Coarse-graining RNA nanostructures for molecular dynamics simulations

2010 Phys. Biol. 7 036001

(http://iopscience.iop.org/1478-3975/7/3/036001)

Phys. Biol. 7 (2010) 036001 (12pp) doi:10.1088/1478-3975/7/3/036001

Coarse-graining RNA nanostructures for molecular dynamics simulations

Maxim Paliy[1,2], **Roderick Melnik**[1,3] **and Bruce A Shapiro**[4]

[1] M^2NeT Lab, Wilfrid Laurier University 75 University Avenue West Waterloo, ON, N2 L 3C5, Canada (http://www.m2netlab.wlu.ca)
[2] Surface Science Western, University of Western Ontario, London, ON, N6A 5B7, Canada
[3] BCAM, Bizkaia Technology Park, 48160 Derio, Spain
[4] Center for Cancer Research Nanobiology Program, National Cancer Institute, Frederick, MD 21702, USA

E-mail: mpaliy@wlu.ca

Received 18 April 2010
Accepted for publication 28 May 2010
Published 24 June 2010
Online at stacks.iop.org/PhysBio/7/036001

Abstract

A series of coarse-grained models have been developed for study of the molecular dynamics of RNA nanostructures. The models in the series have one to three beads per nucleotide and include different amounts of detailed structural information. Such a treatment allows us to reach, for systems of thousands of nucleotides, a time scale of microseconds (i.e. by three orders of magnitude longer than in full atomistic modeling) and thus to enable simulations of large RNA polymers in the context of bionanotechnology. We find that the three-beads-per-nucleotide models, described by a set of just a few universal parameters, are able to describe different RNA conformations and are comparable in structural precision to the models where detailed values of the backbone P-C4′ dihedrals taken from a reference structure are included. These findings are discussed in the context of RNA conformation classes.

S Online supplementary data available from stacks.iop.org/PhysBio/7/036001/mmedia

1. Introduction

Recent progress in understanding RNA structure brought to light a new concept of RNA architectonics—a set of recipes for (self-)assembly of RNA nanostructures of arbitrary size and shape [1, 2]. The smallest RNA building blocks, ‘tectoRNAs’, typically bearing well-defined structural features, such as ‘right angle’ [1], ‘kink-turn’ [2, 3] or ‘RNAIi/RNAIIi’ [4] motifs, were manipulated, either experimentally [1, 2] or via computer modeling [4], into desired 2D or 3D nanostructures (squares, hexagons, cubes, tetrahedrons, etc) that can be further assembled into periodic or quasi-periodic patterns. Compared to DNA nanostructures, RNA as a nano-engineering material brings additional challenging features, such as much larger diversity in tertiary structural building blocks (∼200 versus ∼20 for DNA [2]) and, often, increased conformational flexibility (see e.g. [5], p 320). Useful insights about the behavior of the above-mentioned nanostructures can be gained by all-atom molecular dynamics (MD) simulations [5]. However, presently, the time scales that can be achieved in all-atom MD amount to a few (tens) of nanoseconds only, which is less by many orders of magnitude than the duration of the slowest processes occurring in biomolecules (micro- to milli- to seconds). For example, in a recent study [6], we analyzed, via all-atom MD simulation, a simple RNA nanostructure of about 13 nm in size (330 nucleotides), a hexagon-shaped RNA ring [4], termed ‘nanoring’. It is composed (figure 1) of six RNAIi/RNAIIi complexes, each of which is made up of an A-form double helical side joined by the ‘kissing loop’ motifs at each end (e.g. AACCAUC septaloop is paired with UUGGUAG loop). Figure 2 shows patterns of base pairing and stacking in the kissing loops.

In order to reach at least microsecond time scales in simulations of such nanostructures (hundreds to thousands nucleotides in size), one needs to consider a coarse-grained (CG) treatment, where the groups of atoms are represented by CG interaction centers, ‘beads’, and effective interactions between such beads are set in a way to fit the nanostructure’s atomic connectivity, thermal, mechanical properties, etc. Two kinds of data are often used in the fitting process: (i) the experimentally available structural information as well as other known properties of interest (which can be limited and/or incomplete), and (ii) a host of very detailed atomistic data obtained from all-atom MD simulations. Namely, the parameters for a CG model can be derived from both experimental and full-atom MD data by Boltzmann inversion (BI) [7] of the radial distribution functions (RDFs), using the inverse Monte Carlo scheme [8] or, in the case of all-atom MD simulation only, with the 'force matching' method [9] (for some recent approaches to biomolecules, see e.g. [10–15]). Finally, such a CG model can be further investigated using coarse-grained molecular dynamics (CGMD), which allows one to reach much longer time scales (although the dynamics of the original system is not always adequately represented [11]). The main challenge is to describe the RNA on a CG level with just a few universal parameters, thus adopting the strategy of a 'CG force field' (FF). This proved to be a difficult task, in particular in the sense of transferability of such a CG model to other structures, not used in the fitting. Transferability problems are notorious even for condensed matter, and it is even more true for the biomolecules that show enormous structural diversity, see e.g. [10, 16–18]. In the latter case, instead of a FF approach, a lot of detailed structural information (such as equilibrium values of bonds, angles, dihedrals, nonbonded interatomic distances from the experimentally resolved structures) is supplied to the CG model, thus providing its structural precision, although in describing a given structure only. Such a structurally biased approach is often termed the self-organized polymer (SOP) [19].

In the CG modeling of RNA, several recent advances should be highlighted [19–34]. Besides, many of the RNA modeling approaches focusing on the 3D tertiary structure prediction (reviewed e.g. in [35]) make use of some coarse graining via variety of methods, such as building complex CG energy functions and use of pre-compiled databases of known motifs. However, comparatively few of them are able to produce long-time dynamics once the 3D structure is known. Existing CG models for RNA, similar to those for DNA [16, 36, 37], include one or more CG units (beads) per nucleotide. Simplest one-bead RNA models [20–23] normally use the SOP approach to achieve the desired structural precision (most recent model developed in [23] is of FF type and uses special bonds to fix the tertiary contacts only). Three-beads-per-nucleotide choice gained popularity after having been introduced in [19] in the SOP variant, and it has already been employed both for DNA [36, 37] and for RNA [25, 26, 28]. On the other end of the spectrum, highly precise but more complex CG models have been recently proposed for RNA/DNA, e.g. the one of the FF style with 12–13 beads per nucleotide [34], or the one that includes all heavy atoms with a simplified SOP-style energy function [32, 33]. The most recent three-bead RNA CG model [25, 26] achieves good structural precision (for up 100-nucleotide chains) even with a CG FF approach that includes complex sequence-dependent interaction details and many-body treatment of the stacking, base pairing and hydrophobic interactions (for some larger structures, it still needs the special artificially introduced bonds to fix long-range tertiary contacts [26]). However, the technique employed in [25] to evolve the system, discrete MD, while providing fast folding, at the same time dictates the use of simplified step-wise potentials, which may introduce additional imprecision into the model. By contrast, we use continuous potentials to describe the energetics of RNA, which should make the dynamics of the system more realistic. To put our RNA CG modeling efforts into better perspective relative to previous studies, let us emphasize that we will focus on the development of a simple continuous CG energy function that enables us to do long-time CGMD simulations of very large aggregates such as RNA nanostructures. This simplicity comes at the expense, for example, of needing to know the secondary structure of our models, a requirement that will be lifted in the future.

In this paper we explore avenues from the SOP approach to a RNA CG FF approach by varying the number of beads and amount of atomistic structural information in our series of models. We show that the inclusion of just the dihedral pseudo angles P-C4′ in a SOP manner brings about the same structural precision to the model as the full SOP approach. Besides, a simple modification of the dihedral angle terms to allow an alternative value of each dihedral can render even the RNA FF model sufficiently precise to describe the studied RNA nanoring. These findings are consistent with the existence of the RNA conformation classes, based on P-C4′ dihedrals [38].

2. Coarse-grained model

The development of a CG model consists of two major stages: (i) choice of the groups of atoms to be combined in a single CG bead, and (ii) selection of the functional forms and fitting of the parameters for the effective interactions between the beads.

In the case of nucleic acids, the simplest choice for stage (i) is one bead per nucleotide (beads being placed normally on the phosphate groups which allows one to use experimentally available structural data [15, 22]). However, as has been recently shown in [38], the experimentally available RNA conformations may be described well by just two torsional angles, between the P and C4′ atoms. Connecting the beads placed at the C4′ atomic sites with base-pairing bonds is not suitable in terms of the geometry, as such bonds are too far off the axis of the double helix. Instead, we adopt a representation with three beads per nucleotide that corresponds to the (P)hosphate, (S)ugar and nucleic (B)ase, respectively, which is a natural choice for nucleic acids. Note that to exploit the idea of RNA conformation classes we chose to place beads on the existing atoms rather than on the center of masses of groups of atoms.

The sample configurations of the RNA nanoring in the one- and three-bead variants of our models (denoted as 1B and 3B in what follows) are depicted in figure 3. In the 1B case, all beads of the single type (with the mass $m^{(P)} = 321.5$ amu) are placed on the P atoms in the phosphates. In the 3B case, two types of beads are thus placed on the P atoms and C4′ carbons, while a number of plausible choices are possible for

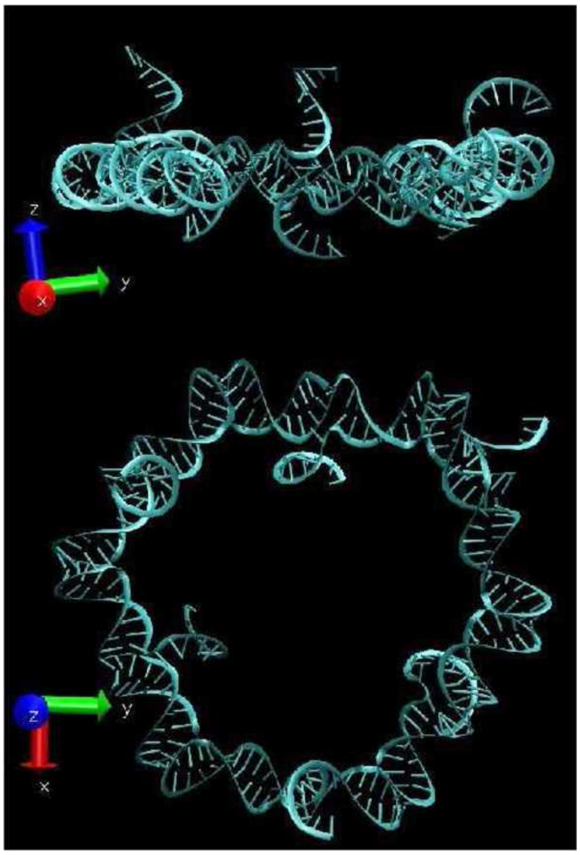

Figure 1. Side and top views of the RNA nanoring structure in the 'new cartoon' representation of VMD.

the placement of the third base bead. We found the following variant to be most convenient: N9 atom of purines and N1 atom of pyrimidines. The masses of beads in the 3B representation are taken as $m^{(\mathrm{P})} = 109$ amu, $m^{(\mathrm{S})} = 120$ amu, $m^{(\mathrm{B})} = 92.5$ amu.

In our model, the beads are organized into several single chains that correspond to the basic building blocks of the studied nanostructure, and the connectivity inside which is never broken in the course of simulation. For example, the nanoring from figure 1 is built up from six chains that form the sides of a hexagon, and are folded into double-helical stems with septaloops at both ends. Dangling 5′ and 3′ ends of the chains, found in the middle of the hexagon sides, are excluded from the CG model in order to focus on the core of the nanoring (264 nucleotides). The total interaction energy has the following form:

$$V = V_{\mathrm{conn}} + V_{bp} + V_{nb}, \tag{1}$$

with the standard chain connectivity contribution V_{conn}:

$$V_{\mathrm{conn}} = \sum_{\mathrm{chains}} \left(\sum_{\mathrm{bonds}} V_b(r - r^{(0)}) + \sum_{\mathrm{angles}} V_a(\theta - \theta^{(0)}) + \sum_{\mathrm{dihedrals}} V_d(\phi - \phi^{(0)}) \right), \tag{2}$$

where $V_b(r)$, $V_a(\theta)$, $V_d(\phi)$ are the intra-chain terms that correspond to the energies of bonds, angles and dihedrals

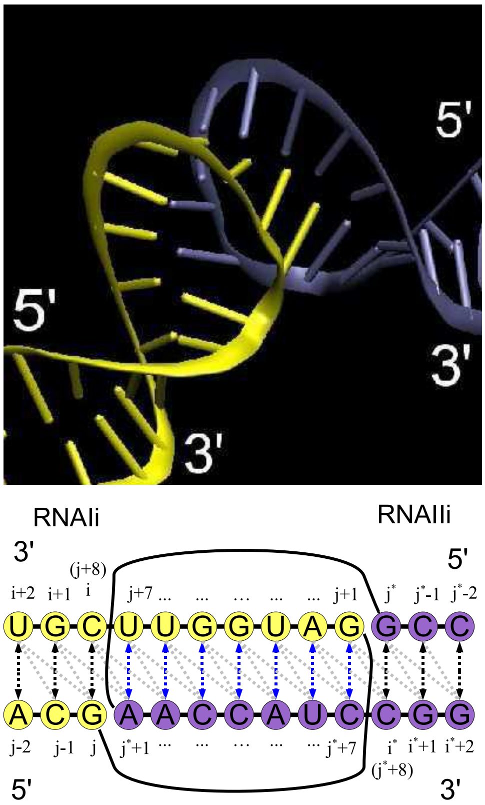

Figure 2. Top: the 'new cartoon' 3D representation of one kissing loop corner of the RNA nanoring illustrating the base pairing and stacking. Bottom: 2D secondary structure sketch of the same kissing loop (colors correspond to the 3D view above); the main base pair bonds are denoted with the arrows (black in the double-helical stems, blue in the kissing loop); two auxiliary bonds (equation (3)) per base pair are shown with gray dashed lines; the indexing scheme used in the simulation is shown; the base pairs are ordered in the sense of stacking that occurs continuously throughout the kissing loop; two long stretched lines connecting consecutive nucleotides correspond to sharp kinks in the nucleic acid backbones visible in the 3D view above.

(often abbreviated '$b/a/d$' in what follows), while $r^{(0)}$, $\theta^{(0)}$, $\phi^{(0)}$ are the equilibrium values for $b/a/d$, respectively.

The energy term V_{bp} accounts for the interactions between the base pairs. In the case of the nanoring, these include the contributions from the base pairs found inside the double-helical part of a single chain, as well as between those septuplets of the base pairs belonging to different chains that form the kissing loops. Following the idea in [15], we express the base pair interactions with three bonds per base pair instead of one:

$$V_{\mathrm{bp}} = \sum_{i,j\in(\mathrm{base\ pairs})} U_{i,j}\left(r_{i,j} - r_{i,j}^{(0)}\right) + U_{i+1,j}\left(r_{i+1,j} - r_{i+1,j}^{(0)}\right) + U_{i+2,j}\left(r_{i+2,j} - r_{i+2,j}^{(0)}\right), \tag{3}$$

which lets us enhance structural accuracy, since it takes into account both the hydrogen bonding between bases and the stacking interactions. As illustrated in figure 2 (bottom: arrows and gray dashed lines), a base j interacts not only with its counterpart i, but also with the neighboring bases $i + 1$, and $i + 2$ from the anti-sense part of the same double-helical stem (if present). Special care is required for laying

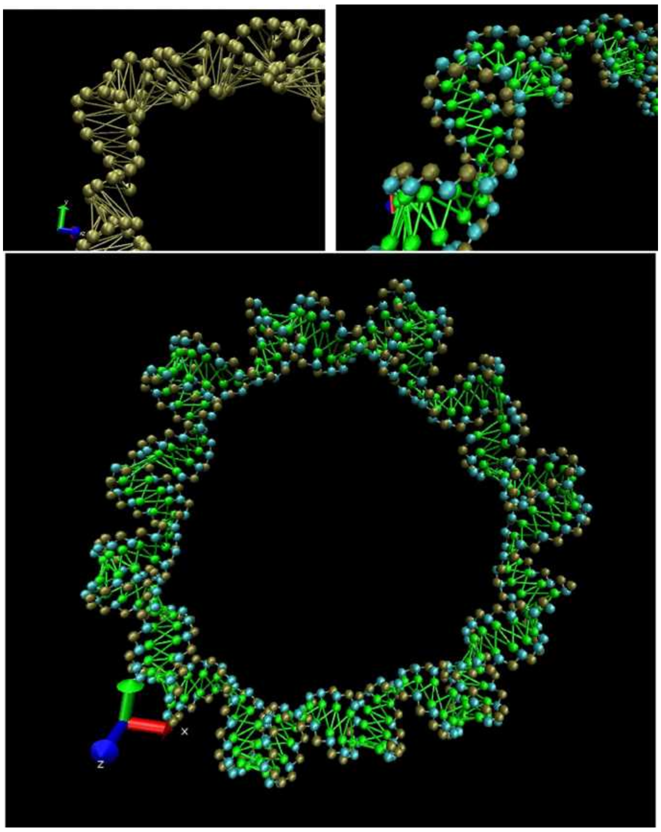

Figure 3. CG representations for the RNA nanoring shown in figure 1. From top to bottom: zoomed views of one 'kissing loop' in 1B and 3B representations, and the full RNA nanoring in the 3B representation. The phosphate (P) beads are shown in brown, the sugar (S) beads in cyan and the base beads (B) in green. The bonding of the backbone and between the bases is shown with the lines.

out the base pair interactions in the kissing loops. Since a base-paired kissing loop section closely resembles the double-helical structure, and the stacking occurs continuously from one stem helix through the loop-loop helix to the other stem helix [39], in the RNA nanoring all the bases have three interacting neighbors (possibly from two different chains). The connectivity between the base pairs is thus maintained throughout the time evolution.

The remaining energy contribution V_{nb} corresponds to the interactions between all bead pairs not involved in the bonded interactions described above. It has the following form:

$$V_{\mathrm{nb}} = \sum_{i,j\in(\mathrm{nonbonded})} v(r_{ij}). \tag{4}$$

In the present paper we take the simplest Weeks–Chandler–Andersen (WCA) form [40] for the nonbonded potential $v(r)$. It consists of the repulsive part of the Lennard–Jones potential, and expresses the steric repulsion between the beads via energy ε and bead diameter σ.

The MD for the CG model is implemented via the DL_POLY 2.19 package [41]. The resulting CG model is relaxed via an energy minimization (conjugate gradient method) and then equilibrated at a constant temperature in an NVT ensemble (Evans algorithm, [42]) with open boundary condition and a time step of 0.01 ps, sufficiently small to conserve the energy of the system in the constant energy runs. A cutoff of 10 Å is applied to nonbonded interactions. Visualization and data processing of both all-atom MD and CGMD simulations are carried with VMD [43], using in-house developed scripts.

3. Fitting of the CG parameters

The total energy of the CG model, equation (1), thus contains a number of parameters for the $b/a/d$, base pair terms (as well as those for the nonbonded interactions). Our general approach to the fitting of these parameters is the following: (i) the histograms of the values of various bonds, angles and dihedrals, as well as the RDF between all sorts of beads are extracted from all-atom MD trajectories; (ii) these distributions are used to fit the CG model parameters via the BI method [7]. Namely, given a probability distribution function $P(q)$ for a degree of freedom q, the corresponding potential of mean force (PMF) $V_{\mathrm{eff}}(q)$ is determined via the following formula:

$$V_{\mathrm{eff}}(q) = -k_B T \ln(P(q)). \tag{5}$$

Note that $V_{\mathrm{eff}}(q)$ thus obtained coincides with the true potential energy only for the case of a single degree of freedom q, and generally it can serve only as a first approximation used in a subsequent iterative procedure, which may not always be successful because many variables have to be fitted simultaneously. Fortunately, different energy contributions usually show a certain hierarchy, which allows their refinement in succession, in order of their decreasing strength, e.g. $V_{\mathrm{bond}} \to V_{\mathrm{angle}} \to V_{\mathrm{van\ der\ Waals}} \to V_{\mathrm{dihedral}}$ [7].

We fit the effective potentials for the bonded degrees of freedom $V_{\mathrm{eff}}(q)$, equation (5), by their Taylor expansions (up to quartic) around their global minima:

$$V_{\mathrm{eff}}(q) = \frac{k}{2}(q-q^{(0)})^2 + \frac{k'}{3}(q-q^{(0)})^3 + \frac{k''}{4}(q-q^{(0)})^4. \tag{6}$$

We thus obtain an equilibrium value $q^{(0)}$ of a bonded degree of freedom q and the coefficients k, k', k''. For the 1B CG model, we used all three terms, while for the 3B CG model the use of only the harmonic term k proved to be sufficient for our purposes (for the latter case, we did some tests with the anharmonic coefficients k' k'' included, and found no important differences; they may be useful in the future, for the fitting of the CG model to the dynamical properties, such as diffusivity). We find that for most of the bonded degrees of freedom (in the 3B case) the initial values of the parameters obtained directly from equations (5) and (6) already reproduce sufficiently well the histograms in the CGMD simulations, so that only seldom subsequent adjustments are required. They are done by manually introducing small changes to the coefficients in equation (6), in order to improve matching between the MD and CGMD distributions. For further refinement of the model, we plan to resort also to a more systematic fitting procedure, involving the iterative procedures and the force matching method [9] with cubic spline potentials [11] in particular.

We used two all-atom MD data sources: (i) a 6 ns 300 K trajectory of a simple RNA double A-helix dodecamer (GCGCUUAAGCGC); (ii) a 2 ns 310 K trajectory of a complex RNA nanostructure—nanoring. Both systems have been simulated in explicit water with Mg and Na counter ions. Further details about these MD runs can be found in the supporting information available from stacks.iop.org/PhysBio/7/036001/mmedia. While the data derived from the dodecamer served as a source for 'double-helical' parameters (due to a longer trajectory they are also more reliable), the set of data derived from the nanoring allowed us to introduce 'non-helicity' into the model in a controllable manner. Note that the all-atom MD runs used are rather short; nevertheless, the studied RNA structures did not show any tendency to change their global conformations [6]. This ensures that our CG model is valid at least in the vicinity of the 'native structure' minimum of the free energy (the transferability of a CG model to a different structure/conformation, and especially to the transition states between different conformations, is still a very challenging problem and needs to be tested).

It is important to stress that, for the RNA nanoring, the above-mentioned degrees of freedom are not distributed according to Boltzmann statistics only, but their distributions also reflect the spatial inhomogeneity of the system. Therefore, an attempt to represent such a degree of freedom via a single potential function using the BI method would lead to instability of the desired structure in the CG model, because such a degree of freedom would be discriminated against energetically in certain regions. Instead, one may introduce some local modifications to the potential functions. The ultimate strategy of this sort is the SOP approach, where each instance of such a degree of freedom has its own equilibrium value depending on its location in the molecule.

Thus, we consider three different parameter sets: (i) 'SOP' parameter set, where the coefficients k, k', k'' of the potential functions $V_{\mathrm{eff}}(q)$ are uniform throughout the system, while the equilibrium values $q^{(0)}$ are unique for each instance of $b/a/d$; (ii) 'SOP-dihedrals', where the SOP approach is applied only to the dihedrals of the nucleic acid backbone, but not to the bonds, base pairing bonds or angles, and (iii) the FF parameter set where each instance of a degree of freedom is described by all uniform parameters (including their equilibrium values) throughout the system. In all cases, for the uniform part, we used the CG parameters k, k', k'' and $q^{(0)}$ extracted from the dodecamer data. However, for the SOP and the SOP-dihedral parameter sets, we replaced the uniform equilibrium values with the full (inhomogeneous) sets found either in the initial non-equilibrated structure of the nanoring (we term such configuration 'ini1' in what follows), or by the averaging of each instance of an equilibrium value over the all-atom MD trajectory of the nanoring ('ini2').

4. Results

4.1. 1B representation of the coarse-grained model

The simplest 1B representation of the model includes the following CG degrees of freedom: (P_iP_{i+1}) backbone bonds, ($P_iP_{i+1}P_{i+2}$) backbone angles and ($P_iP_{i+1}P_{i+2}P_{i+3}$) backbone dihedrals, three base-pairing bonds (P_iP_j), ($P_{i+1}P_j$), ($P_{i+2}P_j$) per (i, j) base pair, as well as the distances for the nonbonded (PP) pairs. The indices i and j denote the nucleotide numbers in the sequence (counting from the 5′ end), and they are omitted in what follows wherever possible without the loss of clarity.

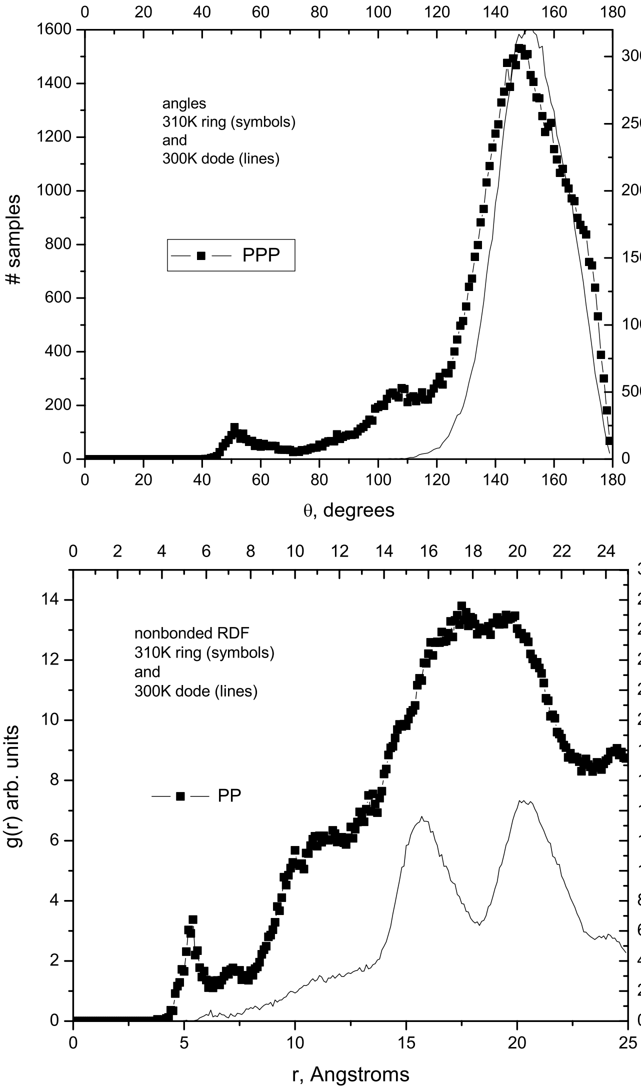

Figure 4. Angle histograms and nonbonded RDFs from all-atom MD runs for the 1B representation of the model. The data for the RNA nanoring are shown with symbols, while those for the RNA dodecamer are shown with thin lines.

The histograms of the backbone angles as well as RDFs for the nonbonded pairs from all-atom MD data are plotted in figure 4 for both the RNA dodecamer and the RNA nanoring in comparison (data for all the 1B CG degrees of freedom can be found in the supporting information, figure S1, available from stacks.iop.org/PhysBio/7/036001/mmedia). The distributions for the RNA nanoring are in general broader; they contain extended tails that reflect the existence of nonhelical regions. While the distribution of the backbone and base pair bond length is simpler and unimodal for both the dodecamer and the nanoring, the remaining distributions (angles, dihedrals and nonbonded pairs) for the nanoring show a number of features that are absent in the dodecamer, but which are crucial for the development of a CG model. Namely, the angular distributions for the nanoring show additional small peaks at ≈60° and ≈110° besides the main peak at 155°, the dihedrals show additional broad peak at ≈−160° besides the main peak at 14.4° and the nonbonded RDF shows a small peak at ≈5 Å due to the closely spaced phosphates. As a close examination of the atomic configurations reveals, these features are associated mostly with the regions of the kissing loops.

Since it does not make sense to fit multi-modal distributions of the bonded CG degrees of freedom with simple potentials of equation (6), in the 1B variant of the model we considered only SOP parameter sets with the coefficients k, k', k'' derived from the dodecamer data. Besides, as the peak of the nonbonded RDF at ≈5 Å dictates that the nonbonded interaction potential does not discriminate such small interbead spacings energetically, we have taken the values $\varepsilon = 0.1$ kcal mol^{-1}, $\sigma = 5.0$ Å for the nonbonded WCA parameters. Table S1 lists the working values of the parameters for the 1B CG model. We subjected the resulting 1B CG representation of the RNA nanoring to 500 ns equilibration at $T = 300$ K. Various distributions from these equilibration runs are plotted in figure S2 together with the analogous all-atom MD data for comparison. While a more detailed account can be found in the supporting information available from stacks.iop.org/PhysBio/7/036001/mmedia, here we emphasize the main result—the 1B CG model fails to reproduce the overall shape of the nanoring, which collapses to various unrealistic configurations, characterized by the abundance of too closely spaced nonbonded beads, figure S2. This is the consequence of the above-mentioned ≈5 Å restriction on the repulsive nonbonded potential (while the all-atom MD nonbonded RDFs suggest that the beads should be ≈10 Å in diameter).

Note that this does not signify an inherent limitation of the 1B CG models for RNA. Ultimately, the choice of the model should be determined by the physics of the problem at hands. The poor performance of the 1B model in our case is merely related to the insufficient space filling due to the combination of chosen bead placement (on phosphates, which is rather far from the double-helix axis, but which we preferred to keep in view of anticipated inclusion of explicit charges in the future versions of the model), bead 'size' and other details of the interactions. For example, in the studies of the forced RNA unfolding [19, 20] both 1B and 3B variants of the CG models used show comparable performance. In 1B case, the model of [20] has larger, ≈7 Å in diameter, beads, placed on the mass-centers of the nucleotides (which is closer to the double-helix axis), and SOP-style 'native' pair interactions. Another recent 1B CG model [23] uses smaller beads (≈5 Å in diameter), placed on C3 atoms in sugars, and this seems to provide adequate space filling for the studied RNA structures. For our phosphate-centered 1B model, it is possible to introduce a more complex nonbonded pair potential, with multiple minima, which would improve its structural precision. However, this opens a number of questions about the relative depths of the minima/heights of the barriers and the existence of unrealistic spurious configurations in the result. While a complex pair potential can be represented with splines and adjusted in a very detailed manner [11], this does not secure us from the latter caveat. Besides, for the FF parameter set in the 1B representation, one needs to introduce the angular and dihedral terms of the fairly complex shapes too. This combination of factors led us to abandon the 1B representation as unsuitable in favor of the 3B representation.

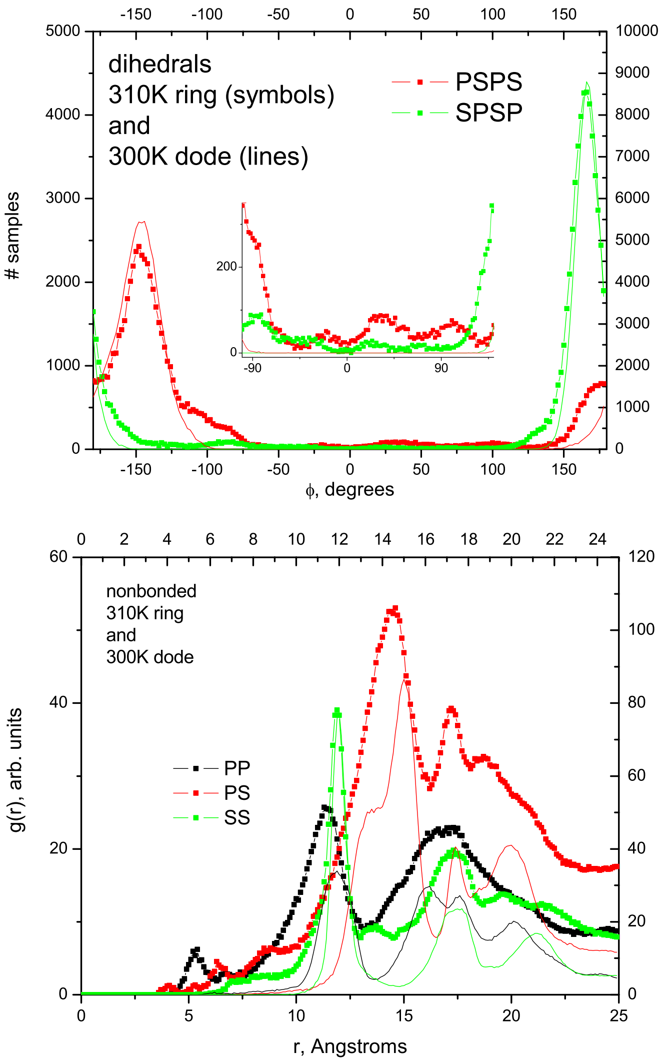

Figure 5. Top: histograms for dihedral angles from all-atom MD runs in the 3B representation of the model. The inset shows the zoomed portion of the figure near the baseline in the interval [−100°, 140°]. Bottom: RDFs for nonbonded interactions from all-atom MD runs for the 3B representation of the model. The data for the RNA nanoring are shown with symbols, while those for the RNA dodecamer are shown with thin lines (the color coding is the same in both cases).

4.2. *3B representation of the coarse-grained model*

The full list of different energy terms for the 3B representation of the CG model includes the following (the layout of the bonding terms is shown in figures 2 and 3). Along the nucleic acid backbone the (P_iS_i), (S_iP_{i+1}), (S_iB_i) bonds between the nearest neighbors, as well as ($P_iS_iP_{i+1}$),

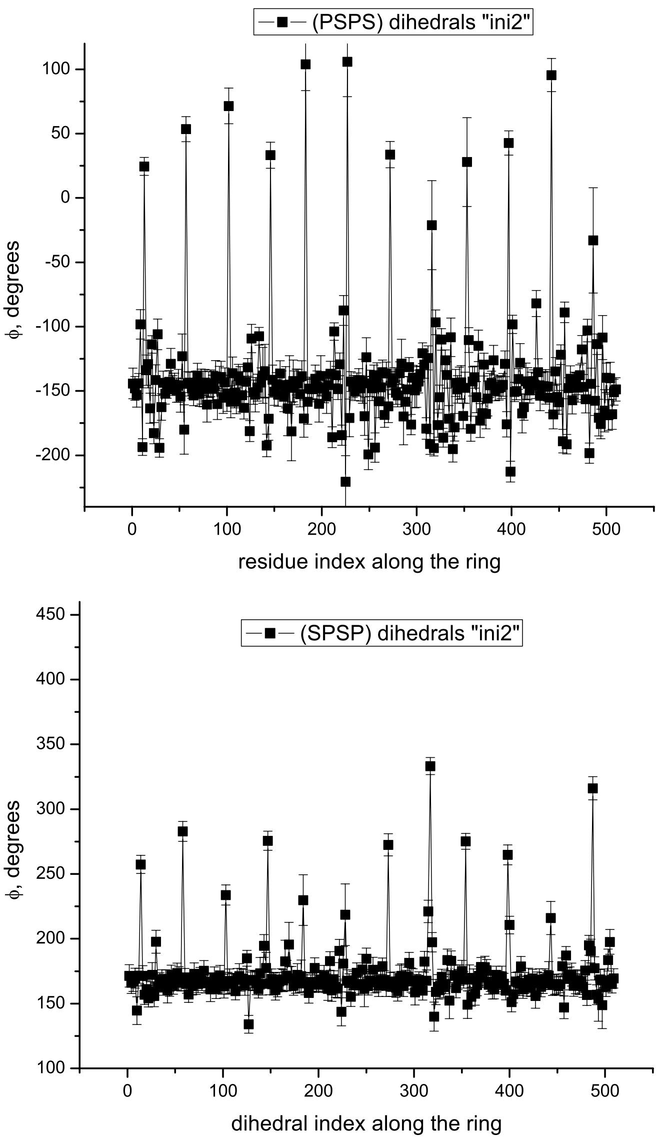

Figure 6. The 'SOP' fields of the dihedral angles introduced in the 3B model. Top: for initial configuration termed 'ini1'. Bottom: averaged over an all-atom MD trajectory (termed 'ini2'). To better represent the deviating dihedrals, the plotting intervals are chosen as follows: PSPS dihedrals are shown within [−240°, 120°]; SPSP dihedrals are shown within [0°, 360°].

Table 1. Parameters of the 3B CG model. The distances are expressed in Å; the angles and dihedrals are expressed in radians (the values in degrees are shown for convenience too). The unit of energy is kcal mol^{-1}, and the units for the coefficients are derived from the unit of energy and Å or radians, respectively. The nucleotide index i is counted from the 5′ end.

3B CG model parameters		
	Backbone bonds	
	r_0	k
(P_iS_i)	3.93	133.4
(S_iP_{i+1})	3.92	107.8
(S_iB_i)	3.38	61.3
	Base-pairing bonds	
	r_0	k
(B_iB_j)	8.99	16.5
$(B_{i+1}B_j)$	6.86	2.83
$(B_{i+2}B_j)$	7.35	2.93
	Backbone angles	
	θ_0	k
$(P_iS_iP_{i+1})$	1.733 (99.3°)	26.3
$(S_iP_{i+1}S_{i+1})$	1.819 (104.2°)	106.9
$(P_iS_iB_i)$	1.728 (99.0°)	32.1
$(B_iS_iP_{i+1})$	1.642 (94.1°)	127.2
	Backbone dihedrals	
	ϕ_0	k
$(P_iS_iP_{i+1}S_{i+1})$	−2.677 (−153.4°)	5.86
$(S_iP_{i+1}S_{i+1}P_{i+2})$	2.948 (168.9°)	16.2
	WCA potential for nonbonded pairs	
	σ	ε
all	5.0	0.1
	Excluded bonds (see the text)	
(S_iB_{i+1})	(B_iS_{i+1})	(B_iB_{i+1})

$(S_iP_{i+1}S_{i+1})$, $(P_iS_iB_i)$, $(B_iS_iP_{i+1})$ angles, $(P_iS_iP_{i+1}S_{i+1})$ and $(S_iP_{i+1}S_{i+1}P_{i+2})$ dihedrals are included. Along the base-paired parts of double helices and kissing loops the (B_iB_j), $(B_{i+1}B_j)$ and $(B_{i+2}B_j)$ bonds are included. Besides, due to the topology of the three-bead nucleic backbone, we introduce dummy 'zero energy' bonds between the nearest (S_iB_{i+1}), (B_iS_{i+1}) and (B_iB_{i+1}) neighbors along the backbone in order to exclude terms from the nonbonded interactions, since these bonds are already restrained by the above-mentioned set of backbone terms.

Our 3B model also includes six different non-bonded bead pairings (PP), (PS), (PB), (SS), (SB) and (BB). The RDFs from the all-atom MD runs for three selected pair types are shown in figure 5 for both studied systems (the full set of data for all nonbonded as well as bonded degrees of freedom can be found in figures S3 and S4). The nonbonded RDFs for the RNA nanoring (figure 5) show well-pronounced peaks/tails in the interval between 5 Å and 10 Å, which are absent in the case of the dodecamer. As in the 1B case, these features, caused by closely spaced beads in the kissing loops, dictate the choice of the same nonbonded parameters, $\varepsilon = 0.1$ kcal mol^{-1}, $\sigma = 5.0$ Å.

The 3B bonded distributions for the RNA nanoring show much less pronounced fine features, compared to the analogous plots for the 1B case. In fact, all the bonded terms (except the dihedrals) have unimodal distributions closely resembling Gaussians, which allows us to retain the harmonic coefficients k only in equation (6). More complex distributions are demonstrated by the dihedral angles. They are plotted in figure 5 for both the RNA nanoring and the RNA dodecamer for comparison. For example, the (PSPS) dihedrals contain two shoulders near the main peak at $\approx -153.4°$ consistent with two RNA conformational classes [38], as further explained in section 5.

Besides, as more careful examination of the dihedral histograms for both (PSPS) and (SPSP) reveals, there exist some dihedral values (mainly in the kissing loops) that deviate strongly from the centers of the distributions (inset in figure 5). Their fraction is not high; however, their presence

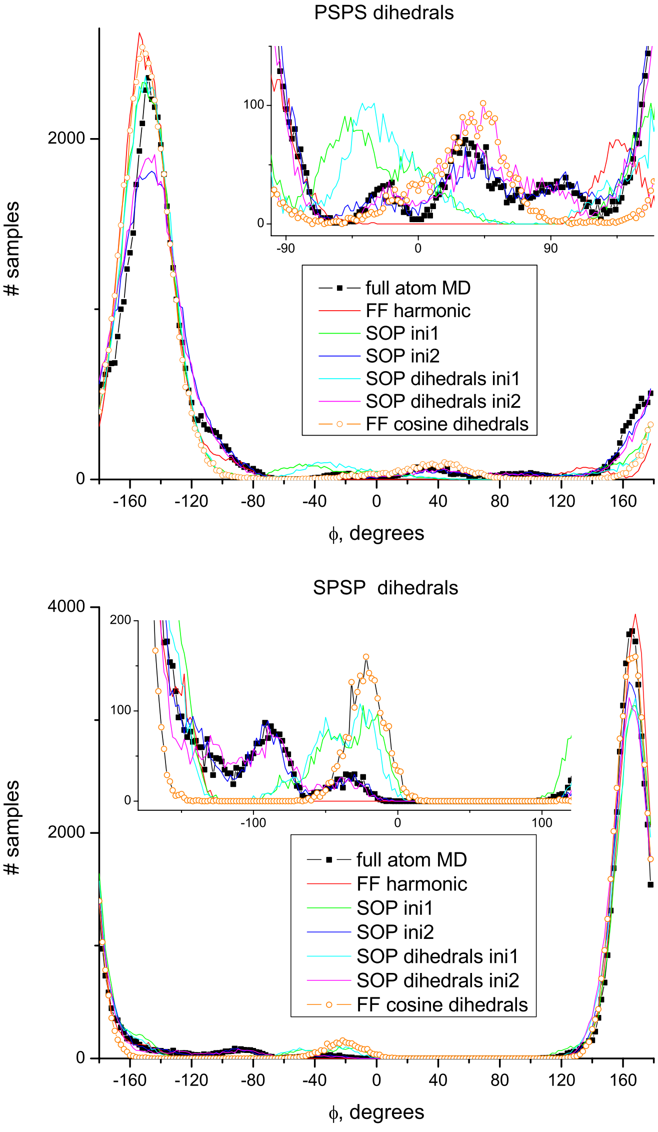

Figure 7. Dihedrals histograms for the RNA nanoring from full atom MD and from CGMD runs for comparison. The insets show the zoomed portions near the baselines in the interval [−100°, 160°] (PSPS) and [−180°, 120°] (SPSP), respectively. The small peaks near the baselines are reproduced equally well by SOP and SOP-dihedral variants of the CG model, and they are also reasonably reproduced by the FF-cosine-dihedral variant, as further discussed in the text.

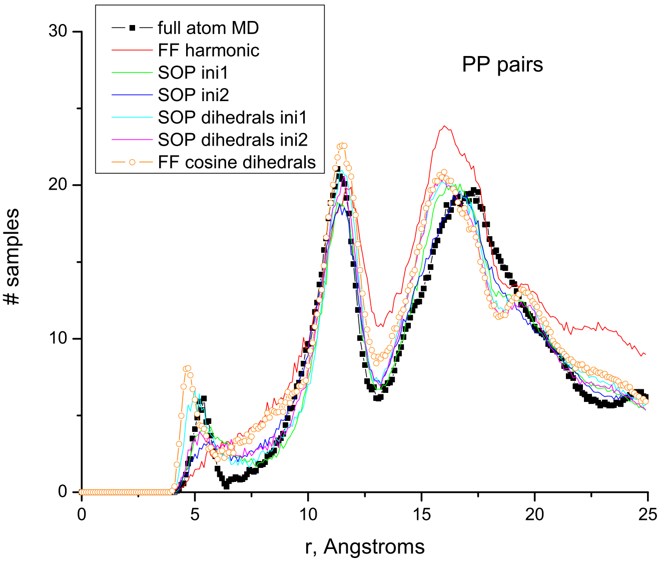

Figure 8. RDFs for nonbonded PP pairs for the RNA nanoring from all-atom MD and from CGMD runs for comparison.

has to be taken into account in a CG model. The detailed dependences of the (PSPS) and (SPSP) dihedrals along the ring versus the dihedral index for the ini2 SOP parameter set are shown in figure 6 (figure S5 shows similar plots for ini1). In both cases, the dihedrals deviating by ∼180° (we term them *cis*) from the distribution centers are clearly visible (the latter correspond approximately to the *trans* orientation). They belong to the 12 localized parts of the nucleic acid backbones (the sharp backbone kinks in figure 2) participating in six kissing loops, i.e. in total there are four such outstanding *cis* dihedrals per kissing loop pair. In the FF variant of the model, where all the dihedrals of the same type should have the same uniform equilibrium value, such dihedrals would be strongly discriminated against energetically by a harmonic or a quartic effective potential of equation (6), and this would strongly distort the equilibrium structure of the kissing loops. As a remedy, we tested the following dihedral function:

$$V_{\text{eff}}(\phi) = \frac{k}{4}[1-\cos(2(\phi-\phi_0))] \qquad (7)$$

that has two minima at $\phi = \phi_0$ and $\phi = \phi_0 + 180^\circ$, both with stiffness k. It turns out that by accommodating the *cis* dihedrals, this simple function provides an excellent performance for the FF variant of the model in describing the RNA nanoring.

Since we intended to describe the RNA nanoring with a simple CG model based on the main body of purely helical parameters, with a controlled amount of non-helicity introduced either via SOP or via a FF with a cosine dihedral term, we have chosen the set of bonded parameters fitted via the BI method from the RNA dodecamer data, and tested it on the RNA nanoring. Table 1 lists the values of parameters for the 3B CG model we thus obtained. To compare the performance of all the considered six variants of the 3B CG model, namely the SOP, SOP-dihedrals (both with ini1 and ini2 detailed sets) as well as two FF variants with the harmonic (6) (termed 'FF-harmonic') and cosine (7) ('FF-cosine-dihedrals') dihedral functions, we performed a series of 750 ns long CGMD equilibration runs at constant temperature of 300 K. The dihedral histograms and RDFs for nonbonded (PP) pairs obtained by the end of these runs are shown in figures 7 and 8, respectively, in comparison to the distributions from all-atom MD (full set of data can be found in the figures S6–S10). After a few minor manual adjustments of the CG parameters, the histograms of the bonded terms show reasonable agreement with those from the all-atom MD simulations (apart from some non-essential discrepancies further discussed in the remarks in supporting information available from stacks.iop.org/PhysBio/7/036001/mmedia). The dihedral distributions show the required extended tails in all cases

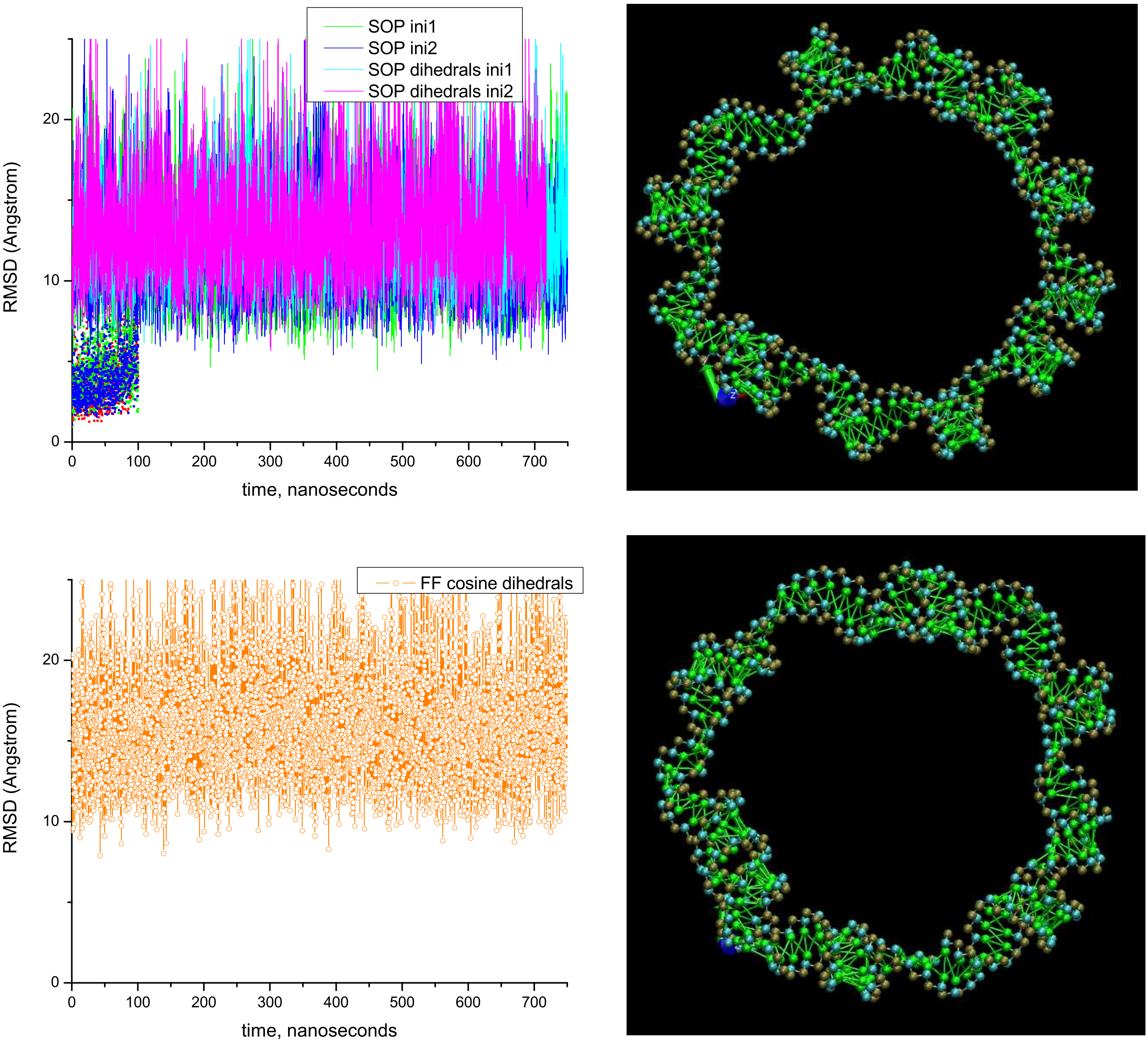

Figure 9. Left: RMSD for different variants of the parameter sets of the CG model in the equilibration CG runs. The data for the nanoring are plotted with lines and open circles ('FF-cosine-dihedrals'), while those for the dodecamer are plotted with small symbols (in the interval up to 100 ns only). Right: the final snapshots of the RNA nanoring after 750 ns equilibration in the CG model with the 'SOP dihedral' (top) and 'FF-cosine-dihedral' (bottom) parameter set.

except, obviously, FF-harmonic (figure 7, the insets). What is even more remarkable, all the variants of the CG model (except the FF-harmonic) are able to capture the fine features of the nonbonded RDFs. The most important of them is the 5 Å peak for the PP pairs (figure 8), which is even slightly over-emphasized by the FF-cosine-dihedral variant.

The root mean square deviations (RMSDs) from the initial structures during these runs are plotted in figure 9. Typical values of RMSD are 13.0 ± 3.0 Å for SOP ini1 and ini2, 13.5 ± 3.5 Å for SOP-dihedrals ini1 and ini2 and the RMSD is about the same (16.1 ± 3.2 Å) for the FF-cosine-dihedral variant. These values are to be compared to the typical values for the dodecamer (also plotted in figure 9 4.5 ± 1.5 Å for all considered variants) and to 35 ± 5.0 Å for the FF-harmonic variant for the nanoring (not plotted) in which case the overall shape of the nanoring is not preserved. The final snapshots of the nanoring for the SOP-dihedrals and FF-cosine-dihedral variants are depicted in the same figure, attesting to the preservation of the helical segments, kissing loop structures and the overall shape.

We conclude, therefore, that the 3B CG model provides an excellent description for the structurally inhomogeneous RNA aggregate—the nanoring. This shows the power of the 3B representation, which, unlike the 1B one, captures adequately the excluded volume effects with small (≈5 Å in size) beads while allowing for closely spaced beads in the kissing loops.

5. Discussion: coarse-grained model and RNA conformation classes

Two findings from the previous section are the most important. First, we observe that the SOP-dihedral variants of the CG model provide about the same performance as the full SOP variants. This means that structural complexity of the RNA nanoring can be handled by including into a 3B CG model the detailed information about P-C4′ dihedral pseudo-angles only. This finding supports, for the case of the RNA nanoring, a more general statement that such a reduced representation of the RNA backbone gives a robust and complete description of the RNA structure [38], similar to the famous $\phi - \psi$ Ramachandran plots for proteins (in the case of RNA the sugar pucker should be specified too [38], it is always C3′-*endo* in our case). Moreover, based on a large body of experimentally available RNA structures, it is shown in [38] that the values of P-C4′ dihedral pairs cluster in a few localized regions only in the 2D pseudo-torsional space, i.e. there exist quasi-discrete *RNA conformation classes*.

Second, in the light of the RNA conformation classes, another important finding is that a simple modification of the dihedral function (7) allowed us to reach the SOP precision in describing the RNA nanoring by properly accommodating the distortions of the dihedrals in the kissing loops. Presently, the function (7) has only one additional dihedral minimum separated from the original one by 180°, i.e. only one additional conformation class (together with the main one for the A-form double helix). In principle, the dihedral function can be made more complex (e.g. to contain multiple minima). Note that the number of observed conformation classes is limited by at most 10 [38]. Therefore, one can hope to keep the dihedral functions (for both PSPS and SPSP dihedrals) reasonably simple yet fairly universal, which opens up a promising avenue toward a FF-like universal RNA CG model.

In this context it is interesting to establish a more detailed connection between the RNA conformation classes from [38] and the ones that we observe in our simulations. According to [38], a conformation class is defined by a pair of ($P_iS_iP_{i+1}S_{i+1}$) and ($S_{i-1}P_iS_iP_{i+1}$) dihedrals centered around a sugar S_i. Note that alternatively one can also select dihedral pairs centered around a phosphate P_i, i.e. ($S_{i-1}P_iS_iP_{i+1}$) and ($P_{i-1}S_{i-1}P_iS_i$). Using the RNA nanoring dihedrals from figure 6, we plotted 2D maps for both variants (figure 10). The main double-helical peak at (−153°, 169°) is clearly visible, along with the two shoulders extending in the PSPS direction (cf figure 5). In the terminology of [38] the shoulders correspond to the classes VI (cross-stem stacking of the purine-purine base pairs) and IV (absent stacking on the 3′ side of the nucleotide). The former is present in the nanoring because of a cross-stem stacking G-G pair just near the base of each loop [39], the latter is probably found in the middle of the nanoring sides (we did not analyze this in detail).

Two more classes are found in the kissing loops (figure 10, top). Namely, there are (i) 12 nucleotides (one per each kissing loop side) that have strongly deviating *cis* PSPS angles and double-helical *trans* SPSP angles, and (ii) immediately following them in the 5′ to 3′ sense, 12 other nucleotides that have the situation reversed, i.e. *cis* SPSP angles and *trans* PSPS angles. These nucleotides (found near sharp kinks of the nucleic backbone visible in figure 2) are the last ones base paired within their own chain and the first ones participating in the cross-chain base pairing in the kissing loops, respectively. While we can relate our class (ii) to the class II from [38], our class (i) seems to be absent from the scheme of [38]. Interestingly, two classes (i) and (ii) collapse into a single one if the 2D dihedral map is replotted with the dihedral pairs centered around the phosphates (figure 10, bottom). The reason for this is clear for the case of the nanoring, where the nucleotides of the classes (i) and (ii) are neighbors in the chain, but one may wonder whether a similar procedure applied to all the volume of data analyzed in [38] would lead to a simplification of the observed conformation classes. Note that the classification based on the phosphates (‘suites’), and not on the nucleotides has already been used for the full RNA conformational space [44].

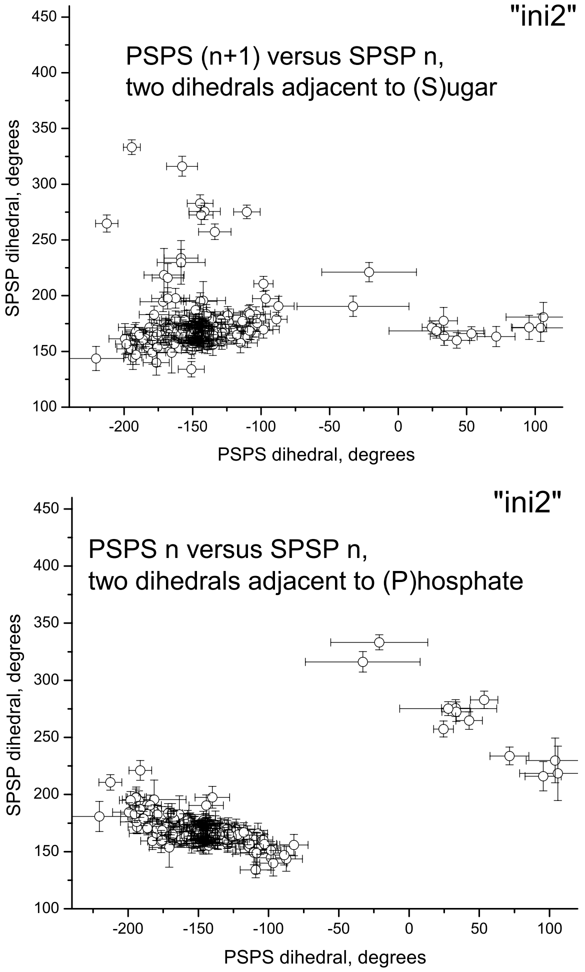

Figure 10. The 2D maps of the dihedral pair distributions of the RNA nanoring. Top: for the dihedral pairs centered around the sugars. Bottom: for the dihedral pairs centered around the phosphates.

Thus, we incorporated only the classes (i) and (ii) (collapsing to a single class if reformulated as specified above) into the FF-cosine-dihedral variant of our CG model, and ignored the remaining two classes present in the nanoring (IV and VI according to [38]). All these classes are obviously accounted for in the SOP-dihedral variant. This explains slightly elevated values of RMSD in the former case, compared to the latter, though one extra class alone proved to be sufficient to model the kissing loops reasonably well.

6. Conclusions

In the present paper we reported a series of bead-based CG models for RNA with varying amounts of atomistic structural information and numbers of beads per nucleotide. They range from SOP ones, where all the specific values of bonds/angles/dihedrals from a reference structure (nanoring) are included in the model, to a FF approach, where the CG model is described by a few universal parameters. We started from the purely double-helical set of parameters derived from all-atom MD data for an RNA dodecamer, and applied it to the RNA nanoring, introducing the non-helical features whenever necessary. We are concluding here that the models with one (phosphate-centered) bead per nucleotide suffer from the effects of insufficient excluded volume, while the models with three beads per nucleotide ((P)hosphate, (S)ugar, (B)ase) are more suitable for the description of the RNA. The inclusion of just the detailed information about the (PSPS) and (SPSP) dihedral angles in the model renders precision similar to the inclusion of all available atomistic structural information, which illustrates the usefulness and robustness of the reduced (P-C4′) representation of the nucleic backbone [38]. Furthermore, the existence of the quasi-discrete RNA conformation classes based on these dihedrals [38] is supported by our data too. For the simple non-helical conformation of the kissing loops we were able to design a dihedral potential function (7), that, while including only a single additional minimum, successfully accommodated local dihedral distortions. This opens up the road toward the development of even more transferable RNA CG models based on the P-C4′ conformation classes. Table 1 lists the values of the parameters (26 in total) for the 3B CG model we obtained. For the systems of thousands of nucleotides, a time scale of microseconds can be easily reached with the developed CG model. The structural precision of our 3B models in terms of RMSD is ∼0.06 Å per nucleotide (to be compared e.g. with ∼0.1 Å for another recent 3B model [25] and ∼0.13 Å for a 1B model [23]).

Finally, we mention several directions for further development of our RNA CG model that are underway now. In the current versions of the model no distinction is made between different bases. However, it is easy to introduce sequence specificity that would enhance structural precision of the model. Besides, if the interaction between base pairs were treated in a non-bonding manner, this would allow one to study the association/dissociation reactions between the RNA nanostructure building blocks. Our present simple scheme (three central-force bonds per base pair), that is well adapted to the geometry of the double helix and treats both base-pairing and stacking interactions, is difficult to transfer to the dissociative scheme. Instead, a more complex layout of the interactions that includes many-body and angular effects explicitly (e.g. in a manner similar to the one used in [25] for simplified discrete potentials) will be incorporated.

Given the importance of electrostatics and counter ions in the RNA folding and structure, it is worthwhile to also treat this physically distinct part explicitly. This would allow one to systematically represent the effects of the ionic strength of the solution as well as local modifications of the environment introduced by ionic cloud. Typical counter ion concentrations provide screening to such an extent that the electrostatics is negligible across the base pairs, while it is still important between neighboring phosphates [19]. Besides, it is well known that counter ions tend to concentrate strongly at certain places (e.g. in the kissing loops for the case of the nanoring). Therefore we plan to include a (screened) repulsion between phosphate (P) beads as well as course-grained representation of the counter ions together with their solvation spheres.

Acknowledgments

MP and RM are grateful to the NSERC CRC Program for its support. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada. The authors are grateful to Valentina Tozzini for helpful comments and suggestions.

Supplementary information

The following supplementary data are available online at stacks.iop.org/PhysBio/7/036001/mmedia: (i) details of the all-atom MD simulations used as sources for CG model fitting; (ii) details of the 1B CG model simulations; (iii) full set of graphs for all CG degrees of freedom.

References

[1] Jaeger L and Chworos A 2006 *Curr. Opin. Struct. Biol.* **16** 531
[2] Jaeger L, Westhof E and Leontis N 2001 *Nucleic Acids Res.* **29** 455
[3] Holbrook S R 2005 *Curr. Opin. Struct. Biol.* **15** 302
[4] Yingling Y G and Shapiro B A 2007 *Nano Lett.* **7** 2328
[5] Sponer J and Lankas F (ed) 2006 *Computational Studies of RNA and DNA (Challenges and Advances in Computational Chemistry and Physics* vol 2*)* (Berlin: Springer)
[6] Paliy M, Melnik R and Shapiro B A 2009 *Phys. Biol.* **6** 046003
[7] Reith D, Pütz M and Müller-Plathe F 2003 *J. Comput. Chem.* **24** 1624
[8] Lyubartsev A P and Laaksonen A 1995 *Phys. Rev.* E **52** 3730
[9] Ercolessi F and Adams J B 1994 *Europhys. Lett.* **26** 583
[10] Voth G A (ed) 2008 *Coarse-Graining of Condensed Phase and Biomolecular Systems* (Boca Raton, FL: CRC Press)
[11] Izvekov S and Voth G A 2005 *J. Phys. Chem.* B **109** 2469
[12] Noid W G, Chu J-W, Ayton G S, Krishna V, Izvekov S, Voth G A, Das A and Andersen H C 2008a *J. Chem. Phys.* **128** 244114
[13] Noid W G, Liu P, Wang Y, Chu J-W, Ayton G S, Izvekov S, Andersen H C and Voth G A 2008b *J. Chem. Phys.* **128** 244115

[14] Tozzini V 2005 *Curr. Opin. Struct. Biol.* **15** 144

[15] Trovato F and Tozzini V 2008 *J. Phys. Chem.* B **112** 13197

[16] Ouldridge T E, Johnston I G, Louis A A and Doye J P K 2009 *J. Chem. Phys.* **130** 065101

[17] Ghosh J and Faller R 2007 *Mol. Simul.* **33** 759

[18] Nielsen S O, Lopez C F, Srinivas G and Klein M L 2004 *J. Phys.: Condens. Matter.* **16** R481

[19] Hyeon C and Thirumalai D 2005 *Proc. Natl Acad. Sci. USA* **102** 6789

[20] Hyeon C and Thirumalai D 2007 *Biophys. J.* **92** 731

[21] Hyeon C, Dima R and Thirumalai D 2006 *Structure* **14** 1644

[22] Trylska J, Tozzini V and McCammon J A 2005 *Biophys. J.* **89** 1455

[23] Jonikas M A, Radmer R J, Laederach A, Das R, Pearlman S, Herschlag D and Altman R B 2009 *RNA* **15** 189

[24] Pincus D L, Cho S S, Hyeon C and Thirumalai D 2008 *Molecular Biology of Protein Folding, Part B* vol 84 ed P M Conn (New York: Academic) pp 203–50

[25] Ding F, Sharma S, Chalasani P, Demidov V V, Broude N E and Dokholyan N V 2008 *RNA* **14** 1164

[26] Gherghe C M, Leonard C W, Ding F, Dokholyan N V and Weeks K M 2009 *J. Am. Chem. Soc.* **131** 2541

[27] Cao S and Chen S-J 2005 *RNA* **11** 1884

[28] Cao S and Chen S-J 2009 *RNA* **15** 696

[29] Das R and Baker D 2007 *Proc. Natl Acad. Sci.* **104** 14664

[30] Parisien M and Major F 2008 *Nature* **452** 51

[31] Tan R K Z, Petrov A S and Harvey S C 2006 *J. Chem. Theory Comput.* **2** 529

[32] Whitford P C, Schug A, Saunders J, Hennelly S P, Onuchic J N and Sanbonmatsu K Y 2009 *Biophys. J.* **96** L7

[33] Whitford P C, Geggier P, Altman R B, Blanchard S C, Onuchic J N and Sanbonmatsu K Y 2010 *RNA* **16** 1196

[34] Gopal S M, Mukherjee S, Cheng Y-M and Feig M 2010 *Proteins: Struct. Funct. Bioinformatics* **78** 2187

[35] Shapiro B A, Yingling Y G, Kasprzak W and Bindewald E 2007 *Curr. Opin. Struct. Biol.* **17** 157

[36] Knotts IV T A, Rathore N, Schwartz D C and de Pablo J J 2007 *J. Chem. Phys.* **126** 084901

[37] Sambriski E, Schwartz D and de Pablo J 2009 *Biophys. J.* **96** 1675

[38] Wadley L M, Keating K S, Duarte C M and Pyle A M 2007 *J. Mol. Biol.* **372** 942

[39] Lee A J and Crothers D M 1998 *Structure* **6** 993

[40] Weeks J D, Chandler D and Andersen H C 1971 *J. Chem. Phys.* **54** 5237

[41] Smith W, Yong C W and Rodger P M 2002 *Mol. Simul.* **28** 385

[42] Evans D J and Morris G P 1984 *Comput. Phys. Rep.* **1** 297

[43] Humphrey W, Dalke A and Schulten K 1996 *J. Mol. Graphics* **14** 33

[44] Murray L J W, Arendall W B, Richardson D C and Richardson J S 2003 *Proc. Natl Acad. Sci. USA* **100** 13904