU and CG (known as Watson-Crick pairs), and a third non-canonical pair UG (known as Wobble pair) has also been considered because it is frequently present[1]. These constraints are accounted for in the algorithm: $P(HB_{ij}) = 0$ if $(B_i, B_j) \notin \{AU, CG, UG\}$, where B_i is the nitrogen base of the bead i. Furthermore, the only possible pairs are between C4'-C4'. Provided that P is further from the Nitrogen base it is more realistic to choose C4' as the particle forming HB.
DPLA uses two main routines: firstly, it determines whether current HBs are considered strong (S), weak(W) or broken(B); secondly, new pairs are created when an HB is not S. Supposing $\mathcal{O} = \{x_1, ..., x_r\}$ a set of local observables that determine $V_{HB}(HB_{ij})$ for a feasible HB_{ij}, then S,W and B are defined as:

$$S = \{(x_1, .., x_r) : \forall k \quad |x_k - x_{1k}^{B_iB_j}| < \delta_1^{B_iB_j} x_k\} \tag{5}$$

$$W = \{(x_1, .., x_r) : \forall k \quad |x_k - x_{1k}^{B_iB_j}| < \delta_2^{B_iB_j} x_k\} \setminus S \tag{6}$$

$$B = \{(x_1, .., x_r) : \exists k \quad \delta_2^{B_iB_j} x_k < |x_k - x_{1k}^{B_iB_j}|\} \tag{7}$$

The parameters $x_{k0}^{B_iB_j}$, for $k = 1, ..., r$ are the values of the observables in the minimum of V_{HB}. In addition, $\delta_1 x_k, \delta_2 x_k$ are adjusted considering the fluctuation of HBs in FF simulations. Moreover, all these parameters depend on the bases that are paired: AU, CG and UG; since their equilibrium value of the observables and its fluctuations vary with the stability of the bond [1].
The second procedure firstly classifies all nucleotides unpaired or with a broken/weak HB as to be Re-bonded (R). Then, we implemented two steps. 1) Two subsets $A, B \subset R$ are selected to find the most probable bond according to $P(HB_{ij})$, in which $i \in A$ and $j \in B$. 2) The nucleotides of the selected pair are removed from R. These steps are iterated until no more bonds can be formed. At the end of DLPA, all nucleotides have been paired or they are in a region Unable to Re-bond (UR), which produces a bulge, mismatch or a loop. The second procedure finishes in a finite number of iterations and the algorithm produces always a pairing that includes the most probable HBs. More details in Appendix IV C.

B. Hydrogen Bonds Potential

This section will motivate the election of a complex HB potential that enables observing flexibility and different morphologies.
The study of different FF simulations determined that given the following set of beads $(C4'_i, P_{i+1})$ (in the same residue) and their counterparts in the HB: $(C4'_{HB[i]}, P_{HB[i]+1})$, 3 observables were stable. Obviously, the most stable was the relative distance between $C4'_i$ and $C4'_{HB[i]}$, $r = |\vec{r}(C4'_i) - \vec{r}(C4'_{HB[i]})|$. The two other were the pairs of angles (ϕ_1, ϕ_2) definded as $\phi_1 = \widehat{P_{i+1}C4'_iC4'_{HB[i]}}$ and $\phi_2 = \widehat{C4'_iC4'_{HB[i]}P_{HB[i]+1}}$, which both remained close to 110° (see Fig.1.B). Therefore, the local observables that determine a feasible HB are $\mathcal{O} = \{r, \phi_1, \phi_2\}$ in this model.
Provided that V_{HB} would depend only on r in a first approach, we expect $V_{HB}^0 = V(r)$. The interaction expected could resemble the one observed between atoms: when they are at a large distance there is no interaction, there is a stable region at an intermediate distance and it diverges to avoid collapse. The initial potential chosen was the standard Morse potential $V_M(r) = V_0\left(1 - e^{-a(r-r_0)}\right)^2 - V_0$ which has a minimum at $r = r_0$ and it diverges for $r < r_0$. However, due to the already existing divergence of V_{LJ} for small r, we had to reduce the integration step to inefficient values for CG simulation. Therefore, we modified V_M to avoid additional divergence and to become it symmetric:

$$V_{HB}^0 = V_{SM}(r) = \begin{cases} V_0\left(1-e^{-a(r-r_0)}\right)^2 - V_0 & \text{if } 0 \le r \le r_0 \\ V_0\left(1-e^{+a(r-r_0)}\right)^2 - V_0 & \text{if } r_0 \le r \le r_B \end{cases} \tag{8}$$

where V_0 is the depth of the well and a regulates its amplitude. r_0 is the equilibrium distance and r_B is the distance where a bond is considered broken and no interactions exist for larger r. This Symmetric Morse (SM) potential does not incorporate any additional divergence as $V_{SM} \to 0$ as $r \to 0$. In addition, for $r > r_B$ there is no interaction and the structure can lose its helicity. This has been observed testing the potential in a small Hairpin (PDB: 1KR8), where a transition between open strand and Hairpin was observed. Nevertheless, this first approach allowed unfeasible bonds in DLPA since RNA is very flexible and during a folding unrealistic bonds occurred. This is fixed by incorporating new terms that take into consideration the rest of observables ϕ_1 and ϕ_2. The new terms should foster those interactions which present ϕ_1 and ϕ_2 close to their mean value observed $\phi_0 = \langle\phi(t)\rangle_{FF}$. This second approach is:

$$V_{HB}^1 = V_{ASM}(r, \phi_1, \phi_2) = V_{SM}(r) e^{-\frac{(\phi_1-\phi_0)^2}{\sigma^2}} e^{-\frac{(\phi_2-\phi_0)^2}{\sigma^2}} \tag{9}$$

incorporating two Gaussian related to the Angular observables. Therefore, these two Gaussians modulate the interaction and the systems tend to ϕ_0. This last potential did produce feasible bonds during the DLPA and no further development was necessary. Moreover, the sets of constants in $V_{HB} = V_{ASM}(r, \phi_1, \phi_2)$ were parametrized by the following procedure. First, $\phi_0 = \langle\phi(t)\rangle_{FF}$ and $\sigma^2 = Var(\phi(t))_{FF}$, then a and V_0 through Replica Exchange Molecular Dynamics (REMD)[8]. This was performed because they could not be determined by statistical inference from previous simulations of the molecule with FF since none allowed transitions between open and closed hairpins. More details on REMD are in the Appendix IV D.

C. Stacking Potential

Beyond HB interaction, Stacking is a phenomenon typically observed in double-helix structures. Stacking

is produced between the bases of two consecutive nucleotides, as a result of hydrophobic, electrostatic and dispersion effects; creating an attracting force[2]. As a result, the structure is more compressed since nucleotides are stacked closely. Helical structures (PDB: 1RNA, 8FCS) were simulated with this model before the stacking development and they had a larger end-to-end distance.
In the first approach, the distance between the bases should determine the interaction between them. However, our CG model does not have any bead associated with the base, only $C4'$ that accounts for the sugar ribose and the base. A potential between two consecutive $C4'$ would alter the dynamics of other local observables. Thus, the proposed potential for these characteristics was two 'Cross-Stacking' potentials. Provided this interaction is mainly observed in helical structures, the stacked nucleotides have an HB to two other bases. Given $(C4'_i, C4'_{i+1})$, there is a related set is $(C4'_{HB[i]}, C4'_{HB[i+1]})$ which is located at the opposite side of the helix. The first approach is: $V^0_{CS+} = V_{SM}(r_+)$, where $r_+ = |\vec{r}(C4'_i) - \vec{r}(C4'_{HB[i+1]})|$; and $V^0_{CS-} = V_{SM}(r_-)$, where $r_- = |\vec{r}(C4'_{i+1}) - \vec{r}(C4'_{HB[i]})|$. The symmetric Morse attracts nucleotides thus compressing the helix globally. A similar potential has been used in other CG Models such as OXRNA [2].
Nevertheless, this interaction should only appear when the HB is formed and should decrease as this breaks apart. Therefore, the local observable that should be used for this potential is $\mathcal{O}_\pm = \{r_\pm, r_1, r_2\}$, where $r_1 = |\vec{r}(C4'_i) - \vec{r}(C4'_{HB[i]})|$ and $r_2 = |\vec{r}(C4'_{i+1}) - \vec{r}(C4'_{HB[i+1]})|$. These magnitudes are incorporated as:

$$V_{CS\pm} = V_{SM}(r_\pm) e^{-|r_1 - r_1^0|^s/\sigma_1^2} e^{-|r_2 - r_2^0|^s/\sigma_2^2} \qquad (10)$$

where parameter $s > 1$, ideally close to 2, and $\sigma_i > 0$ $(i = 1, 2)$ modulates the range of interaction between HB that allows the interaction of V_{CS}. Because of this physical meaning, the exponential may not be Gaussian $(s \neq 2)$ and be of a larger order to provide a more "trapezoidal" shape and allow V_{SM} to act more intensely where there is significant interaction. Once the parameters of exponentials have been characterized: $r_i^0 = \langle r_i(t) \rangle_{FF}$ and $\sigma_i^2 = Var(r_i(t))_{FF}$ $(i = 1, 2)$; the parameter V_0 of V_{SM} can be determined through REMD[8].
These potentials worked fine, but simulations of double-strand RNA, such as A-RNA (PDB: 1RNG), V_0 associated with Morse potential in $V_{CS\pm}$ did not compress enough the helix. Therefore, the Stacking interaction was extended beyond consecutive bases: $V_{CS++}(r_{++}, r_1, r_3)$ with $r_{++} = |\vec{r}(C4'_i) - \vec{r}(C4'_{HB[i+2]})|$, and $r_3 = |\vec{r}(C4'_{i+2}) - \vec{r}(C4'_{HB[i+2]})|$; and its counterpart: $V_{CS--}(r_{--}, r_1, r_3)$ with $r_{--} = |\vec{r}(C4'_{i+2}) - \vec{r}(C4'_{HB[i]})|$. These additional interactions proportionate the expected compression between residues and the simulation of helical structures resembled its FF counterpart. Additional V_{CS} with further nucleotides were studied but the final effect was not as impactful as the second-order interaction and the computational time increased.

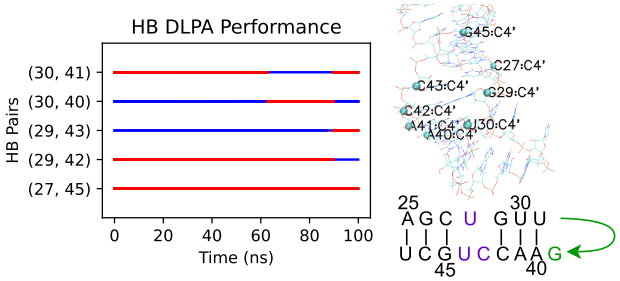

FIG. 2: DLPA performance during 100 ns in bonding residues from 27 to 30 and 40 to 45, which are close to a loop. Each possible pairing between residues is only displayed if there is any bond during this time interval. In case there is, the baseline is blue, and when the interaction occurs, this is highlighted in red. At the right of the chart, a part of the tertiary structure of pre-mir-31 (PDB: 8FCS) with bonded nucleotides are depicted. Below there is the predicted secondary structure for these residues. Purple is used for bulge residues and green for the loop represented by the arrow and G nucleotide.

D. Results

The CG model presented in this TFG simulates using the Eq. 1-3, with the latter extended to include the described interactions:

$$V_{Non-Bonded} = V_{DH} + V_{LJ} + V_{HB} + V_{CS\pm} + V_{CS\pm\pm} \qquad (11)$$

and the integration of the motion equations incorporates the DLPA.
This research has culminated with the test of pre-mir-31 (PDB: 8FCS). This structure is 71 residues long with a loop of 6 residues, a bulge close to the loop and two mismatches along the helical stem. Consequently, this complex molecule requires the proper interaction between all the potential present in the generalized Eq. 1,2,11 and DLPA.
To illustrate our CG model, a 1μs simulation of 8FCS is analysed with specific attention to the new interactions. The model requires as input the initial coordinates, the sequence of base residues and the time step of 0.02 ps.

The first analysis for describing HB interaction is Fig.2 which displays the dynamical behaviour of this interaction between different bonds. The residue 27 is bonded to 45 during the 100ns, which means it is a very stable HB. However, residues 28 and 44 are not able to pair with any other because they are in the bulge. Residue 29 changes its bond from 42 to 43 at 85ns. Additionally, residue 30 is bonded to 41, except for 20ns when it is paired to 40. These small changes in secondary structure allow us to study variations in the morphology during the simulation.

The second analysis in Fig.3 addresses the dynamics of the observables $\mathcal{O}_\pm = \{r_\pm, r_1, r_2\}$ and the variations in $V_{CS\pm}(r_\pm)$ and $V_{HB}(r_{1,2})$. It can be seen that during this 25 ns, the HB is dynamic. The HBs between beads 29-47 and 30-46 act only the first 2 ns and the last 5 ns when the r is close to 15.2Å. During these two inter-

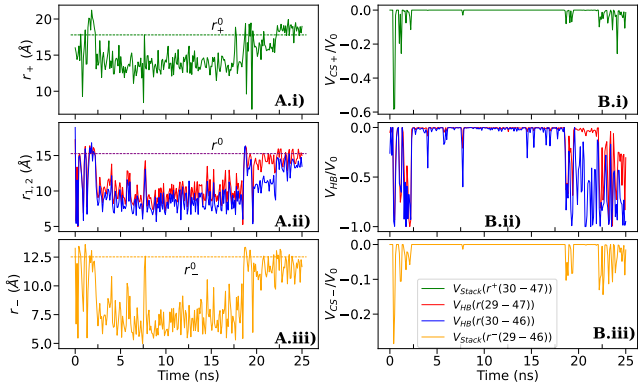

FIG. 3: Stacking Interactions during 25 ns. Panels A) describe the dynamics of the observables involved in $V_{CS\pm}$. A.i) displays r_+ between beads 30-47, with equilibrium distance in the dashed line. A.ii) shows r_1, r_2 between beads (29-47) and (30-46), which are HB interactions. A.iii) describes r_- between beads(29-46). Panels B) illustrate the ratio V/V_0 for Stacking interactions in panels B.i) and B.iii) and HB interaction in B.ii).

vals, the Stacking interaction reveals: a) r_+ and r_- are close to their equilibrium value in A.i and A.iii and b) the Stacking energy is negative only then (panels B.i, B.iii). In all B Panels, V/V_0 is expected to be -1 in the equilibrium position. This is observed in panel B.ii, but the ratio of Stacking interactions is not close to -1, despite $r_\pm$ crossing both their equilibrium position. The reason is that $V_{CS\pm}$ is modulated by the HB which might not be at the equilibrium distance at the same time. This illustrates how HB influences V_{CS} and the non-linearity of this model.

All other observables discussed in the description of the model have been analysed and their expected behaviour occurred. Particularly, the new potentials do not interfere with $V_{S.T}$ which is the most important characteristic of the model. The global behaviour of the molecule is coherent.

III. CONCLUSIONS

The CG Model presented in this TFG has distinct features that contribute to the study of RNA morphologies and, in particular, its torsionals phase space.
On one hand, specific potentials such as Lennard-Jones, Debye-Hückel, and well-established bonded interactions are crucial for capturing chemical interactions. However, to generalize the single-strand model it is required to incorporate HB and Stacking between bases. This TFG explores different potentials, including sophisticated Morse potentials. The physical properties of Morse potential enable the opening and closing of single strands into a hairpin shape while Stacking interactions compress long helical stems to maintain their conformational shape.
On the other hand, RNA's inherent complexity is challenging to represent accurately in a CG model. This complexity is addressed by introducing the two non-linear interactions: the Stacking potential, which is a modulated Morse potential influenced by HBs; and the DLPA for simulating HBs, which are correlated to angular observables. DLPA captures the smooth transition between similar states and naturally establishes the relationship between bond energy and its probability. Moreover, the statistical approach of DLPA and torsionals potential differs from traditional FF, which underscores the importance of these methods to sample complex phase spaces.
In conclusion, this CG Model facilitates simulations over extended periods and explores the targeted torsionals phase space. The newly introduced potentials successfully achieve the desired flexibility for complex morphologies and they allow to study of the stability of bond transitions between morphologies, which are vital in biological processes.

Acknowledgments

I want to thank the support of Alba, my parents, my colleague David Farré-Gil and my two Advisors, all of them essential to this project.

[1] Roy A, Panigrahi S., Bhattacharyya M., Bhattacharyya D. "Structure, Stability, and Dynamics of Canonical and Noncanonical Base Pairs: Quantum Chemical St.udies". J. Phys. Chem. B **112**: 3786-3796 (2008)
[2] Šulc P., Romano F., Ouldridge T.E., Doye J.P., Louis A.A. "A nucleotide-level coarse-grained model of RNA". J Chem Phys. **140**(23): 235102. (2014)
[3] Ben Leihmkuler and Charles Matthews, *Molecular Dynamics With Deterministic and Stochastic Numerical Methods*, (Springer, Cham, Switzerland, 2015)
[4] Grille L., Gallego D., Darré L., da Rosa G., Battistini F., Orozco M., Dans PD."The pseudotorsional space of RNA". RNA. **29**(12):1896-1909 (2023)
[5] Mardia KV, Taylor CC, Subramaniam GK, "Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data". Biometric. **63**(2): 505-12 (2007)
[6] Needleman S., Wunsch C. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. **48**(3): 443-453 (1970)
[7] Smith T., Waterman M. "Identification of Common Molecular Subsequences". Journal of Molecular Biology. **147** (1): 195–197 (1981)
[8] Woods C., Essex J., King M."The Development of Replica-Exchange-Based Free-Energy Methods" J. of Physical Chemistry B **107** (49): 13703-13710 (2003)

IV. APPENDIX

A. Numerical Stability

The code describing all potentials related to angular variables ($V_{Angles}, V_{S.T}$ and V_{HB}) required modifications since numerical instability was encountered when calculating forces. Forces are calculated using Langevin's Equation:

$$m\ddot{q_i} = -\frac{\partial V}{\partial q_i} - m\gamma\dot{q_i} + \sqrt{2\gamma k_B T}R_i(t) \tag{12}$$

Our model was already set in cartesian coordinates: $\{q_i\}_{i=1,...,3N} = \{x_1, y_1, z_1, ..., x_N, y_N, z_N\}$, where N is the number of beads. This avoids time-consuming changes to internal coordinates. Therefore, general α angles or torsionals are defined using dot product between two vectors $\hat{v_a}, \hat{v_b}$ (see ω_i in Fig. 1): $\alpha_{ab} = \arccos(\hat{v_a} \cdot \hat{v_b})$. Provided $\hat{v_a} \cdot \hat{v_b} = F(q_i)$ (other q_j may be involved in F), the acceleration for q_i related to α_{ab} has a factor:

$$-\frac{\partial V(\alpha_{ab})}{\partial q_i} = -\frac{\partial V(\alpha_{ab})}{\partial \alpha_{ab}} \frac{\partial \alpha_{ab}}{\partial(\hat{v_a} \cdot \hat{v_b})} \frac{\partial(\hat{v_a} \cdot \hat{v_b})}{\partial q_i}$$
$$= \frac{\partial V(\alpha_{ab})}{\partial \alpha_{ab}} \frac{1}{\sqrt{1-(F(q_i))^2}} \frac{\partial F(q_i)}{\partial q_i} \xrightarrow[F(q_i)\to\pm1]{} -\infty$$

This divergence as $F(q_i)$ tends to ±1 used to occasion exaggerated accelerations, and the only cause is the mathematics needed to work in the chosen coordinates rather than an observed physics phenomenon. This acceleration can be observed visually, but also considering the temperature fluctuations in time, where the temperature of the system is approximated with Kinetic energy $K = (3/2)Nk_BT$. This issue was solved by incorporating cubic polynomials $P_1(F(q_i)), P_2(F(q_i))$ (as a cubic spline) to modify $\frac{\partial \alpha_{ab}}{\partial F(q_i)}$ that satisfied continuity for all values of $F(q_i) \in [-1, 1]$ and $\frac{\partial \alpha_{ab}}{\partial F(q_i)}(F(q_i) = \pm 1) = 0$.

$$\frac{\partial \alpha_{ab}}{\partial F(q_i)} = \begin{cases} P_1(F(q_i)), & \text{if } -1 \le F(q_i) \le -F_0 \\ -\frac{1}{\sqrt{1-(F(q_i))^2}}, & \text{if } -F_0 \le F(q_i) \le F_0 \\ P_2(F(q_i)), & \text{if } F_0 \le F(q_i) \le 1 \end{cases} \tag{13}$$

The parameter F_0 is to regulate how much the original function is to be deflected, ideally close to 1. Consequently, in case the angle opened too much because of the interaction of different forces, it would simply cross the divergent region. This was introduced into the code incorporating Hermite Interpolation.

B. Bivariate von Mises Distribution

Bivariate von Mises Sine Model probability distribution used in the derivation of $V_{S.T}$ is:

$$\mathcal{P}(\eta, \theta) = ce^{\kappa_1 \cos(\eta-\mu)+\kappa_2 \cos(\theta-\nu)+\kappa_3 \sin(\eta-\mu)\sin(\theta-\nu)} \tag{14}$$

where c is the normalization constant, (μ, ν) is the peak of the pdf, $\kappa > 1, \kappa_2 > 0$ represents the inverse of variance compared to a Gaussian distribution and $\kappa_3 \in \mathbb{R}$ introduces correlation between the two torsional angles. This distribution is in the sample space of a $[0, 2\pi] \times [0, 2\pi]$ and the label of Sine Model is because the correlation term is the product of two sine terms, in comparison to the Cosine model which only uses a single cosine with the two angles as argument:

$$p(\eta, \theta) = ce^{\kappa_1 \cos(\eta-\mu)+\kappa_2 \cos(\theta-\nu)+\kappa_3 \cos(\eta-\mu+\theta-\nu)} \tag{15}$$

The Sine Model was chosen due to it having easier estimators for inducing the parameters that fit data[5].

C. HB DLPA

DLPA operates through 2 procedures. The first one has been accurately explained in Section II A. More details on the second procedure are provided in this appendix.

The second procedure essentially defines two sets to start the pairing: $NR = \{i : \exists j \quad HB_{ij} \in S\}$ which will not be Re-bonded and $R = \{i : \exists j \quad HB_{ij} \in W, B \vee i \text{ unpaired}\}$ which will attempt to Re-bond. Therefore, any residue $\{1, ..., N\}$ is in R or NR. Afterwards, the pairing starts. Firstly, the algorithm will find the maximal continous subset A in $\{1, ..., N\}$ such that $A \subseteq NR$ and $\min NR \in A$; and the maximal contingous subset B such that $B \subseteq R$ and $\max R \in B$. Continguous in this context means no nucleotide is missing between any two nucleotides. Secondly, once A and B are defined, the $(i_M, j_M) \in A \times B$ such that their observables $\mathcal{O}_{i,j} \in S$ and $P(HB_{i_M,j_M}) > P(HB_{i,j})\ \forall (i, j) \in A \times B \setminus (i_M, j_M)$ is formed. Automatically, $i_M, j_M \in NR$ and $i_M, j_M \notin R$. Subsequently, a new pairing begins: new A and B are defined with the resulting R and the pair with maximum probability is bonded; updating R and NR. These two steps are iterated until $\exists i \in R$ such that:(1) $i > \max A$ and $HB[i] > \max B$ or (2) $i < \min A$ and $HB[i] < \min B$. This means that A (in case (1), otherwise B) cannot be Re-bonded (UR) since any bond would cross previous bonds forming a knot. This implies that the subset A (otherwise B) is added to a new category UR. The three categories are exclusive, which means that R is reduced. Thus, a new subset in the updated R must be found to create the new pair. The algorithm iterates until $R = \emptyset$.

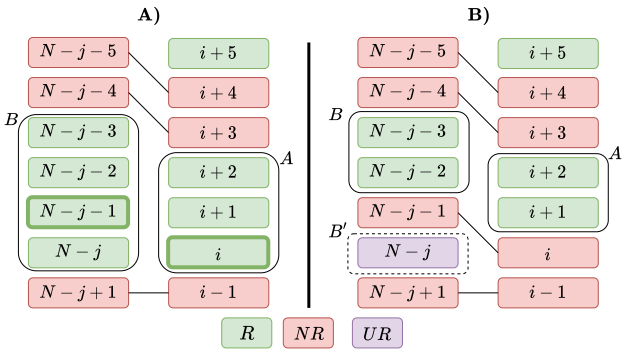

FIG. 4: A) An iteration of DLPA is represented in which $\forall m < i$, m nucleotide has a Strong bond or it cannot be paired. Similarly, $\forall n > N - j + 1$, where N is the number of residues (P are omitted). In this particular example, A and B have different sizes and the most probable bond is $(i, N - j - 1)$. The nucleotide $i+5$ is not included in A, since A must be continuous, as defined previously. B) The next iteration is illustrated after modifying the sets R, NR and UR. In this situation, a bond between $N-j$ and a nucleotide in A would 'cross' the bond $(i, N - j - 1)$ (case (2) explained in the text: bead $i < \min A = i + 1$ and $HB[i] = N - j - 1 < \min B' = N - j$).Therefore, this is not allowed, so $N - j \in UR$ and‘ B is the following set to use in the pairing instead of B'.

Moreover, since the changes in the secondary structure are experimentally observed in longer periods than the integration step, the second procedure can operate only every certain number of steps. This makes faster the simulations because the second procedure is slower as it involves iterating 2 steps, compared to the first procedure which classifies the state of the current HBs. In an integration step without the second phase, this updates whether a nucleotide is still paired or is not ($S \to W \to B$), reducing the number of forces to be calculated.

D. Replica Exchange Molecular Dynamics

REMD is a sampling method typically used to sample the phase space of a system with multiple regions of interest that may not be easy to access. This method consists of creating N replicas of the system and simulating each one at a given Temperature T_n. After a short interval of time, a Metropolis-Hastings Algorithm (described in Appendix IV E) is applied based on the potential energy of HBs (or Stacking, depending on which potential is parametrized) to exchange two structures at consecutive temperatures $T_i \leftrightarrow T_{i\pm1}$ and adapt current speeds of the beads to the new Thermal Reservoir $v' = \sqrt{(T_{i\pm1}/T_i)}v$.

After n iterations of this process, a structure at room temperature has been able to overcome barrier potentials inaccessible at that temperature. In this case, this allowed the structure to access states of opened and closed hairpins. The temperature at which structures are half of the REMD open and the other half helical is considered the Melting Temperature, T_M. Using the following relation for the free energy:

$$\Delta G^\circ = -k_B T_M \ln K \tag{16}$$

When the equilibrium constant $K = 1/2$, then ΔG° accounts for the potential depths V_0 [8]. The parameter a was fixed at the value used by OXRNA[2].
REMD does not only imply developing a specific code for computing this parametrization, but also efficiency is paramount because N simulations must be completed. Thus, it is highly advisable to run each replica in parallel cores in a high-performance computing environment. The library OMP for C-language was used for the parallelization.
Once the parameters were properly determined, the simulation of the tested hairpin (PDB: 1KR8) explored the desired conformations. By small alterations of the parameter V_0, the ratio simulated time of open structure to helical could be altered.

E. Metropolis-Hasting Algorithm

The Metropolis-Hasting Algorithm (MHA) is part of the REMD. The algorithm is a Montecarlo Method typically used to decide whether a new possible state, with energy E' is feasible given that the current state with energy E. Since REMD simulates in different temperature reservoirs, the current state is at temperature T and the next possible state will be in a reservoir at temperature T'. MHA defines a criteria to accept this exchange:

1. If $\Delta E < 0$, the new state is less energetic and therefore the exchange is accepted

2. If $\Delta E > 0$, then it is necessary to compare:

$$\exp\left(\frac{1}{k_B}\left(\frac{E'}{T'} - \frac{E}{T}\right)\right) > s(U(0,1)) \tag{17}$$

where $s(U(0,1))$ is a random number uniformly distributed between 0 and 1. If this last comparison is true, then the exchange is also accepted. Otherwise, it is rejected.