Published online 19 December 2015

Nucleic Acids Research, 2016, Vol. 44, No. 7 ***e63***
doi: 10.1093/nar/gkv1479

SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction

Michal J. Boniecki*, Grzegorz Lach†, Wayne K. Dawson†, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother and Janusz M. Bujnicki*

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

Received October 03, 2014; Revised November 29, 2015; Accepted December 05, 2015

ABSTRACT

RNA molecules play fundamental roles in cellular processes. Their function and interactions with other biomolecules are dependent on the ability to form complex three-dimensional (3D) structures. However, experimental determination of RNA 3D structures is laborious and challenging, and therefore, the majority of known RNAs remain structurally uncharacterized. Here, we present SimRNA: a new method for computational RNA 3D structure prediction, which uses a coarse-grained representation, relies on the Monte Carlo method for sampling the conformational space, and employs a statistical potential to approximate the energy and identify conformations that correspond to biologically relevant structures. SimRNA can fold RNA molecules using only sequence information, and, on established test sequences, it recapitulates secondary structure with high accuracy, including correct prediction of pseudoknots. For modeling of complex 3D structures, it can use additional restraints, derived from experimental or computational analyses, including information about secondary structure and/or long-range contacts. SimRNA also can be used to analyze conformational landscapes and identify potential alternative structures.

INTRODUCTION

Ribonucleic acid (RNA) molecules play crucial roles in living organisms; among many functions, they are carriers of genetic information, regulators of gene expression and catalysts of metabolic reactions (1). While the role of protein-coding RNA in transmission of genetic information encoded in triplets of residues depends essentially just on the ribonucleotide sequence, most of the other roles depend also on the structure of the ribonucleotide chain. Similar to proteins, in which the amino acid sequence determines the structure, the ribonucleotide sequence of RNA directly determines the pattern of base pairs (secondary structure) and the global shape (tertiary structure) that is assumed in a given environment. Many RNA molecules form unique stable tertiary structures, while others form alternative structures or undergo transformations between the structured and unstructured state. For example, riboswitches, regulatory elements located within mRNA that switch protein production on and off, function owing to the ability to undergo conformational changes depending on the binding of specific ligands or on sensing other environmental changes (2). Thus, the understanding of manifold mechanisms of RNA function beyond protein coding requires a detailed knowledge of RNA tertiary structure (3).

Advances in high throughput nucleic acid sequencing resulted in a rapid growth of RNA sequence information. Unfortunately, this growth of sequence information has not been paralleled by structure determination, and for the large majority of known RNA sequences, the three-dimensional (3D) structures remain unknown. The experimental determination of RNA structures is difficult and expensive; currently it is significantly more challenging than protein structure determination (4). This situation resembles a similar problem concerning protein sequences and structures, and both these problems have been approached by the development of computational methods for predicting 3D structures from the sequence information (5).

Previously, we have developed ModeRNA, a method for RNA 3D structure prediction that builds models using information from structures of homologous molecules used as templates (6,7). The major limitation of that method is that it can accurately predict RNA structures only if a similar structure is provided as a template, along with a sequence alignment between the target and the template molecules. However, as mentioned earlier, experimentally determined RNA 3D structures are sparse; hence, homology modeling is currently possible for only a small frac-

*To whom correspondence should be addressed. Tel: +48 22 597 07 50; Fax: +48 22 597 07 15; Email: iamb@genesilico.pl
Correspondence may be also addressed to Michal J. Boniecki. Tel: +48 22 597 07 53; Fax: +48 22 597 07 15; Email: mboni@genesilico.pl
†These authors contributed equally to the paper.

tion of RNA sequences. In addition, homology modeling does not provide information about the RNA folding pathways. For this, one needs to turn to a modeling approach that samples different conformations of the RNA chain and models not only the final structure, but also the folding process. Thus far, various methods for RNA folding simulations have been developed, and they have used a variety of RNA structure representations, conformational sampling schemes and energy/scoring functions (8–12). They have various strengths and limitations, as observed in the recently initiated RNA Puzzles experiment (13). To this end, inspired by the success of coarse-grained methods for protein structure prediction such as REFINER (14) or CABS (15), and based on our experience with protein modeling, we have developed a coarse-grained method for RNA folding simulations and 3D structure prediction dubbed SimRNA. We aimed to develop a method that allows for RNA 3D structure prediction from sequence alone, and that can use additional structural information, if available. Here, we present SimRNA, together with the results of its tests and comparison with other methods for template-free RNA modeling, and we discuss its possible applications.

The history of coarse-grained modeling of RNA is long and multifaceted (16–20), ranging from simple models using one bead per nucleotide with varying levels of sophistication (21,22), two and three bead models (10,23) and sometimes additional beads (24–27). SimRNA uses a statistical potential in the form of a grid and models the essential orientations of the bases along the backbone using five key atomic positions in each nucleotide: two beads (P and C4') define the backbone according to Olson's model (28) and three beads *define the plane* of the nucleotide base. The core atomic coordinates permit a nearly complete one-to-one transformation of trajectories of the base and backbone positions both *from* PDB structures and *to* PDB structures. Moreover, additional aspects of structure can be incorporated into the model in a modular fashion.

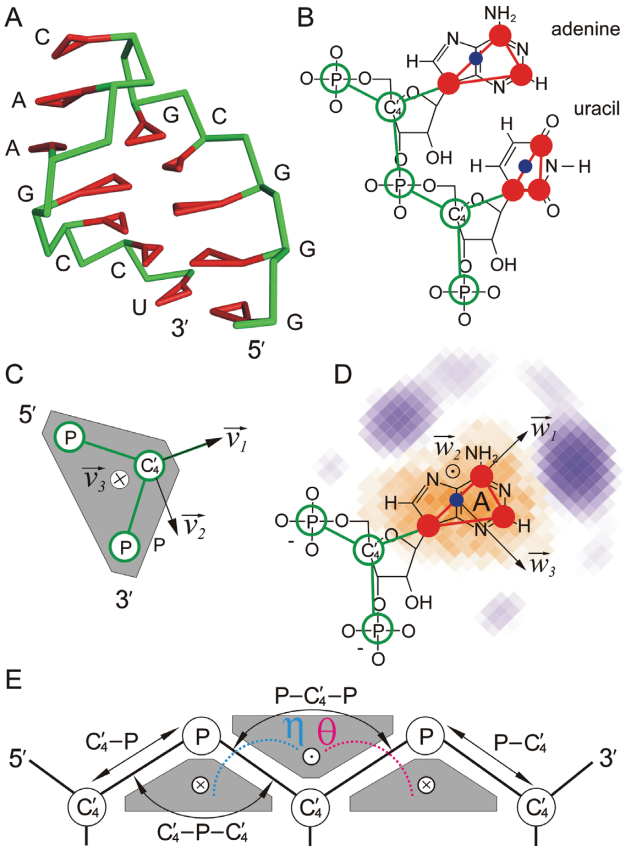

Figure 1. Reduced representation of RNA structure in SimRNA including the relationships between various base and backbone terms. (**A**) An example of an RNA structure (GCAA tetraloop, PDB id: 1zih) shown in reduced representation where green represents the backbone and red represents the base moieties. (**B**) Examples of reduced representation for the adenosine and uridine residues, with base level 1 and level 2 representation shown as red and blue points, respectively. (**C**) The backbone section including the vectors that orient the base relative to the backbone. (**D**) Level 3, the central layer (slice) of the 3D grid for the reference base, where the orange region represents the excluded volume of atoms of the base (repulsive region) and the purple region is an example of the attractive interactions between A and U in the central layer, including base-pairing around the Watson–Crick edge (the largest purple cloud), around the Hoogsteen edge (the second largest purple cloud) and the sugar edge (small purple cloud at the bottom of the diagram). It is worth noting that even though the red triangle covers only part of the base, the 3D grid approximates the volume of all atoms of the base. (**E**) Representation of the bond lengths, flat angles and pseudotorsion angles η and θ.

MATERIALS AND METHODS

Overview of the SimRNA method

SimRNA is a computational method for RNA folding simulations and 3D structure prediction. As virtually every method for simulations of molecular systems, it comprises three main functional elements: a representation of the molecules that are simulated, a scoring function (energy) and an algorithm that controls the moves of the molecular system. SimRNA utilizes a simplified (coarse-grained) representation of a nucleotide chain, a knowledge-based energy function and a Monte Carlo scheme for sampling the conformational space (29,30).

Representation of RNA molecules in SimRNA

In SimRNA, RNA molecules are represented by a coarse-grained model that facilitates the handling of non-bonding base–base interactions (Figure 1). The backbone structure is approximated by two pseudoatoms positioned at P and C4' to represent the phosphate and sugar moieties, respectively. Base moieties are represented at three levels: level 1—three beads, positioned at the following atoms: N1-C2-C4 for pyrimidines and N9-C2-C6 for purines; level 2—the midpoint located between atoms N1 and C4 in pyrimidines and between atoms N9 and C6 in purines; level 3—a 3D cubic grid (lattice spacing of 0.5 Å) that carries information about the excluded volume of all atoms of the base moiety, and, even more so, preferences of the nucleotide residue for non-bonding interactions.

The SimRNA coarse-grained representation reduces the number of explicitly represented atoms from 30 to 34 (20–23 non-hydrogen) per residue, down to five, while it retains the key properties of an RNA chain. In particular, three pseudobonds (level 1 of the base moiety representation) define the position and orientation of the base moieties and

they approximate the Watson–Crick, Hoogsteen and sugar edges that can be used to represent all major interactions made by bases with each other as well as with the backbone (31,32). It is worth emphasizing that the pseudobond that connects the C2-C4 atoms in pyrimidines or the C2-C6 atoms in purines is parallel to the Watson–Crick edge (Figure 1B). This representation not only captures the geometry and stereochemistry of the RNA chain, but also facilitates the visual analysis of complex structures and interactions displayed in a reduced representation.

The backbone representation allows for calculation of the pseudotorsion angles η and θ (spanned on C4'–P–C4'–P and P–C4'–P–C4' atoms, respectively) that can be used to classify all major conformations of the RNA chain in a manner similar to the Ramachandran plot for proteins (33) (Figure 1). Similar backbone representations have been used in other coarse-grained models of RNA, including VFOLD (23) and DMD/iFoldRNA (10).

SimRNA utilizes two kinds of local coordinate systems (Supplementary Figure S1). The coordinate systems of the first kind are defined based on the atoms of the backbone. A local coordinate system is centered on each C4' atom of the backbone. They are used to position each base in relation to the backbone. The local coordinate systems of the second kind are centered on the midpoints of the bases. For each base, three atoms (level 1) are used to define its local system of coordinates, and are used to triangulate the midpoint (level 2) of the interacting bases. Axes of the local system of coordinates serve as the axes of the 3D grids for storing the statistical potential (level 3).

Form and derivation of the SimRNA energy function

The energy function of SimRNA is composed of statistical potential terms, derived from the observed frequencies of occurrence of various proximate structural patterns (base–base contacts, short backbone fragments, etc.). To compute the statistical potential, a manually curated set of RNA 3D structures was selected from the Protein Data Bank. We intentionally separated the set of structures used for the derivation of the potential from those used for testing SimRNA (see below). During the initial step, we selected RNA structures obtained by X-ray diffraction (of resolution higher than 3.2 Å and more than 20 residues long). We analyzed all these structures in detail, and excluded ones that contained large gaps or where the conformation of the RNA molecule was significantly influenced by interactions with other molecules (e.g., proteins). In order to remove sequences that were closely similar to each other, we used the BlastClust tool (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) from the NCBI–BLAST package (34), with a 50% sequence identity threshold. From each cluster, we selected only one structure solved at the highest resolution. For ribosomal RNA, we manually selected five structures solved at highest resolution (PDB ids: 1n32, 3i1m, 3cc2, 3kni, 3i1p). From the resulting data set, we also removed sequences with 50% or more identity to RNA molecules in the two previously published test sets (8,10), which we used for testing of our method (see below). The resulting data set contained 150 structures in total (21238 residues in total) and the data set was used to extract the statistical preferences for base–base and base–backbone interactions, which in turn were used to infer the corresponding terms of the statistical potential.

The energy function of SimRNA is composed of two classes of terms: sequence-independent local terms, associated with the local geometry of the RNA backbone, and sequence dependent long-range terms, associated with pairwise interactions between nucleotide residues. The local terms are functions of bond lengths (one term per virtual bond P–C4' and C4'–P), flat angles (one term per angle defined by the following trios of consecutive atoms: P–C4'–P and C4'–P–C4'), and torsion angles (one 2D term dependent on the subsequent pseudotorsion angles η and θ) (35). Values of 1D terms that control bond lengths or angles are stored in tables (1D arrays), while the values of the 2D term (η-θ) that controls two subsequent torsion angles are stored in a 2D array (Figure 1).

Long–range terms describe base–base, base–backbone and backbone–backbone interactions. Data about interaction preferences for the bases are stored in 3D arrays. The base–backbone interaction terms depend on the positions of the P and C4′ atoms of the interacting backbone moiety in the coordinate system of the reference base. Backbone-backbone interactions are modeled as sums of statistically derived 1D functions of interatomic distances between C4′ atoms. These latter terms are generic (do not depend on sequence or orientation).

The derivation of local terms was done by binning values of bond lengths and angles along the backbones of our curated set of RNA structures. Then the tables of values of counts were smoothed out and normalized by dividing them by their averages. Non-zero values of the tables were subjected to a negative logarithm function. Zero and positive values (from previous step) above 3.0 were set to 3.0.

The initial step of deriving the long-range terms was detecting the base–base contacts and base–backbone contacts. The details of contact classification are described below. For each type of base, points corresponding to the contacts were transformed into the local coordinates of the reference base. In the case of base–base contacts, the points were at the midpoints of the contacting bases. In the case of base–backbone, the points corresponded to the contacting P or C4' atoms. This way, we obtained 16 clouds of points corresponding to base–base contacts between a base of type X and a base of type Y, where X and Y are A, C, G and U. Additionally we obtained 8 clouds of points corresponding to base–backbone contacts of types X-C4' (4 clouds) and X-P (4 clouds), respectively, where X is a base. The clouds of points were then binned into 3D grids (with a lattice spacing of 0.5 Å and the location specified using the lattice indices i, j and k). Then the grids were dispersed by convolution with a symmetric Gaussian function. For normalization purposes, the base–base grids were summed together into a new 3D grid, $\{\mathrm{A}_{ijk}\}$. The mean value of all cells of $\{\mathrm{A}_{ijk}\}$, exceeding the threshold of 0.3, became the normalization constant $\langle a \rangle$. For base–base grids, each non-zero cell of each grid was subjected to the expression:

$$E_{\mathrm{XY}}(ijk) = \min\{-\log\left[\mathrm{XY}_{ijk}/(\langle a \rangle \cdot \chi_X \cdot \chi_Y)\right], 0\} \quad (1)$$

where χ_{X} and χ_{Y} correspond to the mole fractions of the respective bases. For base–backbone grids, each non-zero

cell of each grid was subjected to the expression:

$$E_{\mathrm{XY}}(ijk) = \min\{-\log\left[\mathrm{XY}_{ijk}/(\langle a\rangle \cdot \chi_X \cdot 1)\right], 0\} \quad (2)$$

where χ_X corresponds to the mole fraction of the respective base, and the mole fraction of P or C4' are assumed to be 1.

In the process of developing the statistical potential, we tested many different ways of processing the data for the interacting residues and found that the best results were obtained with the following setup: (i) a term for base–base interactions was derived from canonical and non-canonical base pairs detected with RNAView (36) and base stacking detected with our in-house classifier (i.e., other geometries of physically interacting bases were ignored to reduce the background 'noise'); (ii) a term for base–backbone interactions was derived from residue pairs, in cases where any heavy atom of a base moiety of one residue was at a distance ≤ 5 Å from a P or C4′ atom of the other residue. To obtain a proper balance between the stacking and lateral base–base interactions, the number of points corresponding to stacking was reduced (see Supplementary Information).

The excluded volume of each type of base corresponds to the all-atom representation (including the hydrogens) of the base projected onto the grid with positive values. The size of the atoms was adjusted to reproduce the real volume of the base within the assumed base–base contact model.

Calculation of the energy in SimRNA

The total energy for a specific frame during a simulation is given by Equation (3):

$$E_{tot} = \sum_{bonds} E_{bonds} + \sum_{\substack{flat\\angles}} E_{\substack{flat\\angles}} + \sum_{\eta-\theta} E_{\eta-\theta} + \sum_{base-base} E_{base-base} + \sum_{\substack{base\\bbone}} E_{\substack{base\\bbone}} + \sum_{\substack{bbone\\bbone}} E_{\substack{bbone\\bbone}} \quad (3)$$

where base–bbone is the base–backbone interaction (X-P and X-C4′), and bbone–bbone is the backbone interactions between different C4' atoms of sugar moieties. The energy values for local geometrical terms and long-range terms (E_{bonds}, $E_{flat-angles}$, $E_{\eta-\theta}$, $E_{base-base}$ and $E_{base-backbone}$) are obtained from dedicated tables.

Calculation of the energy in the base–base interactions (e.g., X and a second base in close proximity Y) is as follows. The first base (X) is set as the reference, the center position of Y is transformed to the local coordinates of X to obtain the position *ijk* that is referenced by XY_{ijk}, and the energy $E_{\mathrm{XY}}(ijk)$ is obtained from the corresponding cell of the interaction grid XY. Then the reciprocal procedure is done where Y is set as the reference base, X is transformed to the local coordinates of Y, and the energy $E_{\mathrm{YX}}(i'j'k')$ for $\mathrm{YX}_{i'j'k'}$ is obtained from the grid YX. The total energy for this interaction is the sum of these two energies: $E_{\mathrm{XY}}(ijk)$ + $E_{\mathrm{YX}}(i'j'k')$. This reciprocal operation *reinforces* the geometries that favor strong base–base interactions (Supplementary Information). To help enforce the planarity of the base–base interaction terms, an angle dependent term is also computed for each pair of bases that depends on an additional weight factor $\sqrt{|\cos(\angle\hat{\mathbf{w}}_2\hat{\mathbf{w}}'_2)|}$, where the $|\ldots|$ indicates the absolute value and $\angle\hat{\mathbf{w}}_2\hat{\mathbf{w}}'_2$ is the angle between the normal vectors of the interacting bases: $\hat{\mathbf{w}}_2$ of base X and $\hat{\mathbf{w}}'_2$ of base Y (Supplementary Figure S6). The square root was used because it permits a less constrained planar geometry for the interacting bases. Base–backbone interactions are calculated in a similar way as the base–base interactions except that the backbone P or C4' are transformed into the local coordinates of the reference base. The energy value is also obtained from the corresponding dedicated grid. Backbone-backbone interactions are based on the distance between the two C4' positions.

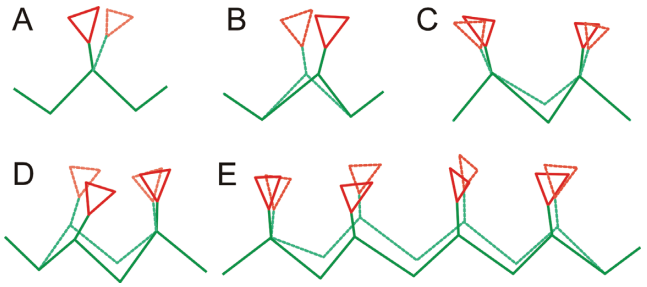

Figure 2. Examples of the Monte Carlo move set. During a simulation, each new conformation is generated as a small modification of a previous conformation: (**A**) a change in the conformation of the base in the local backbone coordinates; (**B**) a change in the backbone position of the C4′ atom; (**C**) a change in the backbone position of P atom; (**D**) a change in the position of two subsequent atoms of the backbone; and (**E**) a change in the direction of a fragment of the backbone.

Conformational sampling method

Sampling of the conformational space is accomplished in SimRNA by the use of an asymmetric Metropolis algorithm (30), which is executed by calling either of two schemes: single thread simulations or replica exchange Monte Carlo. The single thread variant allows for performing isothermal simulations and simulations with a gradual increase or decrease of temperature; e.g., to study RNA unfolding.

Conformational changes are accomplished via a specific set of moves (Figure 2). There are two basic types of moves. First, there is an exchange of a single nucleoside conformer by another one (from an internal database of conformers), which changes the orientation of the base with respect to the backbone. Second, there is an alteration of the backbone conformation, associated with maintaining the conformations of the base moieties in their local backbone coordinates. The latter type of moves may involve a change in the position of a single C4' or P atom of the backbone, a change in the position of two neighboring C4' and P atoms or translation and rotation of a chain fragment. The type of move and the atom or chain fragment to be moved are both selected randomly. Default values of relative frequencies of moves were defined based on a large number of tests (data not shown) and they can be modified by the user. The simulation is conducted in steps that comprise the number of attempted moves (accepted or rejected subject to the Metropolis criterion) equal to the number of residues in the structure.

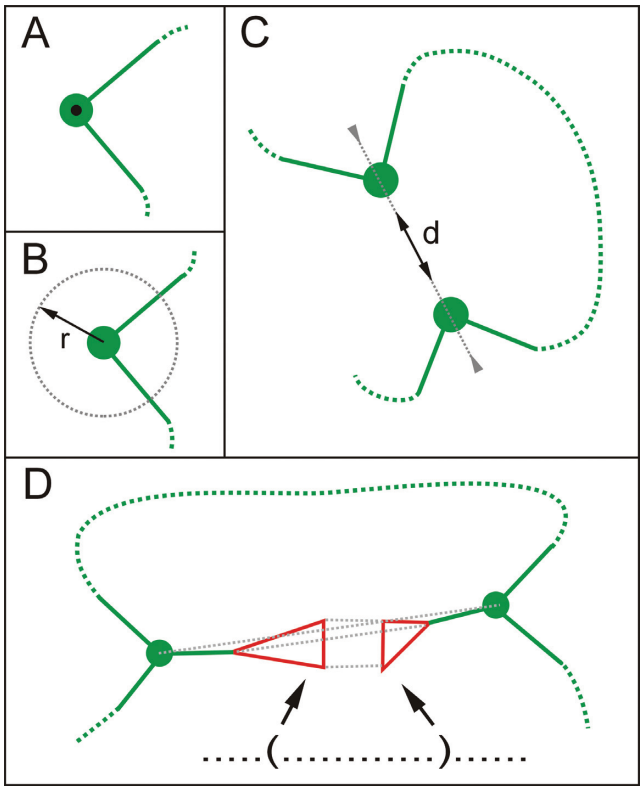

Figure 3. Distance restraints implemented in SimRNA. (**A**) immobilization of one atom; (**B**) flexible pinning of one atom; (**C**) flexible tethering of two atoms; (**D**) canonical base-pairing of two residues.

Restraints

SimRNA can use additional information about the RNA structure, obtained from experimental analyses, from independent computational predictions, or postulated by the user. Three types of user-specified restraints are currently implemented in SimRNA (Figure 3): on atomic positions (immobilization or flexible pinning), on inter-atomic distances (flexible tethering) and on the secondary structure (base-pairing). Positional restraints are used to restrict the movement of selected atoms, which can range from complete immobilization (frozen) to flexible pinning that keeps the atom close to its starting position. Immobilization was implemented as a modification to the sampling algorithm, while flexible pinning, and in fact all other restraints mentioned below, were implemented as additional penalty terms added to the energy function.

Distance restraints serve as pairwise flexible tethers (Supplementary Figure S7). For any pair of atoms, an allowed distance range can be specified. Departure beyond this range results in a penalty that scales linearly with the magnitude of the deviation. The allowed distance range can be based on experimental measurements of intramolecular distances, for example from the Förster Resonant Energy Transfer (FRET), or Electron Spin Resonance (ESR) experiments, or from chemical cross-linking. Further, theoretical predictions of intramolecular contacts can be utilized; e.g., from sequence covariation analysis that may identify important tertiary contacts without specifying the type of contact. This type of restraint may also be used to specify non-canonical base pairs.

The role of secondary structure restraints is to specify the desired canonical Watson–Crick (cis), and wobble base pairs; this type of restraints may include pseudoknots of any type. For specified bases that require pairing, a penalty is associated with a deviation from the reference geometries specific for a given type of contact. Secondary structure restraints are internally represented as distance restraints imposed on the atoms of the interacting bases. By default, SimRNA does not penalize the formation of base-pairs that are not specified in the file with restraints.

Input and output

A typical SimRNA input comprises a starting structure (PDB-formatted) or a sequence (ASCII-formatted) file, a configuration file that contains the basic parameters of the simulation to be performed (e.g., simulation length, temperature range, non-default parameters, etc.), and an optional file with restraints. If no starting structure is provided, then based on the provided sequence, SimRNA generates a circular conformation with the 5′ and 3′ ends close to each other. SimRNA can handle RNA molecules composed of one or multiple chains (up to 52) and it allows for simulations of a part of the system to be performed, with the conformation of the remaining part frozen or restrained. The current version is capable of handling RNA sequences with standard RNA (A, U, C, G) residues only; a representation of modified residues will be implemented in the future. Secondary structure restraints can be specified using the multiline dots-and-brackets format, which allows for defining RNA pseudoknots. The dots-and-brackets input is parsed and internally converted into the dedicated list of restraints.

The output of a simulation is recorded as a trajectory file (or set of files) comprising the lowest-energy conformations selected from a consecutive series of simulation steps. SimRNA is accompanied by a software package for the processing of trajectory files. The content of the trajectory files (in the form of individual frames or a series of such frames) can be visualized, converted to PDB files, searched for structures with desired properties (lowest global energy, lowest RMSD to a reference structure), or subjected to clustering.

The trajectory can be converted to a series of files in PDB format containing models in either the reduced SimRNA representation or models rebuilt to an all-atom representation. The rebuilding is done using a built-in algorithm based on fragment matching. By default, the output also includes information about the energy value and about the secondary structure of the current conformation (expressed in dots-and-brackets format). The secondary structure is detected using a classifier built into SimRNA, which operated on the reduced representation of the 3D structure. SimRNA can be also run in a ‘zero steps’ mode; i.e., take as an input a single PDB file and output the corresponding secondary structure and SimRNA energy value.

SimRNA employs a clustering protocol that is commonly used for protein 3D structure prediction; e.g., in ROSETTA (37). First, the RMSD values are computed between all pairs of structures of the simulation trajectory or for a subset defined by the user. Second, a cluster with the largest

number of structures within a predefined RMSD threshold value is identified, and its members are removed from the initial set. Subsequent clusters are found by iterating these steps until all the structures from the initial set have been assigned to their respective clusters. Based on our experience, we typically use a clustering threshold equal to 0.1 Å times the sequence length; i.e., 5 Å for a sequence of 50 residues, and we consider medoids of the three largest clusters of decoys as well as the decoy with the lowest energy; this procedure was used in this work. However, other protocols of clustering and data retrieval can be used depending on the purpose of the modeling (e.g., for conformational sampling, other thresholds can be used and a larger or smaller number of cluster representatives can be obtained).

Runtime

To predict each RNA structure reported in this article, we have run simulations comprising 8 independent instances of the replica exchange method (10 replicas each), with each thread running on a separate CPU. For each set of simulation we employed 80 CPU cores (AMD Opteron 2.2 GHz) of an in-house computing cluster. Each thread comprised of 16 million Monte Carlo steps. Thus, the runtime of a thread depended mostly on the size of the simulated system. Example runtimes (per thread) for exemplary RNAs with different lengths were as follows: 1zih (12 nt) 3 h, 2tpk (36 nt) 6 h, 1y26 (71 nt) 20 h and 1gid (158 nt) 86 h.

RESULTS

The ability of SimRNA to fold RNA sequences into native-like 3D structures has been tested on five benchmark sets of experimentally determined RNA structures. The first data set (10), hereafter referred to as 'Ding et al. data set', comprises 153 structures of single-chain RNAs. The length of sequences in this test set varies from 20 to 100 nucleotide residues; however, the majority of sequences are shorter than 50 nt, and some of the sequences are redundant (e.g., 1cq5 and 1cql). In this data set, 145 structures were obtained from nuclear magnetic resonance (NMR) spectroscopy, and only eight were obtained from X-ray crystallography. Most of these structures are relatively simple; nonetheless, they contain a variety of structural motifs such as three- and four-way junctions, kink-turns and pseudo-knots. The second benchmark set, taken from (8) and hereafter referred to as 'Das&Baker data set', is composed of 13 RNA structures determined by X-ray crystallography and 7 structures determined by NMR. In this data set, most RNAs are rather small (size 12–41 residues); however, nine structures are composed of two RNA chains and one is composed of four chains, which allowed us to test the ability of SimRNA to simulate and predict structures of RNA–RNA complexes. The third data set, taken from (38) and hereafter referred to as 'Seetin&Mathews data set', comprises only five structures of relatively large RNA molecules (43–158 residues), for which low-resolution experimental data are available that have been used to aid in the structure prediction. This data set allowed us to test the ability of SimRNA to predict RNA 3D structures with the aid of distance restraints. Five structures (1esy, 1kka, 1qwa, 28sp, 2f88) are common to both the Ding et al. and Das&Baker sets, and the structure 1evv is common to the Ding et al. and Seetin&Mathews sets). The fourth data set consists of short 3D motifs used to test the FARFAR method (39), which will be referred to as the 'motifs data set', and the fifth set is taken from the RNA Puzzles challenge (Puzzles 1–6, 8, 10 and 12 (13,40)) and will be called the 'RNA Puzzles data set'.

For all sequences in the benchmark sets, we carried out tertiary structure prediction by *de novo* folding with SimRNA (folding using sequence alone) as well as folding with restraints on the secondary structure, obtained from the target structures using RNAView (36). For the Seetin&Mathews and RNA Puzzles data sets we also predicted structures using restraints on both secondary structure and tertiary contacts, to mimic the predictions reported in these original works (13,38,40). The motifs data set contained only short segments of RNA 3D structures, so the structures could only be tested with the end parts of the structure restrained. For each prediction, we carried out eight independent runs of the Replica Exchange Monte Carlo simulation, each employing 10 replicas. Each run comprised 1000 simulation intervals (16000 steps each) and the lowest energy frame from each interval was recorded. The resulting eight trajectories were combined with each other to yield 80000 conformations per target (1000 conformations from each of the 10 replicas in each of the 8 simulation runs) and the best 1% scored conformations from the set were retrieved and clustered (see Methods for details).

The assessment of RNA structures requires analysis of both the global conformation, and the local features such as interaction patterns (41). To measure the accuracy of the predicted structures, we compared them with the corresponding entries in the PDB; we used the RMSD to describe the global deviation in positioning of the atoms in space and the Interaction Network Fidelity (INF) to describe the agreement of the interactions between the predicted and reference structures (based on the ClaRNA classifier (42), using both canonical and non-canonical pairs as well as stacking). For calculation of the significance of the 3D structure predictions of single chain RNAs, we used the procedure proposed by Hajdin et al. (43). We have also analyzed the accuracy of the predicted secondary structure (in which GU pairs were treated as canonical). The results of the RNA 3D structure predictions on the five above-mentioned benchmarks results (average RMSD and interaction network fidelity values) are summarized in Table 1, and detailed results are provided in Supplementary Table S1. Models generated by SimRNA are available for download from ftp://ftp.genesilico.pl/pub/software/simrna/.

RNA 3D structure prediction without any restraints

The results of the tests clearly show that SimRNA performed well in predicting both simple and complex RNA structures from sequence information alone, without restraints on the secondary or tertiary structure. Predictions generated by SimRNA (see Supplementary Table S1 for details) have largely correct secondary structure (average sensitivity 89%/83% and positive prediction value 82%/77% for the Ding et al./Das&Baker data sets) and recapitulate the majority of contacts including canonical and stacking

Table 1. Summary of average and median (bold font) structure quality measures obtained for RNA structure predictions analyzed in this work

RNA folding method and (optionally) restraints used	lowest energy decoy		First cluster		lowest RMSD	
	RMSD	INF	RMSD	INF	RMSD	INF
	Ding et al. data set (10)					
SimRNA, no restraints	4.74/**3.72**	0.80/**0.82**	4.32/**3.37**	0.81/**0.83**	2.45/**2.18**	0.84/**0.84**
SimRNA, SS restraints	4.46/**3.80**	0.82/**0.83**	4.07/**3.40**	0.82/**0.83**	2.31/**2.13**	0.85/**0.85**
Ding et al. (10)			3.80/**3.25**			
DMD/iFoldRNA server (44)			6.27/**4.46**	0.74/**0.78**		
	Das&Baker data set (8)					
SimRNA, no restraints	4.27/**3.81**	0.80/**0.82**	4.17/**3.60**	0.80/**0.82**	2.81/**2.32**	0.81/**0.86**
SimRNA, SS restraints	4.16/**3.89**	0.81/**0.83**	3.89/**3.47**	0.81/**0.83**	2.48/**2.23**	0.83/**0.85**
Das&Baker (8)			4.91/**3.93**			
	Seetin&Mathews data set (38)					
SimRNA, no restraints	23.90/**24.89**	0.61/**0.66**	23.80/**24.18**	0.60/**0.71**	10.53/**10.72**	0.63/**0.70**
SimRNA, SS restraints	18.47/**18.51**	0.73/**0.81**	17.49/**18.17**	0.71/**0.77**	6.47/**6.94**	0.74/**0.85**
SimRNA, SS+exp. restraints	6.30/**5.91**	0.70/**0.80**	7.70/**5.82**	0.70/**0.79**	4.10/**3.74**	0.72/**0.81**
Seetin&Mathews (38), SS restraints					12.93/**13.28**	
Seetin&Mathews (38), SS+exp. restraints					9.24/**8.58**	
	RNA Puzzles data set (13,40)					
SimRNA, no restraints	21.5/**20.9**	0.64/**0.63**	24.0/**23.3**	0.65/**0.64**	13.2/**13.3**	0.69/**0.66**
SimRNA, SS restraints	17.3/**17.5**	0.75/**0.76**	17.2/**15.1**	0.76/**0.78**	8.7/**7.7**	0.75/**0.77**
SimRNA, SS+exp. restraints	15.1/**14.0**	0.70/**0.72**	16.8/**14.5**	0.71/**0.73**	9.2/**8.5**	0.69/**0.74**
Best models in RNA Puzzles (13,40)					9.21/**9.15**	
	FARFAR motifs data set (39)					
SimRNA, only termini restrained	2.13/**1.66**	0.87/**0.88**	1.50/**1.21**	0.87/**0.89**	1.00/**0.84**	0.87/**0.86**
FARFAR, Das et al. (39) - best out of 5 clusters			3.84/**2.35**		1.98/**1.40**	

Complete detailed results are presented in Supplementary Table S1.

interactions (with an average INF of 80% for both data sets) and for non-canonical interactions about 55%). Tertiary structure is also largely correct. It is worth noting that all pseudo-knotted structures of chain length up to 50 residues were properly predicted in the absence of restraints; hence, SimRNA can be used for *de novo* prediction of pseudoknots. If the best models (medoids of largest clusters) are considered for each RNA across the benchmarks, then using Hajdin et al.'s criterion of significance (HCS) (43), SimRNA proposed significantly correct predictions ($P < 0.01$, according to HCS) for 145/153 (95%) and 9/10 structures (90%) of single-chain RNAs in the Ding et al. and Das&Baker data sets, respectively. It must be emphasized that the HCS was developed for single-chain structures and in its original implementation it cannot be used to evaluate the quality of structures composed of two or more chains. This is particularly relevant for the Das&Baker and motifs benchmarks, which contain multi-chain RNAs.

As expected, the results from *de novo* folding of the Seetin&Mathews and RNA Puzzles data sets were predictably lower: INF roughly 60% for all contacts including stacking, 50% for canonical pairs and 20–30% for non-canonical pairs. These structures are generally very difficult to model, which is why the RNA Puzzles challenge is so important for the community of researchers working on RNA 3D structure prediction (13,40).

For the Das&Baker data set, in 7 cases out of 10, SimRNA generated more accurate predictions (medoids of largest clusters) than the ones reported by Das&Baker. If only one best-scored model is considered per target, 138/153 (90%) and 9/10 (90%) significantly correct predictions were obtained for single-chain RNAs in Ding et al. and Das&Baker data sets, respectively. Only in one case (2a9l structure), the energy criterion alone allowed us to obtain a significantly correct prediction in the absence of a correct prediction in the first cluster; however, in this case the second cluster medoid was significantly correct. On average, models selected by clustering were more accurate than models selected based on energy alone (110/153 cases, and 13/20 cases in Ding et al. and Das&Baker data sets, respectively).

In one case where a single-chain RNA was folded without any restraints (2evy in the Ding et al. data set), SimRNA failed to produce any conformations that could be evaluated as significantly correct. For six cases in the Ding et al. benchmark (1bgz, 1evv, 1k2g, 1oq0, 1xwp, 2f87), and for one case in the Das&Baker benchmark (1zih), SimRNA was able to generate such a conformation in the course of the simulations, but neither the best-scored structure nor the top three cluster medoids were significantly correct according to HCS. It is worth noting that for 2f87 and 1zih structures, SimRNA generated models that were very close to the experimentally determined reference (RMSD 1.20 Å and 1.36 Å, respectively), but these RNAs are very small; hence, the values of RMSD did not meet the HCS.

RNA 3D structure prediction with restraints on secondary structure

With secondary structure provided as restraints, the results of the 3D structure predictions typically improved. Interestingly, the use of secondary structure restraints had a negligible influence on recapitulation of all types of contacts, as the average INF value remained close to 77% for both data sets. For small structures, the improvement in terms of secondary structure and RMSD to the reference structure was usually small. Hence, for the Das&Baker data set, the improvement due to the use of restraints was negligible. However, the use of secondary structure allowed SimRNA to generate significantly correct predictions (both in terms

of the best energy and the first cluster medoid) for some RNAs from the Ding et al. data set that could not be folded without restraints (1bgz, 1evv, 1k2g). In general, the secondary structure restraints significantly improved the predictions for large RNAs. Again, if the medoids of the largest clusters are considered as results of the 3D folding with secondary structure restraints, then SimRNA proposed significantly correct predictions (with a reference to the entire unrestricted search space) for 149/153 structures (97%) in Ding et al. data set. Across both data sets, SimRNA folded correctly 158/163 single-chain structures. It was unable to generate significantly correct predictions only for five very small RNAs: 2f87 (12 nt), 2evy (14 nt), 1oq0 (15 nt) and 1xwp (15 nt) in the Ding et al. data set, and 1zih (12 nt) in the Das&Baker data set. Models were native-like with respect to the secondary structure and tertiary fold, but for such small structures, SimRNA predictions were not precise enough to be evaluated as significant according to HCS.

These results obtained with SimRNA compare well with predictions reported by the authors of the aforementioned benchmarks. In the original work of Ding et al. (10), 149/153 structures were also folded below the level of 'correctness' according to HCS. The RMSD values for predictions reported by Ding et al. (10) (3.8 Å on the average) were slightly better than we could obtain with SimRNA for that data set (4.1 Å on the average for the first cluster medoids). However, when we used the iFoldRNA server developed by the authors that implements their method (44), iFoldRNA generated significantly correct predictions only for 130/153 structures and the RMSD values of the resulting models were in general higher (6.2 Å on average) than models obtained with SimRNA (Table 1 and Supplementary Table S1). Likewise, in the article by Das and Baker for predictions of single-chain RNA structures obtained with FARNA, 9/10 predictions satisfied the HCS. When 10 multi-chain structures from the Das&Baker data set are considered, models generated by SimRNA with restraints on the secondary structure are better in 7/10 cases than results obtained by Das&Baker (in this case the results are not much different from folding without secondary structure restraints). For the entire Das&Baker data set, SimRNA predictions had an average RMSD of 3.9 Å, which compares favorably to the average RMSD 4.9 Å reported by Das&Baker.

RNA 3D structure prediction with restraints on secondary structure and on tertiary contacts from experimental data

The Seetin&Mathews data set and RNA Puzzles data sets comprise only five and nine cases, respectively, and only five the of RNA Puzzles have experimental probing data. However, these are very special in that they are representative of the class of 'real life' challenges faced by researchers studying RNAs with unknown structures, where knowledge of the RNA secondary structure and sparse tertiary structure information are all that is available for 3D structure prediction. Generally, even with such information, it is often difficult to obtain a correct 3D structure for long sequences, as demonstrated by the RNA Puzzles experiment (13). Therefore, we consider these data sets for a separate type of benchmark from the data sets of Ding et al. and Das&Baker described above. The complexity and problems of folding these long sequences and the experimental data sets from Seetin&Mathews are discussed in detail by the authors (38). Here, we attempted to fold each of these RNAs in three distinct modes: without restraints, with secondary structure restraints and with secondary structure restraints as well as additional restraints derived from experimental data (for Puzzle 10, structure 4lck, we folded only the T-box RNA, and kept the homology model of the tRNA frozen).

The Seetin&Mathews data set proved to be the most difficult. Without restraints, SimRNA was able to provide a significantly correct model for only one of the structures (1e8o) in this data set. The use of secondary structure restraints allowed SimRNA to improve the folding of that structure, as well as to generate significantly correct models for 1evv (common with the Ding et al. data set, described above) and 1kh6. For 1gid and 3zd5, models were generated at the borderline of significant correctness. The use of additional restraints further improved the folding of the most difficult cases (1evv, 1gid and 3zd5), resulting in the generation of significantly correct predictions. The final models (first cluster medoids) had native-like secondary structures, in agreement with restraints (average positive predictive value 0.84, sensitivity 0.91), and reasonable overall contacts: including non-canonical interactions and stacking (INF = 0.73). The average RMDS of these models are relatively high (7.7 Å), but they compare well to the RMSD of models generated by the authors of the reference method (9.2 Å).

The RNA Puzzles data set was found to be also very difficult, and most of the models obtained with *de novo* folding with SimRNA had high RMSD values, with the exception of Puzzle 1 (3mei). The use of secondary structure restraints significantly helped folding Puzzle 2 (3p59). The folding with restraints on tertiary contacts, inferred from the publicly available experimental data, resulted in folding structures of most Puzzles to structures with RMSDs between 10 and 17 Å and most of them had *P*-values indicating statistical significance. Not surprisingly, these models were in general somewhat worse than the winning structures submitted by human predictors in the RNA Puzzles competition, with the exception of Puzzle 12, where SimRNA was able to generate a slightly better model than the best human prediction. Nonetheless, they were actually not much worse than predictions submitted by our own group, which has used SimRNA, often in combination with other programs. These results will be analyzed in detail and will certainly influence our strategy for predicting structures in RNA Puzzles, and will also be taken into account in the future development of SimRNA and its possible automated combination with methods for homology modeling and all-atom refinement.

Folding of RNA 3D structure motifs

Finally, we analyzed the ability of SimRNA to predict the structure of short RNA 3D motifs from the data set used by Das&Baker to test FARFAR. In this data set, the structures were relatively small, but many of them comprised multiple chains, and were often derived from larger structures that did not correspond to autonomously folded structural units. For this data set, we performed simulations with

the base-paired termini of all chains restrained to reproduce the context of each motif, scrambled the structure (including the base-pair termini) using a very high initial temperature and allowed SimRNA to predict the internal structure of the motif. Here, the main question was the ability of the program to predict non-canonical interactions. Data from Supplementary Table S1 demonstrate that, in general, models generated by SimRNA had low RMSDs relative to the native structures, and ideal or nearly ideal inferred canonical base pairs. However, whereas these examples are certainly better than *de novo* folding, only roughly half of the structures had an appreciable fraction of non-canonical pairs predicted correctly. Based on this exercise, we conclude that the improving prediction of non-canonical pairs is a major challenge for coarse-grained modeling. While we intend to improve SimRNA with respect to this type of problems, it may be useful to consider the use of independent methods such as RMDetect (45) or JAR3D (46) to predict local structured motifs before the folding, or to use local high-resolution resampling; e.g., with FARFAR (39) after coarse-grained modeling. The RNA Puzzles experiment provides an excellent platform for testing these and other combinations of solutions in the future.

DISCUSSION

Since Anfinsen, it has been an often held view that the 3D structure of biomolecules (proteins and RNAs) is determined by their sequence, and that the formation of the biologically relevant structure is guided by the minimization of the free energy of the system containing the biomolecule (47,48). This assumption provided a basis for the development of computational methods for protein and RNA 3D structure prediction that sample the conformational space, calculate free energies for the sampled conformations and attempt to identify the global free energy minimum (5,49). Ideally, the function with which to calculate the energy should be based on a quantum-mechanical description of the system, however such calculations of even a few hundred atoms are extremely costly and therefore applicable only to very small molecules. Hence, various simplifications must be employed. A particularly successful simplification used for protein structure prediction has been coarse-graining, in which an atomistic description of a molecular system is replaced with a less complex model, where groups of atoms are treated as single interaction centers (50). The development of a coarse-grained model is challenging, because the reduction in detail of the representation must be accompanied by modifications of the energy function to capture the key interactions that are responsible for folding: kinetics and thermodynamics.

SimRNA is a new coarse-grained RNA model, in which the explicit representation has been reduced to five atoms per ribonucleotide residue, and in which the physical energy function has been approximated by a statistical potential derived from a database of experimentally determined structures. The conformational space is sampled by means of Monte Carlo simulation. This approach has been strongly inspired by coarse-grained models developed for protein structure prediction, in particular CABS (15) and REFINER (14). The process of development of SimRNA from the preliminary version with only three atoms per residue (51) to the current one has been greatly aided by blind tests performed in the context of the RNA Puzzles experiment (13). In particular, the development of the three atom description of the base has been dictated by the need to differentiate better between stacking and base-pairing interactions, which is now reflected in an explicit representation of both the base faces and edges. Tests carried out for RNA Puzzles have also prompted the development of various types of restraints that can be used to guide the folding.

We have extensively tested the SimRNA version described in this article by performing RNA folding simulations, and we have compared its performance to other successful models developed previously. The benchmark results suggest that SimRNA runs carried out with only sequence information often recapitulate the native-like secondary and tertiary structure, especially for relatively short RNA sequences, up to ≈50 nt. For the structure prediction of longer molecules, sampling of the vast conformational space becomes a limiting factor, which can be aided by the use of additional restraints on the secondary structure and long-range tertiary contacts. Still, SimRNA exhibits comparable performance (or better) than other methods that use energy functions based on force fields derived from a more directly physical description of intramolecular interactions. It is known that in RNA simulations using pairwise potentials, reliable reproduction of the correct handedness of RNA helices (and possibly other structural motifs) can be a challenge (21). This problem has not been observed in SimRNA, where the energy function terms—especially the torsional angle η-θ in the backbone and the base–base interaction preferences stored in the 3D grids—together energetically favor right-handed A-helices and render left-handed helices unstable over the recommended temperature range.

It is particularly noteworthy that SimRNA can accurately predict both secondary structure and the global conformation of pseudoknots. For all 14 out of 15 pseudoknotted structures in the Ding et al. and Das&Baker data sets, SimRNA generated significantly correct predictions (according to HCS) without the use of secondary structure restraints, and only for the 1evv structure, which contains a very weak pseudoknot, secondary structure restraints were necessary to obtain a significantly correct prediction. SimRNA can also be used to characterize the conformational space and highlight potential alternative structures. Figure 4 illustrates the case of a pseudoknotted RNA: gene 32 messenger RNA pseudoknot of bacteriophage T2, (PDB id: 2tpk, 36 residues). SimRNA was able to identify a native-like 3D structure (largest cluster of solutions), with secondary structure identical to that of the experimentally determined reference. It is noteworthy that alternative non-pseudoknotted hairpin-loop structures also emerged as well (clusters 2 and 3), which exhibited low energies, but could be successfully discriminated from the correct solution. The analysis of the folding trajectories provided useful insight not only into the final folded 3D structure, but also into the structures of potential folding intermediates. In the SimRNA simulations of this RNA, the 5′ hairpin folded first, and in order to form the pseudoknot, the 3′ tail had to bend and form base pairs with residues in the loop formed by the 5′ hairpin. Thus, SimRNA can be used not only for 3D

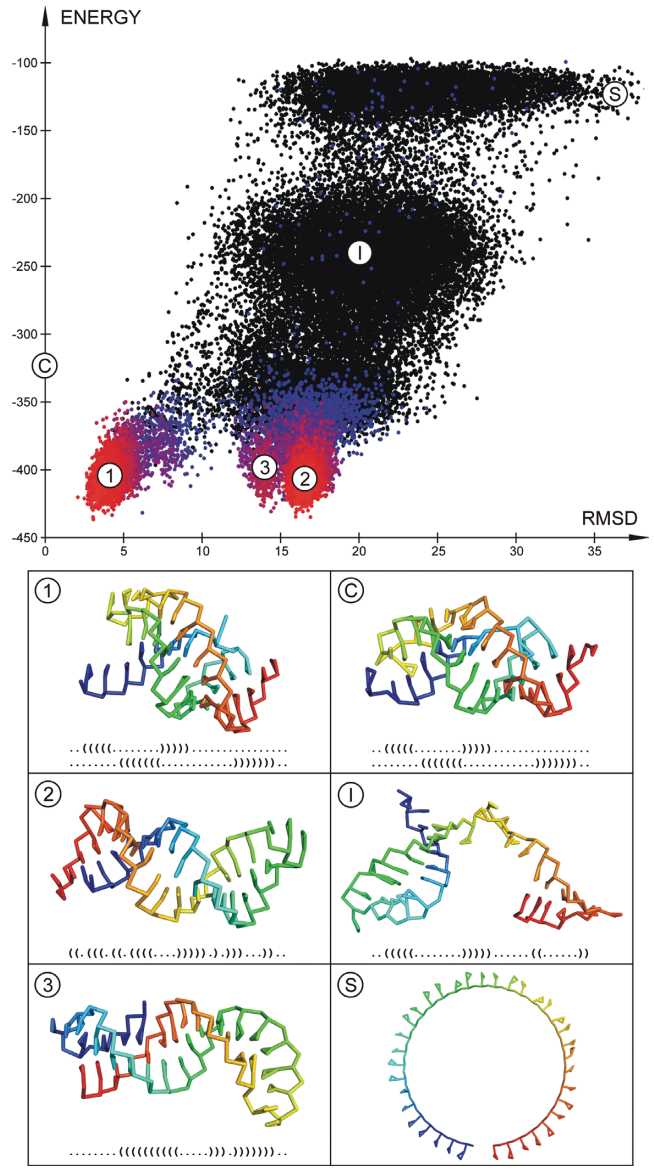

Figure 4. An example of the energy landscape generated in the course of a set of SimRNA simulations. Results are shown for the gene 32 messenger RNA pseudoknot of bacteriophage T2 (PDB id: 2tpk). The upper panel illustrates the relationship between the distance to the reference structure (expressed in RMSD), and the energy of a given conformation (calculated according to the SimRNA statistical potential). Each conformation recorded in the course of the simulation is represented by one dot; where the dots are colored (red to purple to black) according to the conformation's similarity to other conformations. Structures that have many similar conformations are colored red, and structures that have rather unique conformations are colored in black, purple being in-between. The starting conformation is indicated by (S), the reference structure determined by X-ray crystallography is indicated by (C), an example intermediate structure is indicated by (I), and the top three clusters are indicated by (1), (2) and (3). The bottom panel illustrates the tertiary and secondary structure of these conformations. RNA molecules are colored by a spectrum from blue (5′ terminus) to red (3′ terminus) and the secondary structure is shown in dot-bracket format.

RNA structure prediction, but also to investigate intermediate states of folding, structural diversity of intermediate states, and the order of formation of specific parts of the final structure. This can aid in inferring the RNA folding pathways. SimRNA can also be applied to simulations of structure unfolding, and to isothermal simulations that allow determination of the relative stability of different regions of an RNA structure.

SimRNA can be also used to add missing fragments of RNA 3D structures and to remodel uncertain parts of structures obtained with other methods; e.g., by homology modeling. Because of space constraints we have not analyzed these applications in this article, however examples of successful application of SimRNA to such problems have already been published; e.g., for Puzzle 2 in the first edition of RNA Puzzles (13) or for the S6S18CBM RNA motif (52). A practical application of SimRNA for RNA folding with restraints has been also demonstrated (53).

Limitations of the current methodology and prospects for future development

SimRNA is capable of folding RNA molecules of different sizes, with and without additional restraints. However, there are certain limitations of this method that should be taken into account. First, SimRNA, as a coarse-grained method, does not represent all the details of RNA structures ideally. The native-like coarse-grained models are expected to be close to the experimentally determined structures, but they are typically not closer than 2–3 Å in terms of RMSD. Experimentally determined structures often exhibit relatively high energies according to the SimRNA scoring function (see for example Figure 4), and their minimization in the SimRNA force field introduces slight distortions due to 'idealization' of various geometrical parameters inherent to the reduced model. Second, because the energy function is rooted in statistics, SimRNA best recapitulates the structural motifs that are most frequent; i.e., canonical base pairs and stacking. Non-canonical interactions, especially the rare ones, are not scored as highly favorable, and they are very difficult to capture. Both of these issues can be addressed by introducing a high-resolution refinement of SimRNA-generated models, with an energy function that takes into account the true strength of the interactions and does not penalize interactions that are statistically rare, but physically strong. We have already developed an independent computer program QRNAS dedicated to such refinement (J.M.B. and Juliusz Stasiewicz, unpublished data) and we demonstrated its applicability in the context of the RNA Puzzles experiment (40). Another solution to be tested and potentially implemented in the future would be to rescale the relative as well as the absolute strengths of interactions represented in SimRNA according to values determined experimentally as well as values that could be obtained from simulations of RNA molecules with fine-grained methods and high-end physical force fields. Finally, folding of large RNA molecules with SimRNA is computationally demanding, as the program has to sample many different 3D architectures. Thus, modeling of large RNA structures with SimRNA may be jump-started by using starting models generated by other modeling methods developed to predict the

global architecture; e.g., by comparative modeling (6,54) or by sampling of helical topologies (55,56), and the conformational space to be sampled may be restricted by the use of additional restraints (57).

AVAILABILITY

SimRNA is written in C++ and currently is only available for the Linux and MacOSX operating systems. A Windows version is also planned. The source code of SimRNA is not distributed due to intellectual property restrictions. Compiled Linux binaries for Intel and AMD (32 bit and 64 bit) are available from http://genesilico.pl/simrna/. The multiprocessor code requires openmp. MacOSX binaries are compiled with OSX.6/7 support and can run on most MacBook Pro and Air distributions. Users interested in obtaining compiled binaries in some other distribution must contact the authors. The use of a compiled version of SimRNA is free for non-commercial use by academic users. Non-academic users and those interested in commercial use must contact J.M.B. to obtain a commercial license.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Nikolay Dokholyan for providing software to calculate HCS, and François Major for providing software for INF calculations. We would also like to thank Albert Bogdanowicz, Lukasz Kozlowski, Marcin Magnus, Piotr Pokarowski, Claus Seidel and Juliusz Stasiewicz for stimulating discussions and/or for critical reading of the manuscript. We also thank Jan Kogut and Tomasz Jarzynka for maintaining computational facilities in IIMCB and Lukasz Munio for the SimRNA web page.

FUNDING

Polish Ministry of Science [HISZPANIA/152/2006 to J.M.B.; PBZ/MNiSW/07/2006 to M.B.]; European Commission [6FP GA No LSHG-CT-2005-518238 to Reinhard Lührmann, 7FP GA No 316125 to Jacek Kuźnicki]; German Research Foundation (DFG) [GA No SPP 1258 to Claus Seidel]; European Research Council (ERC) [StG grant RNA+P = 123D to J.M.B; Foundation for Polish Science (FNP) [TEAM/2009–4/2 to J.M.B. and 'Ideas for Poland' fellowships to J.M.B.]. Computing power was provided by IIMCB, funded by EU structural funds [POIG.02.03.00-00-003/09 to J.M.B.]. Funding for open access charge: European Commission [7FP grant Fishmed, GA No 316125 to Jacek Kuźnicki].
Conflict of interest statement. Janusz M. Bujnicki is an Executive Editor of *Nucleic Acids Research*.

REFERENCES

1. Atkins,J.F., Gesteland,R.F. and Cech,T.R. (2011) *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.
2. Serganov,A. and Patel,D.J. (2012) Molecular recognition and function of riboswitches. *Curr. Opin. Struct. Biol.*, **22**, 279–286.
3. Leontis,N. and Westhof,E. (2012) *RNA 3D structure analysis and prediction*. Springer-Verlag, Berlin Heidelberg.
4. Doudna,J.A. (2000) Structural genomics of RNA. *Nat. Struct. Biol.*, **7**(Suppl), 954–956.
5. Rother,K., Rother,M., Boniecki,M., Puton,T. and Bujnicki,J.M. (2011) RNA and protein 3D structure modeling: similarities and differences. *J. Mol. Model*, **17**, 2325–2336.
6. Rother,M., Rother,K., Puton,T. and Bujnicki,J.M. (2011) ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.*, **39**, 4007–4022.
7. Rother,M., Milanowska,K., Puton,T., Jeleniewicz,J., Rother,K. and Bujnicki,J.M. (2011) ModeRNA server: an online tool for modeling RNA 3D structures. *Bioinformatics*, **27**, 2441–2442.
8. 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.
9. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
10. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
11. Cao,S. and Chen,S.J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
12. Sijenyi,F., Saro,P., Ouyang,Z., Damm-Ganamet,K., Wood,M., Jiang,J. and SantaLucia,J. (2012) In: Leontis,N and Westhof,E (eds). *RNA 3D structure analysis and prediction*. Springer-Verlag, Berlin Heidelberg.
13. 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.
14. Boniecki,M., Rotkiewicz,P., Skolnick,J. and Kolinski,A. (2003) Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.*, **17**, 725–738.
15. Kolinski,A. (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, **51**, 349–371.
16. Zhang,D., Konecny,R., Baker,N.A. and McCammon,J.A. (2004) Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach. *Biopolymers*, **75**, 325–337.
17. Ponty,Y., Istrate,R., Porcelli,E. and Clote,P. (2008) LocalMove: computing on-lattice fits for biopolymers. *Nucleic Acids Res.*, **36**, W216–W222.
18. 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.
19. Lamiable,A., Quessette,F., Vial,S., Barth,D. and Denise,A. (2013) An algorithmic game-theory approach for coarse-grain prediction of RNA 3D structure. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 193–199.
20. Mustoe,A.M., Al-Hashimi,H.M. and Brooks,C.L. 3rd (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.
21. 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.
22. Sulc,P., Romano,F., Ouldridge,T.E., Doye,J.P. and Louis,A.A. (2014) A nucleotide-level coarse-grained model of RNA. *J. Chem. Phys.*, **140**, 235102.
23. Cao,S. and Chen,S.J. (2005) Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA*, **11**, 1884–1897.
24. Pasquali,S. and Derreumaux,P. (2010) HiRE-RNA: A high resolution coarse-grained energy model for RNA. *J. Phys. Chem. B*, **114**, 11957–11966.
25. Xia,Z., Bell,D.R., Shi,Y. and Ren,P. (2013) RNA 3D structure prediction by using a coarse-grained model and experimental data. *J. Phys. Chem. B*, **117**, 3135–3144.

26. Bernauer,J., Huang,X., Sim,A.Y. and Levitt,M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
27. Denesyuk,N.A. and Thirumalai,D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B*, **117**, 4901–4911.
28. Olson,W.K. and Flory,P.J. (1972) Spatial configurations of polynucleotide chains. 3. Polydeoxyribonucleotides. *Biopolymers*, **11**, 57–66.
29. Metropolis,N. and Ulam,S. (1949) The Monte Carlo method. *J. Am. Stat. Assoc.*, **44**, 335–341.
30. Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
31. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
32. Zirbel,C.L., Sponer,J.E., Sponer,J., Stombaugh,J. and Leontis,N.B. (2009) Classification and energetics of the base-phosphate interactions in RNA. *Nucleic Acids Res.*, **37**, 4898–4918.
33. 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.
34. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
35. Duarte,C.M. and Pyle,A.M. (1998) Stepping through an RNA structure: A novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.
36. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
37. Shortle,D., Simons,K.T. and Baker,D. (1998) Clustering of low-energy conformations near the native structures of small proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11158–11162.
38. Seetin,M.G. and Mathews,D.H. (2011) Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints. *J. Comput. Chem.*, **32**, 2232–2244.
39. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
40. 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.
41. 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.
42. Walen,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.
43. Hajdin,C.E., Ding,F., Dokholyan,N.V. and Weeks,K.M. (2010) On the significance of an RNA tertiary structure prediction. *RNA*, **16**, 1340–1349.
44. Sharma,S., Ding,F. and Dokholyan,N.V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, **24**, 1951–1952.
45. Cruz,J.A. and Westhof,E. (2011) Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nat. Methods*, **8**, 513–521.
46. Zirbel,C.L., Roll,J., Sweeney,B.A., Petrov,A.I., Pirrung,M. and Leontis,N.B. (2015) Identifying novel sequence variants of RNA 3D motifs. *Nucleic Acids Res.*, **43**, 7504–7520.
47. Anfinsen,C.B. and Scheraga,H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.*, **29**, 205–300.
48. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
49. Schlick,T., Collepardo-Guevara,R., Halvorsen,L.A., Jung,S. and Xiao,X. (2011) Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys.*, **44**, 1–38.
50. Saunders,M.G. and Voth,G.A. (2013) Coarse-graining methods for computational biology. *Annu. Rev. Biophys.*, **42**, 73–93.
51. Rother,K., Rother,M., Boniecki,M., Puton,T., Tomala,K., Lukasz,P. and Bujnicki,J.M. (2012) In: Leontis,NB and Westhof,E (eds). *RNA 3D structure analysis and prediction*. Springer-Verlag, Berlin.
52. Matelska,D., Purta,E., Panek,S., Boniecki,M.J., Bujnicki,J.M. and Dunin-Horkawicz,S. (2013) S6:S18 ribosomal protein complex interacts with a structural motif present in its own mRNA. *RNA*, **19**, 1341–1348.
53. Dzananovic,E., Patel,T.R., Chojnowski,G., Boniecki,M.J., Deo,S., McEleney,K., Harding,S.E., Bujnicki,J.M. and McKenna,S.A. (2014) Solution conformation of adenovirus virus associated RNA-I and its interaction with PKR. *J. Struct. Biol.*, **185**, 48–57.
54. Flores,S.C., Wan,Y., Russell,R. and Altman,R.B. (2010) Predicting RNA structure by multiple template homology modeling. *Pac. Symp. Biocomput.*, 216–227.
55. Sim,A.Y., Levitt,M. and Minary,P. (2012) Modeling and design by hierarchical natural moves. *Proc Natl Acad Sci U S A*, **109**, 2890–2895.
56. Kim,N., Laing,C., Elmetwaly,S., Jung,S., Curuksu,J. and Schlick,T. (2014) Graph-based sampling for approximating global helical topologies of RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4079–4084.
57. Magnus,M., Matelska,D., Lach,G., Chojnowski,G., Boniecki,M.J., Purta,E., Dawson,W., Dunin-Horkawicz,S. and Bujnicki,J.M. (2014) Computational modeling of RNA 3D structures, with the aid of experimental restraints. *RNA Biol.*, **11**, 522–536.