A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs

Simón Poblete[*], Sandro Bottaro and Giovanni Bussi[*]

Scuola Internazionale Superiore di Studi Avanzati, 265, Via Bonomea I-34136 Trieste, Italy

ABSTRACT

We introduce the SPlit-and-conQueR (SPQR) model, a coarse-grained (CG) representation of RNA designed for structure prediction and refinement. In our approach, the representation of a nucleotide consists of a point particle for the phosphate group and an anisotropic particle for the nucleoside. The interactions are, in principle, knowledge-based potentials inspired by the $\mathcal{E}$SCORE function, a base-centered scoring function. However, a special treatment is given to base-pairing interactions and certain geometrical conformations which are lost in a raw knowledge-based model. This results in a representation able to describe planar canonical and non-canonical base pairs and base–phosphate interactions and to distinguish sugar puckers and glycosidic torsion conformations. The model is applied to the folding of several structures, including duplexes with internal loops of non-canonical base pairs, tetraloops, junctions and a pseudoknot. For the majority of these systems, experimental structures are correctly predicted at the level of individual contacts. We also propose a method for efficiently reintroducing atomistic detail from the CG representation.

INTRODUCTION

During the last decades, RNA has been found to be much more than a mere messenger and translator of the genetic information in the cell. Its enzymatic and regulatory function has been observed in a variety of cellular processes, conferring it a major role in evolution and cellular metabolism ([1–5]). For the thorough understanding of these functions, an insight on the three-dimensional structure of RNA molecules is of crucial importance. Nevertheless, the reliable prediction of the full structure of a RNA motif based uniquely on its sequence is still a challenging aim.

RNA is dominantly composed by a mixture of the four most common nucleotides. Its alphabet is thus in principle significantly simpler than the one used by proteins. This, together with the simple rule governing Watson–Crick pairing, has suggested RNA folding and structure prediction to be a relatively easy task ([6]). However, RNA structural complexity is considerable due to the large number of backbone conformations ([7,8]) and the rich variety of interactions ([9–11]) which play a substantial contribution in the stability of many biologically relevant structures, and blind structure prediction is difficult to perform ([12–14]).

In this respect, many computational approaches have been proposed to accomplish this goal during the last decades. All-atom molecular dynamics (MD) simulations ([15]) would in principle allow to fold small motifs ([16]). Nevertheless, they are usually limited by the large amount of computational resources required and the reliability of the force fields, which still need certain refinement to be trusted quantitatively ([17,18]). In another vein, bioinformatic assembly tools ([19–28]) can increase the sampling efficiency, but they also possess limitations coming from their energy functions, do not necessarily provide information about the folding pathway and might fail when predicting motifs that are not already characterized experimentally and thus were not used in their training phase. Coarse-grained (CG) models thus emerge as a suitable alternative, at the expenses of resolution and versatility ([29–44]). Relying on different representations and interactions, several models have been proposed for the study of thermodynamic and mechanical properties, and/or to efficiently sample the conformational space for structure prediction and refinement purposes. Depending on the properties to reproduce and the criteria of the authors, CG models focus on particular interactions and structural features. For example, many of them restrict their base–base interactions to stacking and canonical base pairs ([44]), or might model hairpin loops as loose, non-interacting segments of RNA ([37]). However, more recent approaches have put emphasis on tertiary contacts by representing nucleobases as anisotropic objects, able to form directional interactions and non-canonical base pairs (e.g. ([30–32])), although their inclusion is still challenging, and in some cases, limited ([12–14]).

[*]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 Simón Poblete. Tel: +39 040 3787 301; Fax: +39 040 3787 528; Email: spoblete@sissa.it
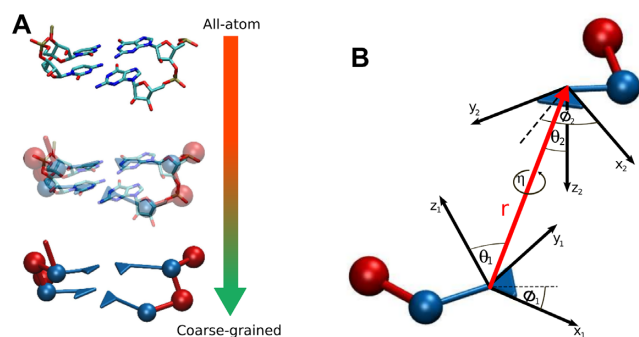
Figure 1. (**A**) Schematic representation of the mapping of four nucleotides with the nucleoside in blue and the phosphate in red. (**B**) The definition of the coordinate system from the oriented base, for two nucleotides. Dashed lines represent the projection of the vector that joins both bases in the *x-y* plane of each nucleoside. The base–base interactions, the base-phosphate interactions, and the interactions along the backbone are defined with respect to this reference frame.

In this paper, we introduce the SPQR (SPlit and conQueR) model, a nucleotide-level CG representation developed for the accurate prediction of the secondary and tertiary structure of small RNA motifs. The mapping of the nucleobases is based on the $\mathcal{E}$SCORE function (45,46), a scoring function which focuses exclusively on the relative arrangement of bases comparing it with the one observed in a structural database. Additionally, sugar and phosphate groups are represented as a virtual site and a point particle, respectively (see Figure 1). The interactions between these elements are designed to sample the geometrical probability distributions obtained from a large set of experimental structures from a base-centered perspective. However, these distributions have been carefully partitioned, identifying several of their contributions and reweighting their corresponding interactions. In this manner, the model focuses on an accurate geometrical description of base–base interactions (planar canonical and non-canonical base pairs) and base–phosphate interactions, while it rescues the conformation of the glycosidic bond angle and the sugar pucker, which are important elements in many relevant motifs.

We show that these interactions suffice for describing a number of structures involving a variety of canonical as well as non-canonical base pairs, like duplexes, tetraloops, three-way junctions and a pseudoknot. We also show that our base-centered approach is useful for the insertion of atomistic details in the predicted structure in a consistent framework.

The paper is organized as follows: the model and the interactions are exposed in the 'Materials and Methods' section, together with the simulation protocol used in the paper. In the following section, the results will be presented over several sets of structures. We will begin with a proof of concept of our method, to continue with a set of tetraloops, double strands (with and without internal loops), a pseudoknot, a subset of motifs already tested in the FARFAR protocol (47) and finally, a small set of junctions. The details of the structures can be found in the Supplementary Data Section 1. We later present the results of the backmapping procedure, which consists in the reintroduction of the atomistic resolution in the predicted structures, applied to

two tetraloops. The paper finishes with the conclusion and discussing how to improve the results presented here.

MATERIALS AND METHODS

Coarse-grained representation

The representation of a nucleotide consists of two elements: a point particle for the phosphate group and a rigid, anisotropic particle for the nucleoside, as shown in Figure 1. The base is represented as a triplet of particles forming a triangle which determines its centroid and orientation, as defined in the $\mathcal{E}$SCORE function (45). Meanwhile, a virtual site rigidly attached to the base represents the geometrical center of the sugar ring. Each nucleotide has a well defined sugar pucker (C2′-*endo* or C3′-*endo*) and a glycosidic bond state, which can be chosen between *anti* and *high-anti* conformations, and additionally *syn* for purines.

Interactions

The energy function between two nucleotides *i* and *j* is defined as:

$$U_{ij} = U_{ij}^{EV} + U_{ij}^{bp} + U_{ij}^{st} + U_{ij}^{bph} + U_{ij,\chi p}^{bb} \qquad (1)$$

which is a combination of excluded volume (EV), base-pairing (bp), stacking (st), base-phosphate (bph) and interactions along the backbone (bb). The latter depends both on the glycosidic bond angle and the sugar pucker, a conformation that will be referred as the χp state of a nucleotide from now on. This energy term is given by:

$$U_{ij,\chi p}^{bb} = U_{ij,\chi p}^{b} + U_{ij,p}^{a} + \frac{1}{2}\left(U_{i,\chi p}^{s} + U_{j,\chi p}^{s}\right), \qquad j = i+1 \qquad (2)$$

where, $U_{ij,p}^{a}$ is an energy function of the angle formed by the sugar–phosphate–sugar triplet of two consecutive nucleotides, and its functional form depends on the sugar pucker conformations of the nucleotides involved. $U_{ij,\chi p}^{b}$ is the interaction between nucleoside *i* and the phosphate group of nucleotide *i* + 1. In addition, the self term $U_{i,\chi p}^{s}$ is a sum of the energy between a nucleoside and its own phosphate group and a shift $\epsilon_i^{\chi p}$ which characterizes the glycosidic bond angle and sugar pucker conformations of the nucleotide. The prefactor 1/2 is not present when considering the terminal nucleotides.

Each energy term is designed to reproduce a probability distribution between two or more types of particles. This distribution is sampled from all their occurrences in a non-redundant list of structures (see Supplementary Data Section 1, (48)). The histograms of the planar base-pairing region obtained are spanned by classifying the points according to the specific kind of interaction using the FR3D package (49), and, therefore, we rely on its definition of base pairing and hydrogen bonding. Thus, it is possible to associate a particular distribution for stacking interactions and non-canonical base pairs, as well as base–phosphate and backbone interactions. The interaction between two bases turns out to be the most complex due to the anisotropy of both particles, but also the most important one to describe the RNA structure. The probability distribution of the spatial configuration of two nucleosides depends on six coor-
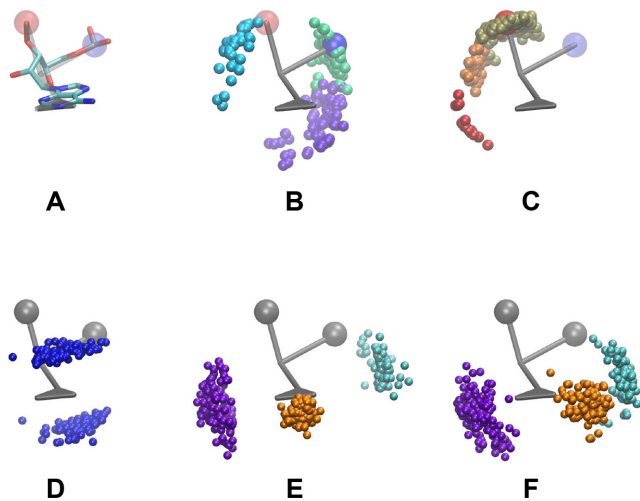
Figure 2. Depiction of selected clouds of points found in structural database: (**A**) CG and atomistic representations of Adenine as found in a typical duplex. The own phosphate group and the one of the following nucleotide are colored in blue and red in the CG representation, respectively. (**B**) Clouds of the nucleotide's own phosphate group; all clouds represent C3′-*endo* cloud under different χ conformations: *anti* (green), *high-anti* (purple) and *syn*(cyan). (**C**) Clouds of the neighboring nucleotide's phosphate group position, χ in *anti* conformation with sugar in C3′-*endo* (gold) and in C2′-*endo* (orange). Also, the *syn* conformation is shown in red. (**D**) Cloud of stacking points, (**E**) cloud of positions of paired bases through sugar (purple), Watson–Crick (orange) and Hoogsteen (cyan) faces. (**F**) represents the positions of the phosphate groups for base-phosphate interactions, with the same color nomenclature of (E).

dinates, which are shown in Figure 1B. The $\mathcal{E}$SCORE reference frame deals only with the angles θ_i, ϕ_i and the distance r. This allows to introduce the probability distribution function $P^{\mathcal{E}}(r, \theta, \phi)$ based on these degrees of freedom. It is therefore reasonable to approximate the full probability distribution as:

$$P(r, \theta_1, \phi_1, \theta_2, \phi_2, \eta) \approx$$
$$r^2 P_\eta(\eta) \times \frac{P_1^{\mathcal{E}}(r, \theta_1, \phi_1) P_2^{\mathcal{E}}(r, \theta_2, \phi_2)}{H_r(r)} \quad (3)$$

where, P_η and H_r are the probability distribution of the orientation η and the histogram of the distance r, respectively. The remaining factors include the probability distribution of the distance r and the corresponding Jacobian.

Once done this approximation, the interaction potential is obtained by Boltzmann inversion, as

$$U(r, \theta_1, \phi_1, \theta_2, \phi_2, \eta) =$$
$$-T_0 \log P(r, \theta_1, \phi_1, \theta_2, \phi_2, \eta) + \epsilon \quad (4)$$

where T_0 is a fictitious temperature. Here, we use the approximation of Equation (3) for the function P and introduce a shift ϵ, which corrects the arbitrary normalization of the probability distributions. This indetermination can be used to control the relative strength of the different interactions present in the system (see Figure 2). In a first instance, in the planar non-canonical base pairs, ϵ is chosen to make the minimum energy proportional to the number of hydrogen bonds present in the Leontis-Westhof tables (see

(10)), while the stacking interactions are adjusted with a fitting procedure over a series of structures (see 'Parametrization' section). This approach is intended to distinguish the multiple interactions that emerge not only between different species, but also between the Watson–Crick, Hoogsteen and Sugar/CH4 faces. Stacking interactions, on the other side, depend on the orientation of the planes defined by the bases involved. This makes that ϵ depends on the kind of interaction, species and faces involved in the pairing or stacking. For the base–phosphate interactions, the energy is given by Equation (4) but using $P^{\mathcal{E}}(r, \phi, \theta)$ as the probability distribution, due to the lack of structure of our phosphate group representation. In this case, the splitting of the probability distribution is also done according to the base faces involved. The position of the base with respect to the phosphate group in its local reference frame is also taken into account to form the base-phosphate interaction and avoid false positives (see Supplementary Data Section 2).

The backbone interactions are designed in a similar way. For the interactions between a nucleoside and its neighboring phosphates along the backbone, Equation (4) is used with its corresponding $P^{\mathcal{E}}$ probability distribution, which is classified according to their backbone conformations using the Suitename software (http://kinemage.biochem.duke.edu/software/suitename.php). The obtained probability clouds are found to depend strongly on the sugar pucker and the glycosidic bond of the involved base (see Figure 2). In this representation, they emerge as a natural partition of the conformational space, which keeps certain resemblance with previous pucker-dependent virtual bonds representations of the backbone (50). The sugar–phosphate–sugar angle shows also a clear dependency on these conformations (see Supplementary Data Section 2). In these cases, the absence of a special distinction of the conformations greatly favors the most populated conformation which corresponds to the C3′-*endo* pucker for the sugar pucker and *anti* conformation of the glycosidic bond torsion χ.

The EV interaction is also taken from the $\mathcal{E}$SCORE representation, which assigns to sugar, bases and phosphates an ellipsoidal geometry with specific parameters.

Simulation protocol

Random moves are proposed and accepted with a standard Monte Carlo Metropolis procedure (51). In addition, each nucleotide can change its χp state by displacing and rotating the nucleobase and remapping the sugar, as described in the Supplementary Data Section 3.

Our method, in its most general form, is intended to be applied in two steps: first, a search of the global minimum of the energy function, which can be performed by Simulated Annealing or Simulated Tempering, while keeping the χp state of all the nucleotides fixed in the *anti* and C3′-*endo* states. These calculations will be referred as χp_c simulations along the paper, to stress the constraint on these variables. Then, a shorter annealing procedure is run on a smaller set of nucleotides, typically a junction or an internal or hairpin loop, without any constraint on the χp states. In this manner, for example, we anneal junctions and internal and hairpin loops by anchoring the molecule to a rigid vicinity, which can be determined from a previous χp_c simulation or

from a crystal structure. However, for larger structures such as duplexes and pseudoknots, we have opted for taking into account their natural flexibility during this refinement step. To this aim, we apply a soft restraint on the secondary structure during the refinement.

For the folding simulations, we used a simulated annealing protocol with 20 initial conditions. For the anchored tetraloops, each step of the annealing procedure consisted of 5×10^7 Monte Carlo trials on each nucleotide, saving conformations every 5000 steps. These parameters had slight modifications in the rest of the systems. The annealing procedure started at temperature $T = 15T_0$. T was multiplied by a factor 0.75 when the minimum energy of an annealing step did not decrease [52]. Once it reached a value smaller than T_0, the temperature was set to zero so as to minimize the energy of the resulting structure and run for a time equal to the one of an annealing step. This is in general more than enough to obtain a converged structure . In addition, we also performed Simulated Tempering simulations for the pseudoknot, which is also implemented in our code. We used the method of [53] for estimating the initial values of the relative weights, with a maximum temperature of 12.5 T_0 and minimum of 0.5 T_0, using 12 temperatures separated by $\Delta T = T_0$.

Parametrization

We parametrize the set of ϵ shifts of stacking and base–phosphate interactions by folding a small set of structures. The base-pairing interactions are scaled according to the number of hydrogen bonds between two nitrogen or oxygen containing groups according to the Leontis-Westhof classification [10]. The energy scale of stacking interactions has been adjusted to obtain an initial value that will be refined in a posterior step. We start with a trial value strong enough to correctly fold the stem of the GCAA tetraloop. Later on, the backbone interactions are multiplied by a prefactor, which is required to obtain the correct arrangement of bases in the loop region. In a following step, we introduce the base–phosphate interactions, and adjust their strength in order to keep the formation of the stem stable. Later, the stacking interactions have been adjusted by folding a duplex which contains the UU stacking (PDB ID: 255D), providing a lower bound to their strength. Thus, the stacking strengths are parametrized to distinguish between purines and pyrimidines (see Supplementary Data Section 2). The parameters obtained were later on refined by annealing a set of 46 internal and hairpin loops from the 1S72 ribosomal structure. On the resulting structures, the native one was not always obtained, and a number of alternative intermediate structures was often obtained as well. We determined the minimum energy structure for a large set of parameters of stacking (see Supplementary Data Section 2), and calculated the INF score only for the stacking interactions. Thus, we chose the set of parameters which maximized this score over the entire training set. In a similar manner, the $\epsilon^{\chi p}$ was obtained by recalculating the energy of these decoys and maximizing the number of correctly predicted cases. The detailed values are reported in the Supplementary Data Section 3, showing that, after the parametrization, the *anti* glycosidic conformation with the C3'-*endo* sugar pucker is still

the most energetically favorable one, although not as drastically as when no shift is applied.

Backmapping

Once the minimum energy CG structure is identified, we produce an all-atom prediction by performing steered-MD simulations. The MD simulation is performed using an atomistic description of RNA in explicit water (TIP3P water molecules [54], Amber99 force field [55] with parmbsc0 [56] and χ_{OL3} corrections [57]) in a truncated dodecahedral box with Na^+ counterions [58]. An external force proportional to the $\mathcal{E}$RMSD [45] with respect to the target structure is applied using the Gromacs code 4.6.7 [59] in combination with PLUMED [60]. This means that the bases are positioned according to the CG structure, while the backbone atoms simply adjust to this arrangement. The temperature was of 350K, while the pulling constant was of 500 kJ/nm^2 and the $\mathcal{E}$RMSD cutoff was of 3.6. The possibility to use the $\mathcal{E}$RMSD as a steered variable has been already discussed in [18] where it was used so as to enforce the correct fold in stem-loop structures.

Assessment of prediction quality

We use the standard root-mean-square deviation after optimal superposition (RMSD [61]) and the Interaction Network Fidelity (INF) to compare the predicted structures with their native counterpart. The INF is given by [62]

$$\mathrm{INF} = \sqrt{\left(\frac{TP}{TP + FP} \right) \times \left(\frac{TP}{TP + FN} \right)} \qquad (5)$$

where, *TP* stands for the correctly predicted contacts while *FP* and *FN* are the false positive and false negative numbers, respectively. This coefficient is calculated separately for stacking (st), canonical pairs (wc), non-canonical pairs (nc) and base-phosphate (bph) pairs. The RMSD was calculated by using the position of the sugar, phosphate and backmapping the C2, C4 and C6 atoms from the CG representation.

RESULTS

In the following we discuss the results obtained using our model on a series of benchmark systems for which the native structure is already known by means of X-ray or nuclear magnetic resonance spectroscopy.

As mentioned in the 'Materials and Methods' section, the simulation protocol might be applied in two steps for large structures, while for small motifs it can be applied without any χp_c constraint, but keeping the RNA motif anchored to its neighbors. These choices will be specified in each subsection.

Validation of the annealing procedure

We have observed including the change of both glycosidic bond angle conformation and sugar pucker increases considerably the complexity of the conformational space and, consequently, the simulation time. We thus perform the annealing simulations without restrictions in the χp space on a
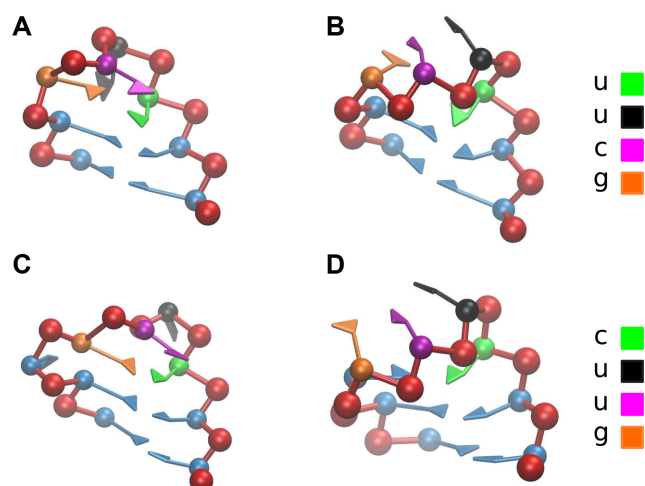
Figure 3. (**A**) UUCG tetraloop predicted by SPQR. (**B**) UUCG tetraloop predicted by χp_c annealing. (**C**) CUUG tetraloop, closest structure to native and (**D**) CUUG tetraloop predicted by χp_c simulations.

selected set of systems only, in order to validate the two-step procedure. In particular, we perform a *de novo* prediction of a duplex formed of the sequence CCCCGGGG (PDB ID: 1RXB) and a hairpin of sequence GGGCGCAAGCCU (PDB ID: 1ZIH). In both cases, the contacts and three-dimensional structure are nicely recovered. Equivalent results were also produced using the two-step procedure, which evidences the robustness of our approach.

Tetraloops

We have tested our approach on a set of tetraloops which include several GNRA occurrences and the UUCG and CUUG sequences. Details about the sequences and native structures can be found in the Supplementary Data Section 4. The loop structures are optimized by a single round of Simulated Annealing while keeping them attached to a fixed stem obtained from the crystal structure. Only the tetraloops and the closing base pair are allowed to move during the simulation (six flexible nucleotides). The total number of nucleotides in the hairpin loops is eight for GNRA and UUCG and nine for CUUG. The change of the χp state is allowed only on the four nucleotides in the tetraloop. The results are shown in Table 1. We can observe the comparison between the systems with fixed χp_c simulations. In the case of GNRA and UUCG, the pucker and glycosidic bond angle conformations are correctly predicted, and the bases orient themselves accordingly leading to an excellent agreement. Particularly, in UUCG, two bases are flipped: a purine (G9), by virtue of its glycosidic bond angle, and a pyrimidine (C8), due to its sugar pucker conformation, which is reconstructed by SPQR as shown in Figure 3A and B. In the case of CUUG, the native structure was observed as the second most energetically favorable structure. The energy difference between the native structure and the one predicted as most stable by SPQR was only of the order of the energy of a single hydrogen bond. Nevertheless, the agreement is remarkable, reproducing both the loop re-
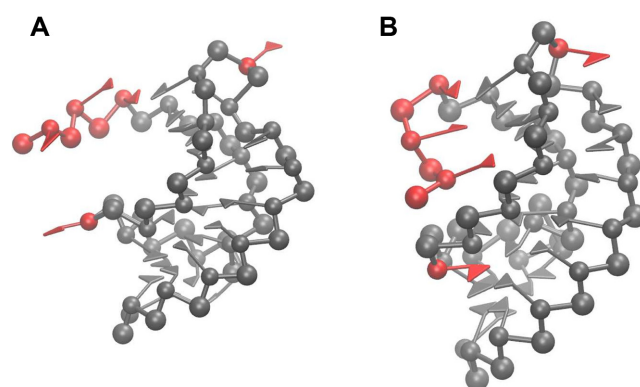
Figure 4. (**A**) Native pseudoknot 1L2X, (**B**) folded after refinement. To facilitate the comparison between the two structures, the three initial nucleotides and two bulges are colored in red.

gion and a bulge as shown in Figure 3C, compared to its χp_c counterpart in Figure 3D.

Double-stranded structures and pseudoknot

We also tested our method on a set of double stranded-structures containing non-canonical pairs, as well as on a pseudoknot. For these systems, the two-step procedure was applied as described in the 'Materials and Methods' section, that is, we performed χp_c simulations which are later refined imposing secondary structure restraints. The results before and after the refinement are shown in Table 2. For the pseudoknot, all the canonical base pairs, with the exception of a pair of bulged bases, were correctly predicted.

Once the duplexes have been successfully folded with constraints, they are remarkably stable. In general, the results are not greatly affected by the additional degrees of freedom. An unfavorable case is 1KD5, where a slightly less favorable conformation appears with a similar energy of the native with two puckers wrongly predicted. A similar case is observed in 402D, where an incorrect structure has practically the same energy as a better structure listed in the table. For the 1L2X pseudoknot structure we obtain a reasonable agreement without sampling glycosidic bond angle and pucker. However, it turns out that the refinement improves the INF corresponding to non-canonical pairs, and reduces the number of false positive base–phosphate interactions from five to two. For this system, the number of true positives is zero, so that INF is zero irrespectively of the number of false positives. We also see that some bases can flip to *syn* state (the only cases on the present set), although they are unpaired or bulged. In addition, the first two unpaired nucleotides turn more flexible under the refinement. By neglecting them from the analysis, the RMSD improves to 3.3Å. A depiction of this structure is found in Figure 4.

Anchored internal loops and junctions

We simulated a small subset of the FARFAR motifs (47). For a detailed description, see the Supplementary Data Section 1. In this case, each motif is completely flexible while the nucleotides of its environment are frozen in space. This included the initial and terminal nucleotides of the

Table 1. RMSD and INF scores of the predicted tetraloops

Structure	RMSD (Å)	INF	INF^{wc}	INF^{nc}	INF^{bph}
GCAA	1.4	1	1	1	1
GAAA	1.7	1	1	1	1
GAGA	1.2	1	1	1	1
GCGA	1.8	1	1	1	1
GGAA	2.1	1	1	1	1
GGGA	1.3	1	1	1	1
GUAA	1.3	1	1	1	1
GUGA	1.5	1	1	1	1
UUCG	0.9	1	1	1	1
UUCG(χp_c)	2.7	0.86	1	0	0
CUUG(s)	1.8	0.94	1	-	0
CUUG(χp_c)	3.5	0.7	0.81	0	-

Results for the tetraloop simulations allowing the change of sugar pucker and χ angle conformation. Results forcing these parameters to the C3′-*endo* and *anti* conformation in all the nucleotides are denoted with (χp_c). Also, the second best result for CUUG is denoted by (s), and included for completeness.

Table 2. RMSD and INF scores for the double-stranded structures and pseudoknot predictions

Structure	RMSD (Å)	INF	INF^{wc}	INF^{nc}	INF^{bph}
157D	1.6	0.98	1	1	-
157D(χp_c)	1.5	0.98	1	1	-
1DQH	2.3	0.93	1	-	-
1DQH(χp_c)	2.1	0.93	0.94	-	-
1KD5	4.4(3.6)	0.81	1	0	-(0)
1KD5(χp_c)	3.6	0.89	1	0	-
1SA9	2	0.92	1	0	-
1SA9(χp_c)	2	0.92	1	0	-
205D	1.4	0.93	1	1	-
205D(χp_c)	1.6	0.95	1	1	-
402D	3.3(1.4)	0.89	1	0(1)	-
402D(χp_c)	1.4	0.89	1	1	-
1I9X	4.2	0.85	0.83	-	-
1I9X(χp_c)	4.5	0.85	0.83	-	-
1L2X	4	0.84	0.94	0.42	0
1L2X(χp_c)	4.4	0.81	0.94	0.22	0

The values correspond to the scores of the minimum energy conformation found on each case. In parentheses are the values of structures that were found with an energy practically indistinguishable to the minimum.

strands involved, plus any interacting nucleotide according to FR3D or a nucleobase with a C4′ atom separated by a distance smaller than 12Å from any atom of the motif. The results are presented in Table 3.

The results are reasonable, showing that in many cases, when the χp_c simulations produce good enough results, the additional degrees of freedom introduced by the χp refinement do not compromise the structure.

On average, the χp refinement improves slightly the values of the INF for non-canonical base pairs. Nevertheless, it makes possible the formation of certain motifs which are technically forbidden without the additional freedom of the sugar pucker and glycosidic bond angle as in 2GDI (see Supplementary Data Section 5). In addition, base flips seem not to cause a major problem; in fact, the only case they appear spuriously is in unpaired bases of 1JJ2. The INF values obtained here are comparable to the values obtained with SimRNA for the same motifs, although the authors have reported a smaller RMSD.

Finally, Table 4 shows the results for the junctions.

Here the results are in general worse than in the previous cases. Although the χp refinement does not improve the results when they are already bad, it does improve them when the χp_c predictions are good. Taken together, these results suggest that in order to be reliably used to predict junctions, the model should likely be further refined.
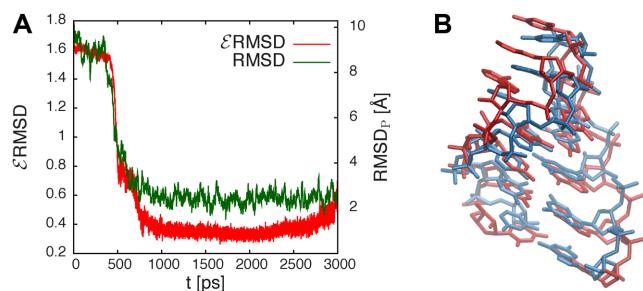
Backmapping

The backmapping procedure has been applied over two representative tetraloops (GCAA, PDB ID:1ZIH and UUCG, PDB ID: 2KOC), using six initial conditions starting from a free strand and MD simulations of 3 ns. The pulling force is proportional to the $\mathcal{E}$RMSD between the bases of the atomistic structure and the lowest-energy CG structure found in the annealing procedure and thus effectively steers the atomistic model toward the structure predicted using the CG model. We stress that since this $\mathcal{E}$RMSD calculation is done with respect to the predicted model, its evaluation does not require any knowledge of the native structure. The $\mathcal{E}$RMSD converges quickly, as seen in Figure 5, reaching a value of 0.4 or lower when the folding is successful. The RMSD of the phosphates also converges in the same time scale. This result is not straightforward, since the $\mathcal{E}$RMSD only measures the relative nucleobase arrangement. By choosing the lowest $\mathcal{E}$RMSD structures of the simulations, we find an all-atom RMSD (without hydrogen atoms) from native of 2.3 Å for 1ZIH and 1.9 Å for 2KOC.

Table 3. RMSD and INF scores for subset of FARFAR motifs

Structure	RMSD(Å)	INF^{st}	INF^{wc}	INF^{nc}	INF^{bph}
$157D_1$	0.7	1	1	1	-
$157D_1(\chi p_c)$	0.7	1	1	1	-
$1D4R_1$	0.7	0.94	1	1	-
$1D4R_1(\chi p_c)$	0.6	0.94	1	1	-
$1JJ2_1$	5	0.55	1	0	0
$1JJ2_1(\chi p_c)$	3.7	0.62	1	0.25	0
$1LNT_1$	1	0.96	1	0.89	1
$1LNT_1(\chi p_c)$	1.1	0.96	1	0.89	1
$1Q9A_1$	1.1	0.91	1	1	1
$1Q9A_1(\chi p_c)$	0.9	0.91	1	1	1
$1U9S_1$	5.9	0.22	0.71	-	-
$1U9S_1(\chi p_c)$	4.6	0.67	0.71	-	0
$2GDI_1$	0.9	1	1	-	-
$2GDI_1(\chi p_c)$	2.3	0.8	1	-	-
$2OUI_1$	1.8	0.75	1	0.5	-
$2OIU_1(\chi p_c)$	2.1	0.77	0.87	0	-
$2R8S_1$	1	0.88	1	1	-
$2R8S_1(\chi p_c)$	1	0.88	1	1	-
$2R8S_3$	3.6	0.53	1	0.29	-
$2R8S_3(\chi p_c)$	5.5	0.5	1	0.29	-

Table 4. RMSD and INF scores of the set of junctions

Structure	RMSD (Å)	INF	INF^{wc}	INF^{nc}	INF^{bph}
$1GID_2$	7.9	0.47	1	0.63	0.71
$1GID_2(\chi p_c)$	3	0.55	1	0.71	0.35
$2QBZ_3$	8.4	0.30	0.58	0	0
$2QBZ_3(\chi p_c)$	6.5	0.41	0.87	0	0
$3R4F_2$	3.1	0.59	1	-	-
$3R4F_2(\chi p_c)$	2.8	0.6	1	-	-
$4P8Z_2$	2.2	0.47	1	-	0
$4P8Z_2(\chi p_c)$	2.9	0.47	1	-	0
$4P9R_2$	3.3	0.57	1	-	-
$4P9R_2(\chi p_c)$	2.7	0.52	1	-	-

Figure 5. (**A**)$\mathcal{E}$RMSD and phosphate RMSD as a function of time for backmapping of 1ZIH. (**B**) Comparison of GCAA tetraloop(1ZIH) at full-atom resolution for native (blue) and annealed (red).

was identified with Suitename. Structures with not classified backbone conformations were discarded in the analysis. We observe that for GCAA in 84% of the cases the full tetraloop has the right glycosidic bond conformation in all its nucleotides, from which 92% had the right conformation in all the sugar puckers at the same time. On the other side, UUCG shows 76% of right glycosidic conformations, from which 85% of them have the right puckers. Note that these loops are not static in solution, so it is not unexpected to observe a variety of rotameric states ([63]).

DISCUSSION AND CONCLUSION

We proposed a CG model with a hybrid parametrization for folding small RNA motifs. The model pays special attention to the non-canonical base pairs, base phosphate interactions, sugar puckers and glycosidic bond torsions, which are the key elements that determine, from our study, the geometrical arrangement of nucleotides from the perspective of a base. It is important to observe that whereas the shape of the distributions associated to each interaction type is obtained from the employed non-redundant list of structures, the strength of the base-pairs interactions follows the classification presented in ([10]). This gives more relevance to certain base-pairs that are not very frequent but could have a significant energetic contribution. It also depends of both faces of the bases when they are confronted. There-

The backmapping procedure is not always successful. In some cases, the base-flipping, like the one of G8 in the UUCG tetraloop, is not observed. This is an indication that the steering procedure was too fast, not letting the atomistic model enough time to relax. Nevertheless, the $\mathcal{E}$RMSD of these structures is above 0.6, which makes these instances easy to detect without any knowledge of the native structure.

By taking the 90 lowest $\mathcal{E}$RMSD structures, that is the atomistic structures that better approximate the predicted CG model, we have analyzed the glycosidic bond torsion and sugar puckers in the tetraloop region. The sugar pucker

fore, the correction also compensates the different normalizations which cannot be captured by the usual approximations that consider, in a first instance, the interactions as in an structureless liquid (64). An equivalent procedure is also used to partition the space according to the χp values.

The procedure used for the parametrization of our model makes it a hybrid in between models which are based on purely statistical potentials (e.g. (21,32)) and methods that are trained based on the energetics of the relevant interactions (e.g. (38)).

We also have opted for leaving out the pseudotorsional space (65) interactions. Although there is evidence of the correlation of these variables and the total RMSD of certain structures, it is not clear whether they are indeed the cause of the form of each loop, and they are not able to distinguish between sugar puckers. We observe that our representation sufficed in many cases for the correct prediction of several motifs. This suggests that the incorporation of additional contributions to the potential is likely not necessary at this time. We also noticed that incorporating additional potentials based on statistics from structural databases may shift the predicted structures to more frequent conformations in the reference database, e.g. A-form duplexes, if not properly orthogonalized to the already present contributions.

Although the parametrization we present is not unique and could be further refined, the simple choice we have used in this paper has been enough for folding a number of structures which include canonical and non-canonical pairs. Even more, we have shown that the inclusion of puckers and glycosidic bond torsions allows for an improvement in the predicted structures in cases where rare conformations (*syn* and C2′-*endo*) are present in the native structure, without significantly destabilizing the most common conformations. Also, crucial base-phosphate interactions have been successfully recovered. Considering the complexity of the interactions, it is also remarkable that the annealing from random conformations is able to reproduce both duplexes and hairpins with good accuracy of bulges and non-canonical pairs.

The blind prediction tests give results that are comparable to those obtained with SimRNA (32). However, it should be mentioned that an intrinsic advantage of our model is that it allows base flipping to be explicitly modeled changing both sugar puckers and glycosidic bond angle. We thus think that SPQR could be optimally used side-by-side with other modeling tools that do not take these degrees of freedom explicitly into account.

We have also implemented a backmapping procedure able to introduce atomistic detail on the candidate structures. Although the procedure is designed to only enforce the position and orientation of the bases, it is able to backmap reasonably well the whole structure, with a good agreement of the sugar puckers and glycosidic bond torsion conformations in most cases. This allows our protocol to be used for structure prediction at atomistic resolution. We recall that atomistic MD is still impaired by sampling issues and by the accuracy of force fields and it is not able to predict correctly such a variety of motifs. On the opposite side, structural bioinformatics tools are typically designed so as to quickly model or fold larger structures with an accuracy that does not allow individual contacts to be reliably pre-

dicted. Our model is just optimally suited in the middle, and might be useful for refining structures obtained with other tools so as to bring them at atomistic resolution.

The CG simulations were performed using an in-house Monte Carlo code, which can be downloaded at http://github.com/srnas/spqr. The folding of a 12-nt hairpin may take around 6 h in a desktop PC. Depending on the number of resources available, this procedure can be parallelized for a better exploration of the folding space, although both the code and the annealing procedure can be improved. The typical cost of a backmapping simulation with our procedure is 4 h on twenty processors for a 12-nt hairpin. This time might seem large, considering that most of the available CG models allow for the atomistic representation to be constructed in a single step, typically by assembling fragments that are compatible with the CG representation. However, the timescale of the backmapping simulation itself is comparable or even shorter than that of the structure prediction using the CG model, so that the total simulated time is less than doubled if only the best prediction is to be refined. In addition, if one is willing to use the resulting structure as an input for an atomistic simulation, our procedure has the advantage of producing a fully solvated and equilibrated structure. Finally, the backmapping procedure itself, being based on a force field with a reasonable accuracy, can help in identifying nucleotides that have been wrongly placed by the CG simulation. The similarity between the timescales of the CG simulation and of the backmapping procedure naturally calls for a synergic simulation protocol where both representations are simultaneously evolved, that will be the subject for further investigation.

In this paper we presented a novel protocol for RNA structure prediction and refinement. The employed parametrization and the separation of the variables along the backbone were shown to be the key ingredients for the formation of tetraloops belonging to three families, as well as for the correct prediction of a number of hairpins, duplexes, internal loops, junctions and a pseudoknot, featuring multiple non-canonical interactions. Future work will contemplate a further refinement of the base–base interactions and an extensive validation on a larger number of motifs taken from structural databases.

DATA AVAILABILITY

The CG simulations were performed using an in-house Monte Carlo code, which can be downloaded at http://github.com/srnas/spqr.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Lehman,N. (2010) RNA in evolution. *Wiley Interdisc. Rev. RNA*, **1**, 202–213.
2. Doudna,J.A. and Cech,T.R. (2002) The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
3. Mattick,J.S. (2004) RNA regulation: a new genetics?*Nat. Rev. Genet.*, **5**, 316–323.
4. Serganov,A. and Nudler,E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.
5. Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.
6. Tinoco,I. Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
7. Murray,L.W., Arendall,W.B. III, Richardson,D.C. and Richardson,J.S. (2003) RNA backbone is rotameric. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 13904–13909.
8. Richardson,J.S., Schneider,B., Murray,L.W., Kapral,G.J., Immormino,R.M., Headd,J.J., Richardson,D.C., Ham,D., Hershkovits,E., Williams,L.D. *et al.* (2008) RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA*, **14**, 465–481.
9. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
10. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The nonWatsonCrick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
11. Zirbel,C.L., Šponer,J.E., Šponer,J., Stombaugh,J. and Leontis,N.B. (2009) Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.*, **37**, 4898–4918.
12. Cruz,J.A., Blanchet,M.F., Boniecki,M., Bujnicki,J.M., Chen,S.J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **14**, 610–625.
13. 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.
14. Miao,Z., Adamiak,R.W., Antczak,M., Batey,R.T., Becka,A.J., Besiada,M., Boniecki,M., Bujnicki,J.M., Chen,S.J., Cheng,C.Y. *et al.* (2017) RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA*, **23**, 655–672.
15. Šponer,J., Banas,P., Jurecka,P., Zgarbova,M., Kuhrova,P., Havrila,M., Krepl,M., Stadlbauer,P. and Otyepka,M. (2014) Molecular Dynamics Simulations of Nucleic Acids. From Tetranucleotides to the Ribosome. *J. Phys. Chem. Lett.*, **5**, 1771–1782.
16. 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.
17. Bergonzo,C., Henriksen,N.M., Roe,D.R. and Cheatham,T.E. 3rd (2015) Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA*, **21**, 1578–1590.
18. Bottaro,S., Banaš,P., Šponer,J. and Bussi,G. (2016) Free energy landscape of GAGA and UUCG RNA tetraloops. *J. Phys. Chem. Lett.*, **7**, 4032–4038.
19. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
20. Reinharz,V., Major,F. and Waldispuhl,J. (2012) Towards 3D structure prediction of large RNA molecules: an integer programming framework to insert local 3D motifs in RNA secondary structure. *Bioinformatics*, **28**, i207–i214.
21. 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.
22. Flores,S.C., Wan,Y., Russel,R. and Altman,R.B. (2010) Predicting RNA structure by multiple template homology modeling. *Pac. Symp. Biocomput.*, **2010**, 216–227.
23. Jossinet,F., Ludwig,T.E. and Westhof,E. (2010) Aseemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*, **26**, 2057–2059.
24. Martinez,H.M., Maizel,J.V. Jr and Shapiro,B.A. (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. *J. Biomol. Struct. Dyn.*, **25**, 669–683.
25. Bida,J.P. and Maher,L.J. III (2012) Improved prediction of RNA tertiary structure with insights into native state dynamics. *Bioinformatics*, **18**, 385–393.
26. Eggenhofer,F., Hofacker,I.L. and Höner zu Siederdissen,C. (2016) RNAlien—unsupervised RNA family model construction. *Nucleic Acids Res.*, **44**, 8433–8441.
27. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **36**, 1227–1236.
28. Popenda,M., Szachniuk,M., Antczak,M., Purzycka,K.J., Lukasiak,P., Bartol,N., Blazewicz,J. and Adamiak,R.W. (2012) Atumated 3D structure composition for large RNAs. *Nucleic Acids Res.*, **40**, e112.
29. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
30. Xia,Z., Gardner,D.P., Gutell,R.R. and Ren,P. (2010) Coarse-grained model for simulation of RNA three-dimensional structures. *J. Phys. Chem. B*, **114**, 13497–13506.
31. Pasquali,S. and Derreumaux,P. (2010) HiRE-RNA: a high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B.*, **114**, 11957–11966.
32. Boniecki,M.J., Lach,G., Dawson,W.K., Tomala,K., Lukasz,P., Soltysinski,T., Rother,K.M. and Bujnicki,J.M. (2015) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res.*, **44**, e63.
33. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
34. Tan,R.K.Z., Petrov,A.S. and Harvey,S.C. (2006) YUP: A molecular simulation program for coarse-grained and multi-scaled models. *J. Chem. Theory Comput.*, **2**, 529–540.
35. Li,J., Zhang,J., Wang,J., Li,W. and Wang,W. (2016) Structure prediction of RNA loops with a probabilistic approach. *PLoS Comput. Biol.*, **12**, e1005032.
36. Liwo,A., Baranowski,M., Czaplewski,C., Golas,E., He,Y., Jagiela,D., Krupa,P., Maciejczyk,M., Makowski,M., Mozolewska,M.A. *et al.* (2014) A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions. *J. Mol. Model.*, **20**, 2306.
37. Xu,X., Zhao,P. and Chen,S.-J. (2014) Vfold: A web server for RNA structure and folding thermodynamics prediction. *PLoS One*, **9**, e107504.
38. Denesyuk,N.A. and Thirumalai,D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B*, **117**, 4901–4911.
39. Mustoe,A.M., Al-Hashimi,H.M. and Brroks,C.K. III (2014) Coarse grained models reveal essential contributions of topological constraints to the conformational free energy of RNA bulges. *J. Phys. Chem. B*, **118**, 2615–2627.
40. Bernauer,J., Huang,X., Sim,A.Y.L. and Levitt,M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
41. Kim,N., Zahran,M. and Schlick,T. (2015) Computational prediction of riboswitch tertiary structures including pseudoknots by RAGTOP: a hierarchical graph sampling approach. *Methods Enzymol.*, **553**, 115–135.
42. Kerpedjiev,P., Höner zu Siederdissen,C. and Hofacker,I.L. (2015) Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, **21**, 1110–1121.
43. Jost,D. and Everaers,R. (2010) Prediction of RNA multiloop and pseudoknot conformations from a lattice-based, coarse-grain tertiary structure model. *J. Chem. Phys.*, **132**, 095101.
44. Sulc,P., Romano,F., Ouldridge,T.E., Doye,J.P.K. and Louis,A.A. (2014) A nucleotide-level coarse-grained model of RNA. *J. Chem. Phys.*, **140**, 235102.
45. Bottaro,S., Di Palma,F. and Bussi,G. (2014) The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.*, **42**, 13306–13314.
46. Bottaro,S., Di Palma,F. and Bussi,G. (2015) Towards de novo RNA 3D structure prediction. *RNA Dis.*, **2**, e544.

47. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.

48. Leontis,N.B. and Zirbel,C.L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In: Leontis,N and Westhof,E (eds). *RNA 3D Structure Analysis and Prediction*, Vol. **27**, Springer, Berlin, pp. 281–298.

49. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.

50. Olson,W.K. (1975) Configurational statistics of polynucleotide chains. A single virtual bond treatment. *Macromolecules*, **8**, 272–275.

51. Frenkel,D. and Smit,B. (1996) *Understanding Molecular Simulation: from Algorithms to Applications*. Academic Press, San Diego.

52. Snow,M.E. (1991) Powerful simulated-annealing algorithm locates global minimum of protein-folding potentials from multiple starting conformations. *J. Comput. Chem.*, **13**, 579–584.

53. Park,S. and Pande,V.S. (2007) Choosing weights for simulated tempering. *Phys. Rev. E*, **76**, 016703.

54. 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.

55. 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) Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.

56. 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.

57. Banas,P., Hollas,D., Zgarbová,M., Jurecka,P., Orozco,M., Cheatham,T.E. III, Sponer,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.

58. 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.

59. Hess,B., Kutzner,C., van der Spoel,D. and Lindahl,E. (2008) Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

60. Tribello,G.A., Bonomi,M., Branduardi,D., Camilloni,C. and Bussi,G. (2014) PLUMED2: New feathers for an old bird. *Comp. Phys. Comm.*, **185**, 604–613.

61. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, **A34**, 827–828.

62. 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.

63. Bottaro,S. and Lindorff-Larsen,K. (2017) Mapping the universe of RNA tetraloop folds. *Biophys. J.*, **113**, 257–267.

64. Skolnick,J., Kolinski,A. and Ortiz,A. (2000) Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, **38**, 3–16.

65. Wadley,L.M., Keating,K.S., Duarte,C.M. and Pyle,A.M. (2007) Evaluating and learning from RNA pseudotorsional space: Quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol.*, **372**, 942–957.