J Mol Model (2014) 20:2389
DOI 10.1007/s00894-014-2389-z

ORIGINAL PAPER

De novo tertiary structure prediction using RNA123—benchmarking and application to Macugen

Emma S. E. Eriksson · Lokesh Joshi · Martin Billeter · Leif A. Eriksson

Received: 14 February 2014 / Accepted: 22 July 2014 / Published online: 10 August 2014

Abstract The present benchmarking study utilizes the RNA123 program for de novo prediction of tertiary structures of a set of 50 RNA molecules for which X-ray/NMR structures are available, based on the nucleic acid sequence only. All molecules contain a hairpin loop motif and a helical structure of canonical and non-canonical base pairs, interrupted by bulges and internal loops to various degrees. RNA molecules with double helices made up purely by canonical base pairing, and molecules containing symmetric internal loops of non-canonical base pairing are, overall, very well predicted. Structures containing bulges and asymmetric internal loops, and more complex structures containing multiple bulges and internal loops in the same molecule, result in larger deviations from their X-ray/NMR predicted structures due to higher degree of flexibility of the nucleotide bases in these regions. In a majority of the molecules included herein, the RNA123 program was, however, able to predict the tertiary structure with a heavy atom RMSD of less than 5 Å to the X-ray/NMR structure, and the models were in most cases structurally closer to the X-ray/NMR structures than models predicted by MC-Fold and MC-Sym. A set of RNA molecules containing pseudoknot tertiary structure motifs were included, but neither of the programs was able to predict the folding of the single-stranded stem onto the helix without additional structural input. The RNA123 program was then applied to predict the tertiary structure of the RNA segment of Macugen®, the first RNA aptamer approved for clinical use, and for which no tertiary structure has yet been solved. Four possible tertiary structures were predicted for this 27-nucleic-acid-long RNA molecule, which will be used in constructing a full model of the PEGylated aptamer and its interaction with the vascular endothelial growth factor target.

Keywords RNA · RNA123 · Tertiary structure prediction · Macugen

This paper belongs to Topical Collection 9th European Conference on Computational Chemistry (EuCo-CC9)

Electronic supplementary material The online version of this article (doi:10.1007/s00894-014-2389-z) contains supplementary material, which is available to authorized users.

E. S. E. Eriksson (✉) · M. Billeter · L. A. Eriksson
Department of Chemistry and Molecular Biology, University of Gothenburg, 405 30 Göteborg, Sweden
e-mail: emma.eriksson@chem.gu.se

L. Joshi
National Center for Biomedical Engineering Science, National University of Ireland, Galway, Ireland

Introduction

RNA

RNA is one of the most essential biological molecules in nature. Its complex structure is based on a sugar-phosphate backbone, π–π nucleobase stacking and hydrogen-bonded base pairing. RNA comprises two purine type bases, guanine (G) and adenine (A), and two pyrimidine type nucleotide bases, cytosine (C) and uracil (U), which, in different combinations and with different interactions, generate a specific fold of the backbone. The three-hydrogen-bonded G–C pair and the two-hydrogen-bonded A–U pair represent the canonical Watson-Crick pairs that result in an A-form double helix (Fig. 1a). The A-form helical structure that RNA adopts differs from the common B-form of DNA in having a narrower deep major groove and a wider shallow minor grove. The RNA fold is, however, also very dependent on the less stable non-canonical base pair interactions, i.e., mismatched hydrogen bonding between pairs other than G–C and A–U (Fig. 1b). Non-canonical base pairs deform the shape of the RNA helix, and can serve as recognition sites for interactions

with proteins and other molecules. The most common non-canonical base pair in ribosomal RNA (rRNA) [1] is the wobble G–U pair [2], which is found to be almost as thermodynamically stable as canonical base pairs [3–5] and constitutes recognition sites for cleavage in self-splicing introns [6, 7]. Extensive information and discovery of previously unknown non-canonical pairs have been achieved mainly through the solution structures of the large and small ribosomal subunits [8–10]. Bases do not only interact in pairs, but triples [11] and quadruples [12] also commonly occur.

The nucleotide bases can interact with each other through one of three edges of each base, named Watson-Crick edge, Hoogsteen edge, and sugar edge (nomenclature and classification of possible base pair interactions was introduced by Leontis and Westhof [13]). The Watson-Crick edges are those involving hydrogen bonding in the regular canonical Watson-Crick base pairs. An edge of one base can potentially interact with any edge of a second base, and this together with the two possibilities of the orientation (cis/trans) of the glycosidic bonds, generates in total 12 different possible edge–edge interactions.

The variation in base pair interactions and mismatches alter the fold of the RNA, and this generates various specific RNA motifs to be identified. Common RNA motifs include hairpin loops, bulges, internal loops, and pseudoknots (Fig. 1). Hairpin loops occur in all RNA molecules that comprise a single strand folded into a double helix, thus resulting in a turn region with a number of unpaired bases. Hairpin loops with four unpaired nucleotides (tetraloops) are found prevalently in rRNA [14], and GNRA (*N*=A, C, G, U; R=A, G) is the most commonly observed tetraloop in RNA [14, 15]. Bulges, sometimes referred to as bulge loops, are composed of one or several unpaired bases on one of the stems, and surrounded by canonical base pairs on both sides (Fig. 1c). Internal loops are herein defined as regions of non-interacting or non-canonically paired bases, either equally (symmetric internal loop; Fig. 1d,e) or unequally (asymmetric internal loop; Fig. 1f,g) distributed between the two stems. Symmetric internal loops with non-canonical base pair interactions (Fig. 1d) thus constitute a continuation of the double helix structure, yet distorted to various degrees (cf. [16–19]). According to the same categorization we define an asymmetric loop as either an internal loop of non-canonical base pairs with one or multiple bulged nucleotides (Fig. 1f), or a more disturbed structure with non-interacting nucleotides on both stems (Fig. 1g). Asymmetric internal loops that cause sharp turns in the helical structure are important structural motifs in, for example, translation, and involve specific motifs such as the kink-turn [20] and hook-turn [21]. Another type of RNA structure contains two hairpin loops and the 5′and 3′ends of the sequence are subsequently located in the center of the helix (Fig. 1h).

The tertiary structure of RNA is determined by interaction between distant regions of motifs containing unpaired bases, and this thus determines the overall conformation of the molecule. Pseudoknots (Fig. 1i) are common tertiary structure motifs and are formed by hydrogen bond interactions between bases in a hairpin loop of a double helix with bases in a single stranded region that folds upon the first helix. This results in a complex and highly stable two-helical structure. It was first identified in the turnip yellow mosaic virus [22], and it has since been found that many viruses use pseudoknots to induce ribosomal frameshifting in order to alter gene expression [23].

RNA structure prediction

The complex three-dimensional (3D) structure of RNA is difficult and time-consuming to determine using X-ray crystallography or nuclear magnetic resonance (NMR). A majority of structures in the Protein Data Bank (PDB) contain proteins, whereas only a very small fraction of the structures contain RNA alone, many of which are also redundant. However, the number of extracted RNA sequences surpasses the number of available tertiary structures by far. This motivates the application of advanced computer-based techniques to predict structures of the increasing number of sequences that are extracted experimentally.

Theoretical prediction and evaluation of RNA structures is, in general, fast and can be applied to a large set of sequences in a short amount of time, and can hence be an efficient alternative to the more expensive and time-consuming experimental techniques. Prediction of RNA tertiary structures is, however, a challenging task, and shares many similarities to protein folding prediction, for which methods have been extensively developed during the past decades; however, for RNA the development has not reached as far. The number of possible conformations of an RNA molecule is ~3^{7N}, where N is the number of nucleotides, 7 is the total number of backbone and base dihedrals per residue, and 3 is the assumed number of conformational minima for each dihedral angle. It should, however, be noted that this number also includes impossible conformations in which steric overlap would occur. Hence, we herein define a conformation as any possible geometrical structure. The backbone of each RNA nucleotide can adopt six different dihedral angles, compared to two for each amino acid in proteins. This gives 3^{6N} possible backbone conformations, a number that is reduced to 3^{3N} under the assumption that half of the residues adopt an A-form geometry. This still makes prediction of RNA structures difficult, in particular when considering molecules with a large number of residues. However, RNA inherently carries some features that make its structure prediction easier than the protein folding problem. As discussed above, RNA contains only four different nucleotides, which in addition are fairly similar in structure. This can be compared to the 20 possible amino acids in proteins,

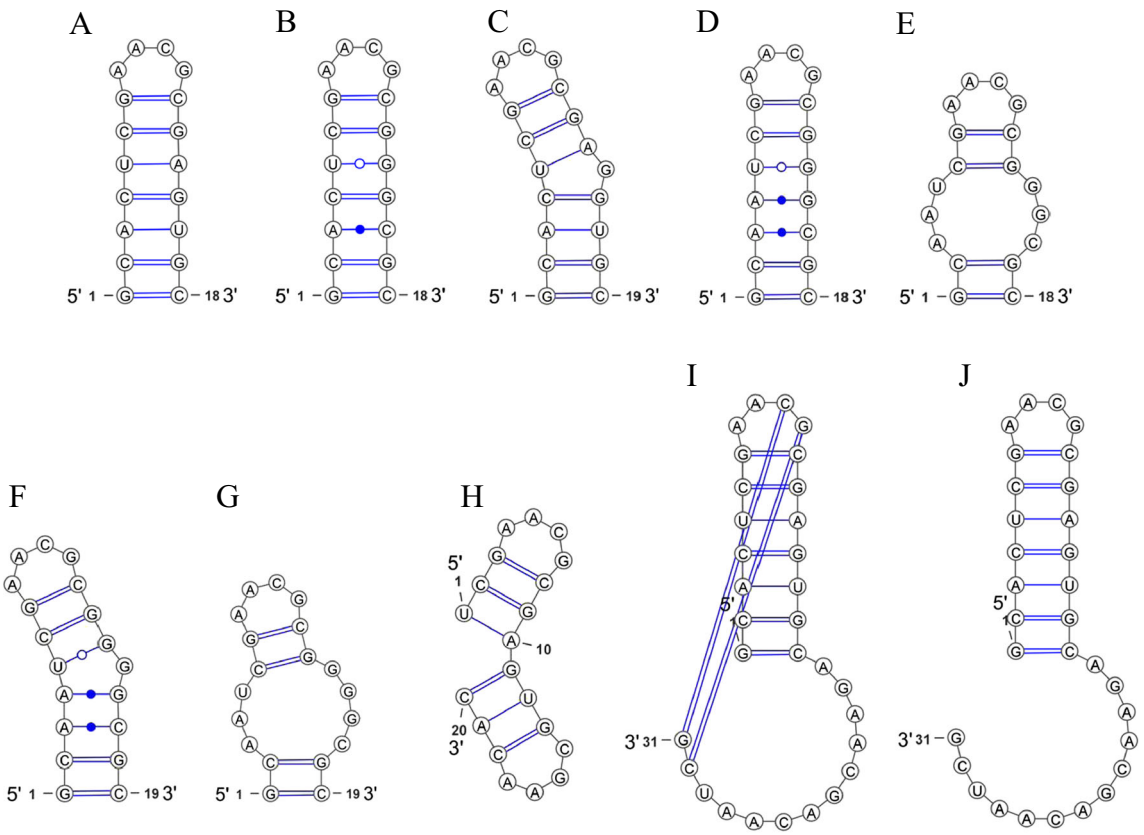


Fig. 1 Schematic drawings of secondary and tertiary structure motifs in RNA, represented by a hairpin structure with **a** purely canonical base pair interactions; **b** canonical and non-canonical base pair interactions; **c** canonical base pair interactions and a non-interacting base in a bulge; **d** a symmetric internal loop of non-canonical base pair interactions; **e** a symmetric internal loop of non-interacting bases; **f** an asymmetric internal loop of non-canonical base pair interactions and a bulge; **g** an asymmetric internal loop of non-interacting bases; **h** a double hairpin structure; **i** a pseudoknot; and **j** erroneously predicted structure by RNA123, missing the pseudoknot interaction. Non-canonical base pair interactions are indicated with an open (wobble G–U pair) or closed circle. Figures created with VARNA software [34]

which have much more varying properties and sizes. RNA folding is also simplified by the hydrogen bonded base pairs that largely determine the secondary structure of the RNA, and result in formation of rigid double helices.

Development of algorithms for predicting the two-dimensional fold of RNA has been progressing for several years [24]; however, techniques for determining the more complex tertiary folds have not been developed as extensively. Predicting the tertiary fold can be performed by including various amounts of experimental data. Naturally, the most challenging prediction is the one in which only the nucleic acid sequence is used as input and no additional data is available (de novo structure prediction).

Among the programs available for tertiary structure prediction, MC-Fold and MC-Sym [25], Assemble [26], RNA2D3D [27], and FARFAR [28] can be mentioned. Classical molecular dynamics (MD) simulations can rarely find the correct tertiary structure from a random starting structure due to insufficient time to search the entire conformational space; however, coarse-grained MD simulations, such as those applied in programs such as YUP [29], iFoldRNA [30, 31], which employs discrete molecular dynamics (DMD) simulations, and NAST [32] have been found to be useful. RNA123 [33] is a fragment and motif de novo assembly program that can also perform homology modeling by using experimental input of a related structure. RNA123 applies a nucleic acid force field (NA_FF) optimized for RNA, and allows for modeling of systems that are initially far from the native geometry.

We herein present a benchmarking study of the de novo algorithm in the RNA123 program where a set of 50 RNA structures are predicted and compared with the corresponding X-ray or NMR structures available in the PDB. The selected structures are 10–70 nucleotides in length, and include parts extracted from, e.g., ribozymes, rRNA, telomerase, and viral pseudoknots.

We then move on to predicting the tertiary structure of the RNA fragment of pegaptanib (Macugen®)—an RNA aptamer that is in clinical use to treat age-related macular degeneration, but for which no tertiary structure has yet been determined experimentally.

Table 1 summarizes the secondary and tertiary structure motifs found in the RNA structures included in the present benchmarking study, based on the X-ray/NMR structures.

Table 1 Summary of secondary and tertiary structure motifs in the X-ray and nuclear magnetic resonance (NMR) structures of the RNA structures included in the current benchmarking study

Secondary/tertiary structure motif	Number of structures	Number of nucleotides in molecule
Hairpin loop (all)	50	10–70
Symmetric internal loop	10	20–55
Asymmetric internal loop	10	27–70
Bulge	14	19–70
Pseudoknot	7	26–47

Two or more continuous non-canonical base pairs are considered an internal loop, whereas a single mismatched base pair surrounded by canonical base pairs is not; see definitions above and Fig. 1. Mismatched bases at the 5′ and 3′ ends are not accounted for. The tertiary structure pseudoknot elements are also listed. Pseudoknot structures are only counted to contain hairpin loops.

Methods

RNA123 [33] (version 1.1) was used to predict secondary and tertiary structures of the 50 RNA molecules in the benchmarking set. The nucleic acid sequence was used to initially predict an optimal and a maximum of 20 suboptimal secondary structures. Tertiary structures were built using the optimal predicted secondary structure.

Secondary structure prediction in RNA123 is carried out by applying a dynamic programming algorithm (DPA), which searches for global minimum and suboptimal structures—a method used in several programs. The prediction is performed at 310 K, with sodium concentration of 0.1 M and magnesium concentration of 0.001 M, and generates secondary structures corresponding to optimal and suboptimal folds. RNA123 further uses a motif library of conditioned structures extracted from the PDB for tertiary structure prediction. Structure conditioning is required in order to detect and correct errors in X-ray and NMR structures, which are otherwise transferred to the predicted models. To date, there are 453 conditioned RNA structures present in the RNA123 library, from which motifs are identified and extracted during the modeling procedure. It is thus important to note that the current motif library is not complete and hence does not contain all possible motifs present in native RNAs. The secondary structure serves as a base for the BUILDER algorithm that assembles matching motifs from the database, energy minimizes all possible conformers with a discrete sampling of torsion angles (DSTA) algorithm, and finally ranks the conformers based on the RNA123 force field energies. The conformer with the lowest NA_FF energy is used to build the final tertiary structure corresponding to this secondary structure, and is then energy minimized with the DSTA algorithm [33].

The phosphate group in the first residue of each predicted structure was replaced by a hydroxyl group, as this is lacking in a majority of the X-ray/NMR structures. Additional energy minimization and MD simulations were performed with the YASARA [35] program (version 10.7.20). The RNA molecules were inserted in boxes that were extended 10 Å in all directions from the molecule, and then made cubic by the length of the longest axis. The boxes were filled with water and neutralized with NaCl to a concentration of 0.9 %. The pH was set to 7.4 and the temperature to 298 K. A short MD simulation was run for the solvent, during which the density was adjusted to 0.997 g/L. The total system was then energy-minimized with the AMBER03 [36] force field, using a 7.86 Å force cutoff and the particle mesh Ewald (PME) algorithm [37] to treat long-range electrostatic interactions, since periodic boundary conditions were applied. After removal of conformational stress by a short steepest descent minimization, the procedure continued by simulated annealing (time step 2 fs, atom velocities scaled down by 0.9 every 10th step) until convergence was reached, i.e., the energy improved by less than 0.05 kJ mol^{-1} per atom during 200 steps. MD simulations were performed for 500 ps and a final energy minimization was performed after completion. A selection of RNA models that were hard to correctly predict was also simulated for 100 ns in order to determine if an extended simulation could adjust the structures.

The predicted tertiary structures were superposed with the X-ray/NMR structures after the initial prediction, and again after energy minimization and after MD simulation. If more than one structure was available in the NMR file, the predicted model was superposed with all structures, and the lowest root mean square deviation (RMSD) reported (see table below).

In order to determine whether the predicted optimal secondary structure represents the native fold of the RNA molecule, tertiary structures were also built based on the second and third predicted secondary structures, and superposed with the X-ray/NMR structure initially and after energy minimization. In a few cases, however, no more than one or two secondary structures were possible to predict.

Tertiary structures of a set of RNA structures in the benchmarking set was also predicted using the MC-Fold and MC-Sym programs [25] for comparison of the performance of the RNA123 program. MC-Fold predicts RNA secondary structures using free energy minimization that takes into account also non-canonical base-pairs, as opposed to several common prediction tools. An energy function based on nucleotide cyclic motifs extracted from the PDB is applied. A maximum of 20 secondary structures were predicted, and the energetically most optimal predicted structure was used as

input to the tertiary structure prediction by MC-Sym. MC-Sym applies a process of fusing nucleotide cyclic motifs from the PDB in the generation of tertiary structures. A total of 1,000 tertiary structures was generated and roughly refined in a steepest descent energy minimization with the AMBER99 force field and a root-mean square gradient of 100 kcal mol^{-1} $Å^{-1}$. The refined structures were scored based on the built-in scoring function, and the structure with lowest energy was extracted and superposed on the X-ray/NMR structure. As our aim was to evaluate the performance of the programs to predict RNA structures based on the nucleic acid sequence only; we do not apply the option of manual structural input or constraints. No pseudoknots were predicted with MC-Fold and MC-Sym as this requires structural input data. In addition, we performed a more thorough analysis of the MC-Sym tertiary structures generated for a number of the RNA structures in the benchmarking set. This analysis takes into account base entropy, radius of gyration, score and a knowledge-based P-score, and selected models were clustered by structural similarity. The model with the lowest score value in the largest cluster was selected to represent the best model, and superposed on the X-ray/NMR structure.

Results and discussion

Benchmarking study

The benchmarking study was based on the difference in structure between the predicted models and the X-ray/NMR structures available in the PDB. A word of caution regarding the difference in X-ray and solution structure determination is advised here. A majority of the RNA structures available have been determined by NMR, although this technique is difficult to apply to RNA due to low proton density as compared to proteins. Improvements in NMR structure determination include, for example, residual dipolar couplings, which are discussed below.

The NMR structure files used in the present study are furthermore composed of sets of structures, from 1 to 32 separate structures. The presence of multiple structures—commonly a set of lowest energy structures—is a problem when a predicted model is to be compared thereto, as they result in different RMSDs to the predicted model. The individual NMR structures can differ significantly and result in RMSD differences of several Ångströms when the predicted model is superposed with these. The reported heavy atom RMSDs herein represent the lowest values obtained when the initially predicted models were superposed with the NMR structures. RMSDs of these same NMR structures after energy minimization and MD simulation are also reported. Overall, the NMR structures with the lowest initial RMSDs to the predicted models also gave the lowest RMSDs after energy minimization. It should also be pointed out that some of the NMR predictions represent averaged structures. In such cases, this is indicated specifically in the tables, and is discussed further below.

The aim of the present study was to evaluate the performance of the de novo prediction program RNA123, and we begin by comparing tertiary structures predicted by this program to structures predicted by MC-Fold and MC-Sym. In Table 2 we compare heavy atom RMSDs of RNA123 predicted initial tertiary structures to corresponding structures predicted by MC-Sym. For both programs, the energetically most optimal secondary structure was used to build tertiary structures, and the lowest energy (based on each program's scoring function) tertiary structure was selected. In a clear majority of cases, the tertiary structure predicted by RNA123 showed a lower or equal RMSD to the X-ray/NMR structure compared to the tertiary structure predicted by MC-Sym. For structures that were better predicted by MC-Sym, the RMSD difference to the RNA123 predicted structure was at most 2.9 Å. These results indicate that the RNA123 program overall predicts the RNA tertiary structures in this benchmarking set to a higher accuracy than MC-Fold and MC-Sym, when only the nucleic acid sequence is used as input. It should be stressed that the MC-Sym tertiary structures most likely could be improved by manual structural input; however, the aim of the present study was to evaluate the de novo performance of programs that can be used to study RNA molecules for which no structural data is available.

A more thorough analysis of the tertiary structures generated by MC-Sym was made by taking into account base entropy, radius of gyration, score and a knowledge-based P-score. In the selected set of structures for which the analysis was performed we did not observe any overall improvement in the structures extracted by this analysis when superposed with the X-ray/NMR structure. In most cases, the extended analysis generated RMSD values in the range of±0.4 Å from those obtained for the structure with the lowest score, and in a few extreme cases we found that the RMSD was increased or decreased by as much as 5 Å. Clearly, no consistency in either an increase or decrease in RMSD was noted when the extended analysis was applied, hence we present the data for only the lowest scored structures in Table 2.

As we concluded that the RNA123 program performed overall better in the de novo prediction, the remaining part of this paper deals only with the evaluation of this program. Heavy atom RMSDs between the RNA123 tertiary structures and the corresponding X-ray/NMR structures are displayed in Table 3, together with RMSDs after energy minimization and after MD simulation.

In the cases where the predicted model is significantly different from the X-ray/NMR structure—commonly for pseudoknots (and one other structure; PDB ID: 1R7Z)—a set of residues was always correctly predicted (cf. helix in

Table 2 Heavy atom root mean square deviation (RMSD) of predicted RNA models by RNA123 and MC-Fold/MC-Sym based on the most optimal predicted secondary structure to X-ray/NMR structure[a]

PDB ID	Number of nucleotides	X-ray/NMR (total number of structures)	RMSD RNA123; Å [b]	RMSD MC-Fold/MC-Sym; Å [b]
1A51	41	NMR (9)	3.4 (1)	3.4 (6)
1AFX	12	NMR (13)	2.6 (13)	3.3 (11)
1ANR	29	NMR (20)	5.2 (10)	6.0 (5)
1BVJ	23	NMR (21)	3.2 (6)	3.7 (18)
1CQ5[c]	43	NMR (1)	4.2 (1)	17.7[e] (1)
1CQL	43	NMR (10)	4.2 (2)	17.5[e] (6)
1E4P	24	NMR (20)	2.0 (13)	2.2 (14)
1EBQ	29	NMR (5)	3.3 (4)	4.0 (4)
1ESY	19	NMR (20)	4.9 (12)	3.4 (8)
1F6X[c]	27	NMR (1)	2.5 (1)	2.9 (1)
1F84	29	NMR (25)	2.4 (1)	3.3 (17)
1HS1[c]	13	NMR (1)	2.3 (1)	3.0 (1)
1HS8[c]	13	NMR (1)	0.8 (1)	2.9 (1)
1HWQ	30	NMR (20)	3.1 (13)	3.4 (4)
1JOX[d]	21	NMR (24)	3.9 (3)	3.4 (3)
1JP0[d]	21	NMR (28)	3.8 (13)	3.8 (18)
1K2G[c]	22	NMR (1)	3.8 (1)	3.8 (1)
1KXK	70	X-ray (1)	6.0 (1)	9.1 (1)
1L1W[c]	29	NMR (1)	3.4 (1)	3.4 (1)
1MFY	31	NMR (16)	3.5 (9)	4.6 (14)
1MNX	42	NMR (13)	5.1 (11)	3.3 (5)
1MT4	24	NMR (17)	5.5 (12)	3.0 (13)
1P5M	55	NMR (22)	3.8 (9)	7.4 (20)
1Q93	27	X-ray (1)	1.9 (1)	2.1 (1)
1R2P	34	NMR (10)	6.7 (3)	3.8 (8)
1TBK	17	NMR (11)	4.2 (8)	3.4 (6)
1TXS	38	NMR (20)	4.0 (1)	8.3 (20)
1U3K	38	NMR (10)	7.3 (8)	8.0 (7)
1ZC5	41	NMR (20)	2.3 (18)	4.3 (13)
2F88	34	NMR (10)	2.5 (1)	3.6 (8)
2U2A	20	NMR (1)	3.2 (1)	2.4 (1)
3PHP	23	NMR (10)	2.9 (3)	3.5 (8)
17RA	21	NMR (12)	1.9 (12)	2.9 (11)
361D	20	X-ray (1)	3.4 (1)	3.7 (1)
480D	27	X-ray (1)	1.2 (1)	6.5 (1)

[a] In cases where more than one structure is determined by NMR, the lowest RMSD is listed

[b] Numbers in parenthesis indicate which X-ray/NMR structure the RMSD refers to

[c] NMR averaged structure

[d] 1JP0 and 1JOX correspond to the same RNA molecule with the difference that the NMR predicted structure 1JOX was refined with residual dipolar couplings

[e] All tertiary structures generated by MC-Sym for this RNA molecule were extremely different from the NMR structures, and none of the 1,000 generated structures showed a RMSD value of less than 16.0 Å to the NMR structures

Fig. 1j) and RMSDs are reported for both this set of residues (specified in tables below) as well as for the full system. For pseudoknots, the initial hairpin and helix is generally well reproduced, whereas the back-folding of the additional single strand over the hairpin poses considerable problems.

Table 4 shows heavy atom RMSDs for the tertiary structures of the three energetically most optimal secondary structures after energy minimization compared to the corresponding X-ray/NMR structure. Corresponding initial RMSDs of the predicted models are included in the Supporting Information.

Table 3 Heavy atom RMSD of predicted RNA models based on the most optimal predicted secondary structure to X-ray/NMR structure[a]

PDB ID	Number of nucleotides	X-ray/NMR (total number of structures)	Initial RMSD; Å [b]	RMSD after energy minimization; Å	RMSD after 500 ps MD simulation; Å	RMSD after 100 ns MD simulation; Å
1A51	41	NMR (9)	3.4 (1)	3.5	4.1	
1AFX	12	NMR (13)	2.6 (13)	2.6	2.9	
1ANR	29	NMR (20)	5.2 (10)	5.2	6.0	5.8
1ATV	17	NMR (4)	2.1 (3)	2.1	2.6	
1BN0	20	NMR (11)	2.9 (4)	2.9	3.8	
1BVJ	23	NMR (21)	3.2 (6)	3.2	2.6	4.3
1CQ5[c]	43	NMR (1)	4.2 (1)	4.2	4.0	4.5
1CQL	43	NMR (10)	4.2 (2)	4.1	3.9	5.1
1E4P	24	NMR (20)	2.0 (13)	2.0	3.6	
1E95	36	NMR (15)	19.6 (14)	19.7	19.8	
	Res 1-17		1.7 (14)	1.8	2.1	
1EBQ	29	NMR (5)	3.3 (4)	3.3	3.6	
1ESH[c]	13	NMR (1)	1.5 (1)	1.5	1.8	
1ESY	19	NMR (20)	4.9 (12)	5.0	4.6	5.8
1F6X[c]	27	NMR (1)	2.5 (1)	2.6	3.3	
1F84	29	NMR (25)	2.4 (1)	2.5	3.3	
1HS1[c]	13	NMR (1)	2.3 (1)	2.4	2.7	6.1
1HS8[c]	13	NMR (1)	0.8 (1)	1.0	1.7	
1HWQ	30	NMR (20)	3.1 (13)	3.1	3.0	
1IDV	10	NMR (10)	3.2 (1)	3.2	3.2	
1JOX[d]	21	NMR (24)	3.9 (3)	3.9	3.8	
1JP0[d]	21	NMR (28)	3.8 (13)	3.9	4.1	3.8
1K2G[c]	22	NMR (1)	3.8 (1)	3.8	2.8	4.4
1KAJ	32	NMR (1)	26.1 (1)	26.4	28.6	
	Res 1–20		4.9 (1)	5.0	5.2	
1KKA	17	NMR (8)	6.6 (2)	6.6	5.6	
1KPD[c]	32	NMR (1)	26.9 (1)	27.1	29.6	
	Res 1–19		6.3 (1)	6.3	6.3	
1KXK	70	X-ray (1)	6.0 (1)	6.0	6.6	
1L1W[c]	29	NMR (1)	3.4 (1)	3.3	3.0	
1MFY	31	NMR (16)	3.5 (9)	3.4	4.0	
1MNX	42	NMR (13)	5.1 (11)	5.1	3.9	
1MT4	24	NMR (17)	5.5 (12)	5.5	5.0	
1OQ0	15	NMR (20)	2.7 (12)	2.7	3.1	
1P5M	55	NMR (22)	3.8 (9)	3.7	4.3	4.0
1Q93	27	X-ray (1)	1.9 (1)	1.9	1.7	
1R2P	34	NMR (10)	6.7 (3)	6.7	7.0	6.7
1R7Z	34	NMR (20)	18.3 (19)	18.4	17.4	
	Res 14-27		3.2 (19)	3.2	3.3	
1TBK	17	NMR (11)	4.2 (8)	4.2	4.6	
1TXS	38	NMR (20)	4.0 (1)	4.1	4.0	
1U3K	38	NMR (10)	7.3 (8)	7.3	4.7	6.1
1UUU	19	NMR (15)	3.4 (1)	3.3	3.8	
1VOP	13	NMR (32)	3.9 (18)	3.9	4.1	
1XSG[c]	27	NMR (1)	3.9 (1)	3.9	3.2	
1YMO	47	NMR (20)	27.8 (18)	27.9	28.2	
	Res 15–47		1.9 (18)	2.0	2.1	
1ZC5	41	NMR (20)	2.3 (18)	2.3	4.1	

Table 3 (continued)

PDB ID	Number of nucleotides	X-ray/NMR (total number of structures)	Initial RMSD; Å [b]	RMSD after energy minimization; Å	RMSD after 500 ps MD simulation; Å	RMSD after 100 ns MD simulation; Å
2A43	26	X-ray (1)	17.0 (1)	17.1	19.6	
	Res 3–17		0.9 (1)	1.1	1.6	
2F88	34	NMR (10)	2.5 (1)	2.5	3.7	
2U2A	20	NMR (1)	3.2 (1)	3.2	3.4	
3PHP	23	NMR (10)	2.9 (3)	3.0	3.7	
17RA	21	NMR (12)	1.9 (12)	1.9	2.2	
361D	20	X-ray (1)	3.4 (1)	3.4	3.4	
387D[e]	26	X-ray (1)	20.6 (1)	21.0	21.6	
	Res 1–16		1.3 (1)	1.4	2.0	
437D	28 (27)[f]	X-ray (1)	22.3 (1)	22.3	23.0	
	Res 2–18		1.5 (1)	1.5	2.4	
480D	27	X-ray (1)	1.2 (1)	1.1	1.4	

[a] In cases where more than one structure is determined by NMR, the lowest RMSD is listed

[b] Numbers in parenthesis indicate which X-ray/NMR structure the RMSD refers to

[c] NMR averaged structure

[d] 1JP0 and 1JOX correspond to the same RNA molecule with the difference that the NMR predicted structure 1JOX was refined with residual dipolar couplings

[e] A sub-optimal secondary structure was used to build the tertiary structure since the optimal energy structure did not predict the non-hydrogen-bonding tail as the 3′ end of the sequence (cf. Fig. 1j)

[f] As the X-ray structure contains only residues 2–28, the first residue in the predicted structure was removed

In cases where a suboptimal secondary structure results in a predicted tertiary structure that has a completely different fold compared to the native structure, thus making it impossible to find common structural features to superpose, RMSDs are not reported and are instead indicated with '*' in Table 4.

Energy minimization generally has a small total effect on the RMSD to the X-ray/NMR structures in the benchmarking set; in some cases it lowers the RMSD and in other cases increases it. MD simulation, however, increases the RMSD for a majority of the structures. After the extended 100 ns simulations, the structures were in general not significantly improved. An explanation for this observation is that the AMBER03 force field is not parameterized specifically for RNA. In addition, hairpin RNA structures are highly flexible, in particular the end region, and this can subsequently be illustrated by an elevated RMSD in the final snapshot of the MD simulation. The reported RMSD is generated after superposition to the same NMR structure that gave the lowest RMSD to the initially predicted structure.

Canonical base pairs and hairpin loops

All RNA structures included in the present study contain hairpin structural elements, as a result of folding of the single-stranded RNA sequence into a double helix structure. A characteristic feature of a hairpin loop is its unpaired nucleotides. In the present benchmarking study, the RNA structures contain hairpin loops with three to six nucleotides. A majority of the hairpin loops are tetraloops, and ten structures contain the frequently occurring GNRA tetraloop. The nucleotides in the hairpin loop are flexible and more likely to move compared to the base-paired ones. This results in problems in determining the location of these nucleotides in space, using both experimental and computational methods. The nucleotides in the hairpin loop therefore contribute disproportionally much to the RMSDs; this feature is clearly seen in the RMSD per residue graphs displayed in Fig. 2.

RNA structures that contain only a helix with purely canonical base pairs and a hairpin loop, and free of bulges or internal loops, are overall very well predicted by RNA123. There are, in total, ten such RNA structures included in this benchmarking study. One additional molecule containing a G–U base pair in the middle of the sequence (PDB ID: 1VOP), and four molecules with a single mismatched base pair prior to the hairpin loop (PDB ID: 1HS1, 1HS8, 1MT4, 1OQ0) were also included in this group of structures, as these structures do not contain any additional internal loops or bulges. The group considered here thus contains in total 15 structures, made up of 10–24 nucleotides. Larger RNA molecules generally contain bulges and internal loops to a higher degree. In a small number of the structures, the helix starts with a single non-interacting nucleotide on one of the two stems. The heavy atom RMSD for the structures ranges between 1.0 and 6.6 Å after energy minimization for the tertiary

Table 4 Heavy atom RMSD of predicted RNA models based on the three most optimal predicted secondary structures after energy minimization to X-ray/NMR structure[a]

PDB ID	Number of nucleotides	X-ray/NMR (total number of structures)	1 RMSD; Å [b]	2 RMSD; Å [b]	3 RMSD; Å [b]
1A51	41	NMR (9)	3.5 (1)	3.7 (2)	5.6 (4)
1AFX	12	NMR (13)	2.6 (13)	4.9 (2)	4.3 (2)
1ANR	29	NMR (20)	5.2 (10)	4.8 (20)	5.3 (19)
1ATV	17	NMR (4)	2.1 (3)	–	–
1BN0	20	NMR (11)	2.9 (4)	4.5 (4)	–
1BVJ	23	NMR (21)	3.2 (6)	4.0 (15)	4.2 (18)
1CQ5[c]	43	NMR (1)	4.2 (1)	5.9 (1)	6.7 (1)
1CQL	43	NMR (10)	4.1 (2)	5.5 (7)	5.9 (3)
1E4P	24	NMR (20)	2.0 (13)	3.9 (13)	3.2 (11)
1E95	36	NMR (15)	19.7 (14)	35.3 (13)	20.8 (14)
	Res 1–17		1.8 (14)	1.7 (13)	5.5 (14)
1EBQ	29	NMR (5)	3.3 (4)	3.2 (1)	4.0 (5)
1ESH[c]	13	NMR (1)	1.5 (1)	–	–
1ESY	19	NMR (20)	5.0 (12)	4.9 (9)	13.9 (8)
1F6X[c]	27	NMR (1)	2.6 (1)	2.9 (1)	5.2 (1)
1F84	29	NMR (25)	2.5 (1)	5.2 (10)	2.8 (1)
1HS1[c]	13	NMR (1)	2.4 (1)	5.3 (1)	2.4 (1)
1HS8[c]	13	NMR (1)	1.0 (1)	5.2 (1)	7.0 (1)
1HWQ	30	NMR (20)	3.1 (13)	4.9 (13)	3.5 (4)
1IDV	10	NMR (10)	3.2 (1)	2.7 (10)	4.9 (7)
1JOX[d]	21	NMR (24)	3.9 (3)	4.4 (3)	4.3 (3)
1JP0[d]	21	NMR (28)	3.9 (13)	4.2 (13)	3.6 (13)
1K2G[c]	22	NMR (1)	3.8 (1)	10.3 (1)	13.9 (1)
1KAJ	32	NMR (1)	26.4 (1)	—[g]	—[g]
	Res 1–20		5.0		
1KKA	17	NMR (8)	6.6 (2)	9.1 (1)	6.1 (1)
1KPD[c]	32	NMR (1)	27.1 (1)	—[g]	—[g]
	Res 1–19		6.3		
1KXK	70	X-ray (1)	6.0 (1)	7.5 (1)	7.7 (1)
1L1W[c]	29	NMR (1)	3.3 (1)	2.6 (1)	5.2 (1)
1MFY	31	NMR (16)	3.4 (9)	3.8 (9)	4.0 (14)
1MNX	42	NMR (13)	5.1 (11)	5.0 (5)	5.2 (11)
1MT4	24	NMR (17)	5.5 (12)	5.2 (13)	4.2 (13)
1OQ0	15	NMR (20)	2.7 (12)	2.7 (12)	5.0 (20)
1P5M	55	NMR (22)	3.7 (9)	3.5 (9)	13.6 (13)
1Q93	27	X-ray (1)	1.9 (1)	8.6 (1)	13.0 (1)
1R2P	34	NMR (10)	6.7 (3)	4.0 (8)	5.9 (8)
1R7Z	34	NMR (20)	18.4 (19)	20.3 (19)	12.3 (19)
	Res 14–27		3.2 (19)	3.2 (19)	3.2 (19)
1TBK	17	NMR (11)	4.2 (8)	3.2 (5)	–
1TXS	38	NMR (20)	4.1 (1)	3.9 (1)	9.1 (2)
1U3K	38	NMR (10)	7.3 (8)	9.7 (8)	7.9 (4)
1UUU	19	NMR (15)	3.3 (1)	–	–
1VOP	13	NMR (32)	3.9 (18)	4.5 (21)	8.2 (21)
1XSG[c]	27	NMR (1)	3.9 (1)	4.1 (1)	3.7 (1)
1YMO	47	NMR (20)	27.9 (18)	—[g]	23.2 (18)
	Res 15–47		2.0 (18)		1.9 (18)

Table 4 (continued)

PDB ID	Number of nucleotides	X-ray/NMR (total number of structures)	1 RMSD; Å [b]	2 RMSD; Å [b]	3 RMSD; Å [b]
1ZC5	41	NMR (20)	2.2 (18)	8.0 (3)	2.2 (18)
2A43	26	X-ray (1)	17.1 (1)	17.2 (1)	23.5 (1)
	Res 3–17		1.1	1.1	4.4
2F88	34	NMR (10)	2.5 (1)	3.1 (10)	3.3 (3)
2U2A	20	NMR (1)	3.2 (1)	3.9 (1)	–
3PHP	23	NMR (10)	3.0 (3)	4.7 (3)	2.3 (3)
17RA	21	NMR (12)	1.9 (1)	2.3 (8)	5.1 (9)
361D	20	X-ray (1)	3.4 (1)	4.0 (1)	3.5 (1)
387D	26	X-ray (1)	21.0[e] (1)	—[g]	–[g]
	Res 1–16		1.4		
437D	28 (27)[f]	X-ray (1)	22.3 (1)	23.0 (1)	22.1 (1)
	Res 2–18		1.5	2.1	4.4
480D	27	X-ray (1)	1.1 (1)	2.3 (1)	1.1 (1)

[a] In cases where more than one structure is determined by NMR, the lowest RMSD is listed

[b] Numbers in parenthesis indicate which X-ray/NMR structure the RMSD refers to

[c] NMR averaged structure

[d] 1JP0 and 1JOX correspond to the same RNA molecule with the difference that the NMR predicted structure 1JOX was refined with residual dipolar couplings

[e] A sub-optimal secondary structure was used to build the tertiary structure since the optimal energy structure did not predict the non-hydrogen-bonding tail as the 3′ end of the sequence (cf. Fig. 1j)

[f] As the X-ray structure contains only residues 2–28, the first residue in the predicted structure was removed

[g] Predicted tertiary structure too different from the native fold to superpose

structures predicted from the most optimal secondary structures. The largest RMSD is observed for a structure that is predicted to have a significantly different fold of the helix compared to the NMR structure (PDB ID: 1KKA). However, a majority of the predicted structures in this group show RMSDs below 3.5 Å after energy minimization.

An example of an RNA molecule containing a short helical structure with canonical base pairs, a U–U mismatched base pair prior to the hairpin loop, and a tetraloop, is displayed in Fig. 2a (PDB ID: 1HS8). The largest RMSD to the X-ray/NMR structures are generally seen in the hairpin region due to the non-base-paired nucleotides, and the end parts of the helix can also be a problem to predict due to possible flexibility of these nucleotides. However, in some cases, such as with the short RNA molecule shown in Fig. 2a, the predicted model shows an almost perfect match to the NMR structure, also in the hairpin region.

We next compared the tertiary structure built from the optimal secondary structure with the tertiary structures based on the next two suboptimal secondary structures (Table 4). It was found that in 7 out of 12 structures in this group, for which one or more suboptimal secondary structures were identified by RNA123, the RMSD to the X-ray/NMR structure is lowest for the tertiary structures corresponding to the predicted optimal secondary structure. In cases where the RMSD for any of the suboptimal structures is lower, the difference to the optimal structure is at most 1.3 Å.

Bulges and internal loops

Twenty-eight RNA molecules in the benchmarking set contain bulges and internal loops. Some structures contain either bulges or internal loops, whereas other structures contain both or multiples of each. Many of the molecules also contain single mismatched base pairs surrounded by canonical base pairs; however, these are not defined herein as internal loops, as discussed in the Introduction. The RNA molecules in this group contain 19–70 nucleotides, and the tertiary structures predicted from the most optimal secondary structures display heavy atom RMSDs to the X-ray/NMR structures in the range of 1.2–7.3 Å after energy minimization.

RNA molecules containing symmetric internal loops made up of multiple serial non-canonical base pairs (Fig. 1d) often form helix structures highly similar to those with pure canonical base pairs. Depending on the size of the loop, the base types, and the type of edge-to-edge interactions between the bases, the helix structure can be more or less distorted, often affecting the structure of the deep and shallow grooves. The non-canonical base pairs are often well predicted in the models, although differences do occur. As these base pairs

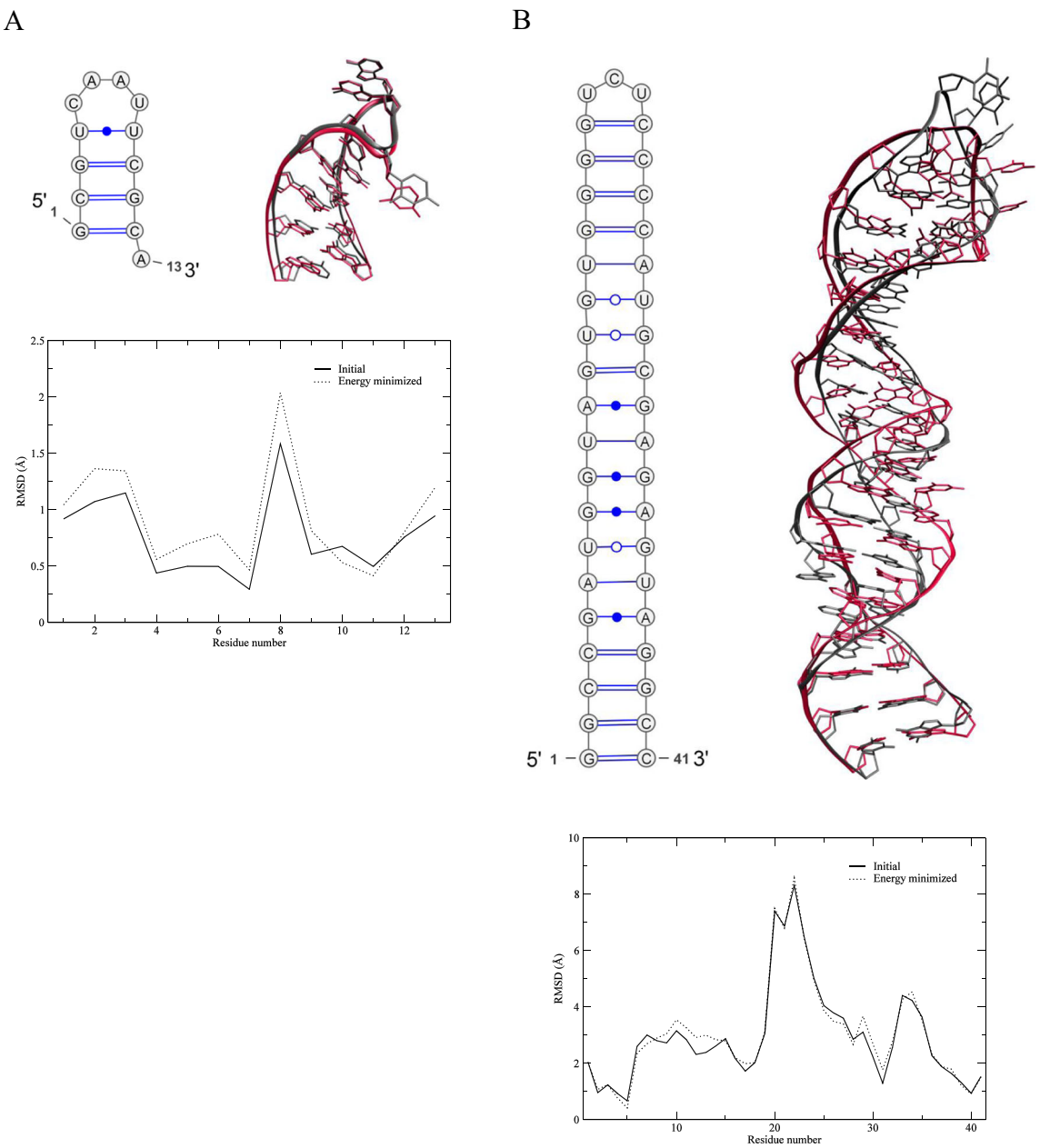


Fig. 2 RNA123-predicted secondary and tertiary structures of (**a**) RNA molecule containing a helix with canonical base pairs and a hairpin loop; PDB ID: 1HS8 (RMSD 1.0 Å) and (**b**) RNA molecule containing a symmetric internal loop; PDB ID: 1A51 (RMSD 3.5 Å). Tertiary structures after energy minimization (*red*) superposed with NMR structures (*grey*). Hydrogens and selected backbone atoms are not displayed for reasons of clarity. The *lower plots* show heavy atom RMSD per residue

are less stable compared to canonical ones, they are naturally more flexible and form less rigid structures.

Figure 2b shows a 41-nt long RNA molecule extracted from the 5S rRNA in *Escherichia coli* (PDB ID: 1A51) that contains a symmetrical internal loop. The loop region contains palindromic motifs of three base pairs each in the two ends of the internal loop. This loop has been indicated to be important for binding to a ribosomal protein [38]. The predicted model shows a good match with the NMR structure, and the total heavy atom RMSD is 3.5 Å. Similar to the structures that contain only a helix with purely canonical base pairs, the largest RMSDs to the X-ray/NMR structures for structures containing symmetric internal loops with interacting non-canonical base pairs are seen overall in the hairpin regions (cf. RMSD per residue plot in Fig. 2b).

Regions with bulges and internal loops are, in most cases, easily identified by the RNA123 program and included in the predicted built models. However, the local geometry in regions of bulges, asymmetric loops, and non-interacting nucleotides in symmetric loops, where the nucleotides are more

flexible compared to base paired ones are more difficult to predict, similar to the hairpin loop regions, resulting in elevated RMSDs when the predicted models are superposed with the X-ray/NMR structures. Bulged nucleotides can form triple base pairing with canonical or non-canonical base pairs. Base-triples are found in several of the structures containing bulged nucleotides included herein. Nucleotides within a hairpin loop can also form triple base interactions with the closing base pair below the hairpin loop. Triple base interactions are in some structures well predicted by RNA123, whereas in other structures the third nucleotide is instead modeled as a bulged nucleotide with no interactions with the other nucleotides. MD simulation did not manage to adjust the faulty structures so as to get the bulged nucleotide to approach any base pair and initiate triple base interaction.

Two examples of RNA molecules containing bulges and asymmetric internal loops are displayed in Fig. 3. Figure 3a shows an RNA fragment extracted from the group I intron in *Tetrahymena* (PDB ID: 1K2G), containing a different geometry compared to the other RNA structures included herein. The molecule is made up by a double hairpin structure with the 5′- and 3′-ends of the sequence instead in the middle of the helix (Fig. 1h). This RNA structure is highly stable due to the crossed base pairing of the two ends (see NMR structure in grey in Fig. 3a). C-1 forms a pair with G-13, and G-22 forms a triple base pair with the canonical pair of G-3 and C-12 [39]. A bulged adenosine (A-2) induces a twist in the helix that affects

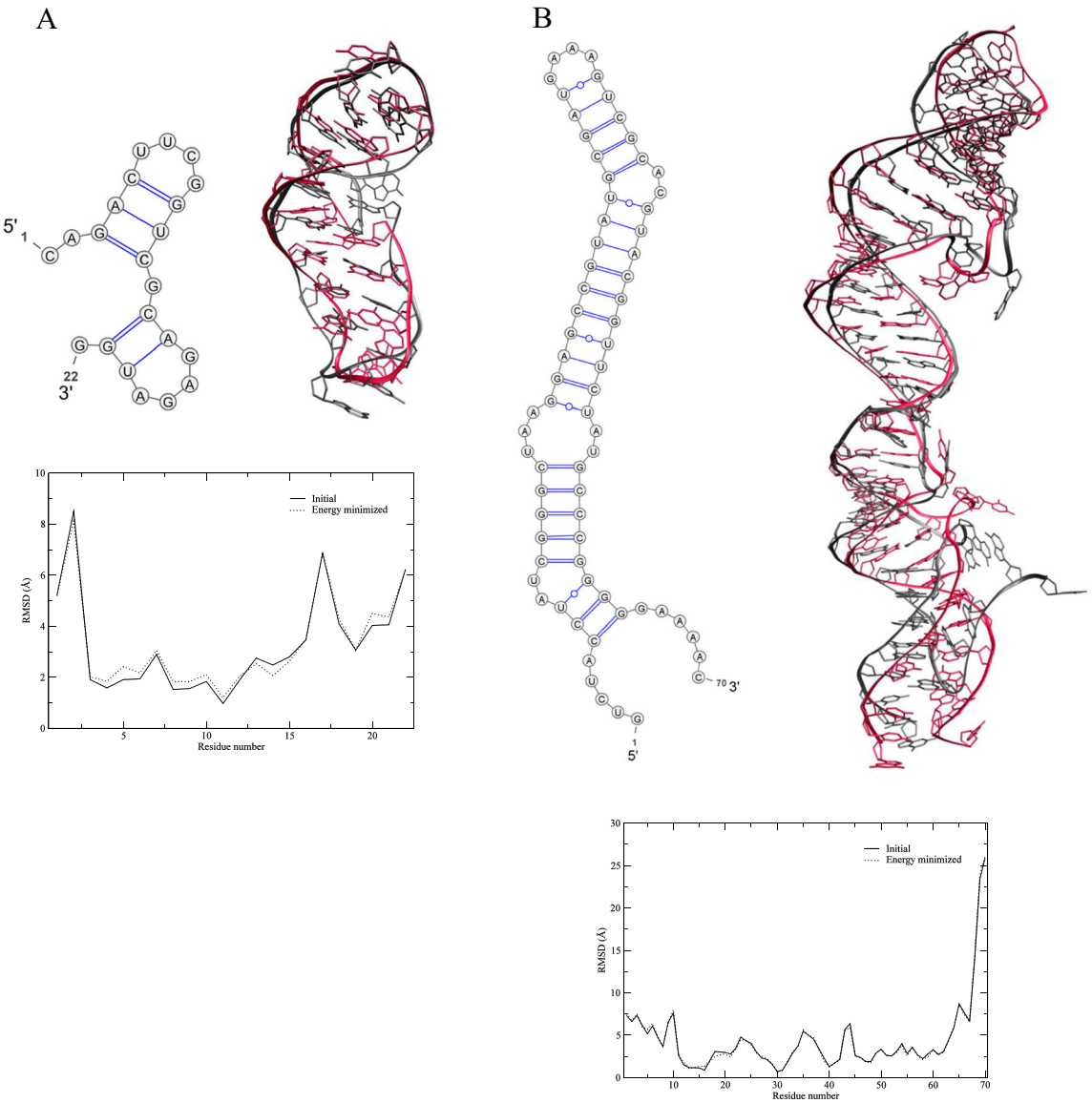


Fig. 3a,b RNA123-predicted secondary and tertiary structures of RNA molecules containing bulges and internal loops. Tertiary structures after energy minimization (*red*) are superposed with X-ray/NMR structures (*grey*). Hydrogens and selected backbone atoms are not displayed for reasons of clarity. The lower plots show heavy atom RMSD per residue. **a** PDB ID: 1K2G (RMSD 3.8 Å), **b** PDB ID: 1KXK (RMSD 6.0 Å)

the base pairs closing the structure. Neither the triple base pair nor the bulged adenosine is caught in the predicted model. G-22 is located too far away to have any interactions with the G-3 and C-12 base pair, and C-1 at the other end also does not show any base pairing. A 100-ns MD simulation did not improve the structure by enabling base pairing of the end regions. The two hairpin loops in this molecule are better predicted, despite a twist in the structure of the 3′ loop, which increases the RMSD (see RMSD per residue graph in Fig. 3a). The other loop shows a very good match with the NMR structure, and the total heavy atom RMSD is 3.8 Å.

The largest molecule included in this benchmarking study is a 70-nt long RNA containing domain 5 and 6 of the group II intron from yeast, shown in Fig. 3b (PDB ID: 1KXK). The helical structure in the top of the figure represents domain 5 and the lower part belongs to domain 6. The long helical structure is interrupted by a bulge and three asymmetric internal loops, each containing two bulged non-interacting nucleotides on one stem adjacent to a G–U pair [40]. The secondary structure predicted by RNA123 (Fig. 3b) contains only three asymmetric internal loops; however, the middle one is in fact an asymmetric internal loop and a one-nucleotide bulge separated by an A-U pair formed in both the X-ray structure and the subsequent tertiary structure predicted by RNA123. In the X-ray structure the bulged A-9 and U-10 in domain 6 stacks with A-69 in the 3′ end of the sequence [40]. This pseudoknot-related structural feature is not seen in the predicted tertiary model; the 5′ and 3′ ends do not interact with either each other or the same stem, which contributes to a large increase in RMSD in this region (see RMSD per residue graph in Fig. 3b). The remaining RNA, including the bulges and internal loops, is well predicted, and the overall heavy atom RMSD is 6.0 Å.

Suboptimal secondary structures could be predicted for all structures in the set of RNA molecules containing bulges and internal loops. In 16 out of 25 structures (structures for which two X-ray/NMR structures were used for comparison; 1CQ5/1CQL and 1JOX/1JP0, and the case of PDB ID: 1R7Z, are not counted here but are instead discussed separately below) the tertiary structure built from the optimal secondary structure generated a lower RMSD to the X-ray/NMR structure compared to the suboptimal structures. The RMSD was decreased by at most 2.7 Å (PDB ID: 1R2P) when a suboptimal secondary structure was used to build the tertiary structure, but in all other cases the improvement in RMSD is at most 0.7 Å.

For PDB ID: 1R7Z, the predicted model displays a very different fold of the helix compared to the NMR structure. This is because the predicted bulge possesses a markedly different geometry, which results in an incorrect twist of the helix. The top part of the helix and the hairpin loop are, however, well predicted and this region was thus superposed before RMSD was calculated (Tables 3, 4), similar to the pseudoknot structures (see below). None of the suboptimal structures were able to better represent the correct structural fold of this molecule. Structures for which the predicted models were compared to two different X-ray/NMR structures require special attention and are discussed below.

A small fraction of the NMR structures included in this benchmarking study are averaged structures, commonly the average of a set of lowest energy conformations. In order to determine how an average structure would compare to an individual NMR structure superposed with a predicted model, one such case was investigated (PDB ID: 1CQ5 is an average structure of the 10 models available in PDB ID: 1CQL). The predicted model of this RNA molecule has a heavy atom RMSD of 4.2 Å from the average structure and 4.1 Å from the most similar of the ten separate NMR structures. The difference is clearly minor. It is however important to note that the ten NMR structures are fairly different in structure, although the average and the lowest RMSD structure are very similar. For the tertiary structures built from the second and third predicted secondary structures the difference in RMSDs are in this case larger than both the lowest RMSD structure and the average structure.

As discussed earlier, various refinement methods can be applied for NMR determinations. In the benchmarking set of RNA structures determined by NMR, several of these were refined with residual dipolar couplings that might improve the structure [41–43]. In order to determine if a predicted model is more similar to the NMR structure refined using residual dipolar couplings compared to the non-refined one, we include an RNA molecule that has been determined with (PDB ID: 1JOX) and without (PDB ID: 1JP0) residual dipolar coupling refinement, and compare our predicted model with these. It should be emphasized that both NMR files contain multiple structures. The lowest heavy atom RMSD of the predicted model is 3.9 Å, when superposed with either the refined or the non-refined NMR structure. The predicted model built from the second predicted secondary structure also displays similar RMSD (4.2 vs 4.4 Å), whereas in the case of the third predicted secondary structure the difference is larger; 3.6 Å for the non-refined structure and 4.3 Å for the refined structure.

Pseudoknots

The third group of RNA molecules that are discussed separately are those that contain pseudoknot structures. Pseudoknots are made up by a single-stranded tail that forms base pairing interactions with nucleotides in the hairpin of a helix. It is clearly seen that the predicted models do not show any pseudoknot features (compare secondary structures in Fig. 1i, j). This is a known issue in RNA123 [33]. Despite this problem, we included seven pseudoknot structures in order to investigate how well the program is able to predict

the double helix region of the molecules. The helices are overall very well predicted as is clearly seen in Fig. 4, where the helices are superposed with the corresponding X-ray/NMR structures. In two out of the seven pseudoknot structures (see Fig. 4a with PDB ID: 2A43), the tail folds back onto itself and forms a short hairpin structure, although not interacting with the helix, whereas in the five other structures the tail shows no interaction with either itself or the initially formed helix segment. The heavy atom RMSD for only the helices is less than or equal to 2.0 Å for five of the structures after energy minimization, whereas two structures show larger deviations in their folds compared to the NMR structures, and have RMSDs of 5–6 Å. For three of the molecules only one secondary structure was predicted by RNA123. For the ones where more than one secondary structure was predicted, in only one case did one of the suboptimal structures correspond to a tertiary structure with lower RMSD of the helix to the NMR structure, although the difference was only 0.1 Å. For PDB ID: 387D the most optimal secondary structure did not provide a structure possible to fold into a pseudoknot, but instead the fourth predicted secondary structure was used to build the tertiary structure. The remaining tertiary structures for this system, built from other secondary structures, are not included in the results presented here.

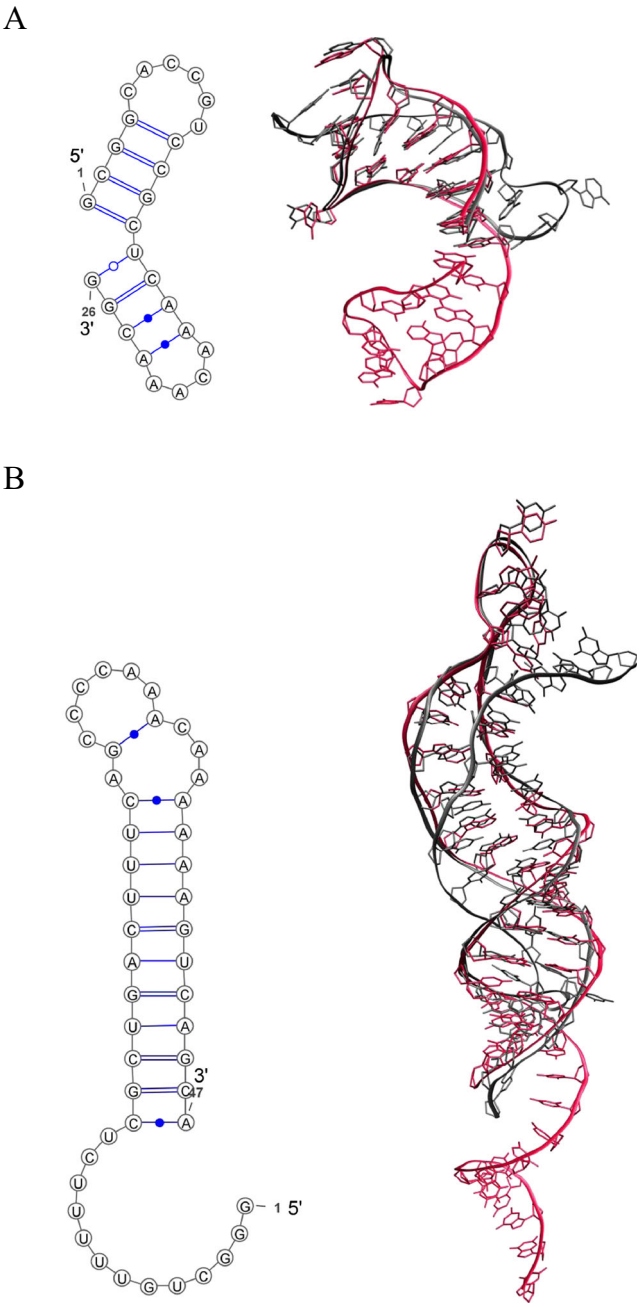


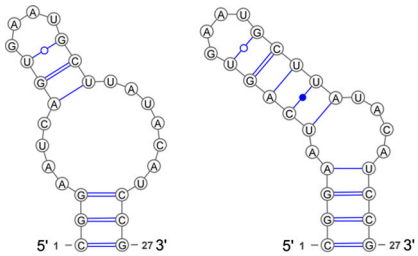
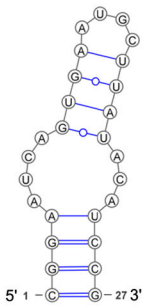
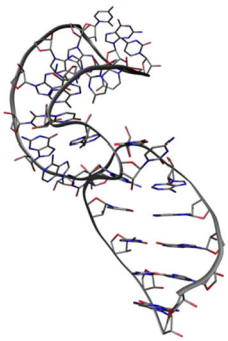
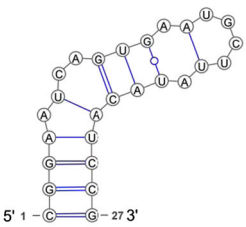
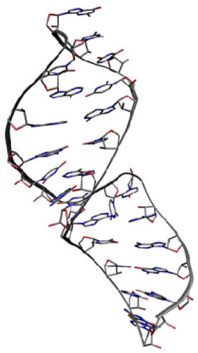
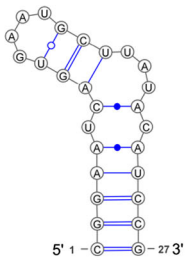
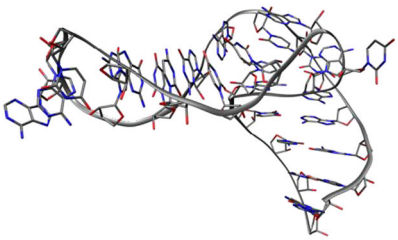
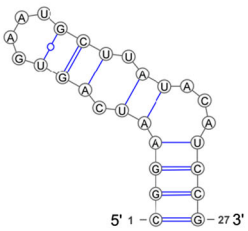
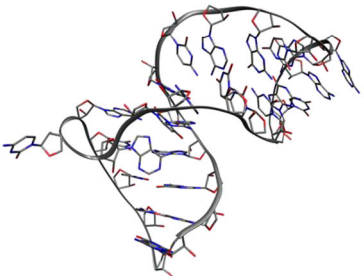
Fig. 4a,b RNA123-predicted secondary and tertiary structures of RNA molecules containing pseudoknots. Tertiary structures after energy minimization (*red*) superposed with X-ray/NMR structures (*grey*). Hydrogens and selected backbone atoms are not displayed for reasons of clarity. **a** PDB ID: 2A43, **b** PDB ID: 1YMO

Prediction of the tertiary structure of Macugen®

Confident that the RNA123 program is able to satisfactory predict the tertiary structure of RNA structures not containing pseudoknots, we apply this to Macugen®—an RNA aptamer used in medical treatment of age-related macular degeneration. Possible secondary structures of Macugen have been predicted by NMR (Fig. 5) [44, 45]. However, further studies would be required in order to determine the accuracy of those. No tertiary structure of Macugen has yet been determined. Such information would be essential in elucidating the detailed interactions with the vascular endothelial growth factor (VEGF) to which the aptamer binds and subsequently inhibits. The proposed secondary structures indicate that an asymmetric internal loop exists in the center of the helix and a hairpin tetraloop motif consisting of G–11–U–14 [44, 45]. Early experimental studies identified U-14 as crucial for the interaction with the VEGF as it forms a photo-crosslink with a cysteine residue in the protein [46]. Tertiary structure data could reveal vital information useful for development of drugs for a wide range of conditions involving growth factors.

Macugen is designed with 2′-fluorine substituted pyrimidines and 2′-O-methylated purines to prevent cleavage by endonucleases, a 40-kDa 5′ polyethyleneglycol (PEG) moiety, and a 3′–3′-linked deoxythymidine residue. These modifications may induce conformational changes in the structures but were not included in the initial model studied herein; instead, we focus here on the pure nucleic acid structure. The actual binding to VEGF most likely involves an induced-fit mechanism that results in stabilizing the structure of the aptamer

Fig. 5 NMR- and RNA123-predicted secondary structures and ▶ corresponding tertiary structures of the RNA sequence in Macugen. The helix ends of the tertiary structures are aligned in order to show the different folds. Hydrogens and selected backbone atoms are not displayed for reasons of clarity

	Secondary structure	Tertiary structure
NMR [44,45]		
1		
2		
3		
4		

[45]. This would generate a slightly different fold of the aptamer compared to that predicted in the absence of the VEGF; however, we do not expect any global conformational changes to occur upon binding.

Figure 5 displays the four most optimal secondary structures predicted by RNA123 and their corresponding tertiary structures after energy minimization. The secondary structures clearly show a double helix and hairpin motif, but the different structures differ in the fold of the helix, determined by the nucleotide base interactions between the two stems originating after the four outmost canonical base pairs that make up an identical initial helix structure in all four structures. The tertiary structure corresponding to the optimal predicted secondary structure (**1**) displays a helix structure with an asymmetric internal loop containing non-interacting bases and closed by a G–U pair. The internal loop induces a significant twist in the helix structure. The continuation of the helix is made up by two canonical base pairs and one G–U pair, and the helix ends up with a tetraloop of A–13–C–16.

The first suboptimal structure (**2**) has a similar helix structure, but shows a more pronounced twist due to three bulged residues on the 5′ stem. U-6 and A-23 forms a canonical pair surrounded by the two bulges that are made up by one and two nucleotides, respectively. A-23 is also able to form hydrogen bonds with A-5 and A-8 as a result of its location. The tetraloop is shifted one nucleotide compared to the previous structure, and is here located at U–14–U–17. The two structures differ by 6.0 Å in heavy atom RMSD.

The next two suboptimal structures (**3** and **4**) show significantly different folds compared to the first two due to the presence of bulged non-interacting nucleotides on the opposite 3′ stem compared to in **1** and **2**. However, all four structures share the initial region of the helix with the four canonical base pairs. Structures **3** and **4** share the same hairpin tetraloop structure and two canonical and one wobble G–U base pairs prior to the hairpin loop. The G–11–U–14 hairpin loop in these structures is identical to the one in the secondary structures proposed by NMR experiments [44, 45]. The difference in the two structures arises from the predicted locations of non-interacting bases, and this subsequently determines the fold of the helix. In **3**, U-18, A-19, and U-20 form a bulge of non-interacting bases, whereas in **4**, a bulge is instead formed by A-21, C-22 and A-23 further down the helix, similar to one of the suggested secondary structures determined by NMR [45]. The two structures show very different overall folds, and the heavy atom RMSD between the two structures is 11.9 Å.

All four structures have U-14 in the hairpin loop, i.e., the nucleotide that has been identified to be crucial for interaction with VEGF [46]. This nucleotide is, however, positioned differently in the four structures. Structure **4** possesses a structure that has U-14 in the region of the hairpin loop that points in the direction of the spacious pocket that arises due to the substantial fold of the helix, which is likely the region in which VEGF tightly binds. U-14 is hence located to enable interaction if the VEGF binds in the pocket. This, together with the similarity in secondary structure to the ones suggested from NMR experiments, leads us to conclude that the tertiary structure of **4** holds high probability to closely represent the native fold of the aptamer. In **3**, U-14 is located in the region of the hairpin loop that points away from the pocket, thus an interaction between the VEGF in the pocket would not be possible without significant re-folding of the structure. Structures **1** and **2** possess less pronounced pockets and none of them have U-14 in a location for which it could easily interact with the VEGF if it binds directly in the pocket, in the static state in which it is at present. However, as mentioned above, an induced fit procedure is likely to occur when the VEGF binds the aptamer, and this conformational change could result in U-14 to move into a position so as to interact with the VEGF. A more thorough study of the substituted nucleic acid sequence, the impact of the PEG moiety on the structure and availability of U-14, and the exact interactions between the models and the VEGF could likely elucidate more clearly which aptamer structure constitutes the native fold.

Conclusions

We have performed an initial benchmarking study in which we found that the RNA123 program is able to accurately predict 3D models of a set of 50 hairpin RNA molecules based only on their nucleic acid sequence. Compared to MC-Fold and MC-Sym, RNA123 performed overall better in de novo prediction of the benchmarking set of structures. For tertiary structures predicted by RNA123, lowest heavy atom RMSD to the X-ray/NMR structures were found for structures made up by only a continuous helix of canonical base pairs or symmetric internal loops of interacting non-canonical nucleotides, whereas larger differences were observed for more complex structures containing bulges and asymmetric internal loops. For a majority of structures, the program was able to predict models with a heavy atom RMSD of less than 5 Å to the X-ray/NMR structures. Tertiary structures built from predicted sub-optimal secondary structures showed, in a majority of cases, larger deviations, indicating that the predicted optimal secondary structure corresponds to the native tertiary structure. For RNA molecules containing pseudoknots, the program was not able to predict the folding of the single-stranded stem onto the helix; however, this is a known problem in many RNA folding programs. It is important to stress that X-ray and NMR predicted structures are not always fully reliable. They often contain errors such as

incorrect base conformations and may therefore not necessarily correspond to the native fold of a molecule. This can thus have a significant impact on the analysis when a computationally predicted model is judged based on an X-ray or NMR structure.

From the predictions of possible tertiary structures of the RNA sequence present in Macugen we conclude that further studies of the full molecule and its interactions the VEGF are required in order to fully determine whether the optimal predicted structure corresponds to the natural fold of the molecule.

Acknowledgments The Faculty of Science at the University of Gothenburg and the Swedish research council (VR) are gratefully acknowledged for financial support.

References

1. Gautheret D, Konings D, Gutell RR (1995) GU base pairing motifs in ribosomal RNA. RNA 1(8):807–814
2. Crick FHC (1966) Codon-anticodon pairing: the wobble hypothesis. J Mol Biol 19:548–555
3. Jaeger JA, Turner DH, Zuker M (1989) Improved predictions of secondary structures for RNA. Proc Natl Acad Sci USA 86(20): 7706–7710
4. Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288(5):911–940. doi: 10.1006/jmbi.1999.2700
5. Strazewski P, Biala E, Gabriel K, McClain WH (1999) The relationship of thermodynamic stability at a G x U recognition site to tRNA aminoacylation specificity. RNA 5(11):1490–1494
6. Doudna JA, Cormack BP, Szostak JW (1989) RNA structure, not sequence, determines the 5′ splice-site specificity of a group I intron. Proc Natl Acad Sci USA 86(19):7402–7406
7. Hur M, Waring RB (1995) Two group I introns with a C.G. basepair at the 5′ splice-site instead of the very highly conserved U.G. basepair: is selection post-translational? Nucleic Acids Res 23(21): 4466–4470
8. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. Science 289(5481):905–920
9. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A (2000) Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. Cell 102(5):615–623
10. Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V (2000) Structure of the 30S ribosomal subunit. Nature 407(6802):327–339. doi:10.1038/35030006
11. Abu Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB (2012) Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic Acids Res 40(4):1407–1423. doi: 10.1093/nar/gkr810
12. Agarwal T, Jayaraj G, Pandey SP, Agarwala P, Maiti S (2012) RNA G-quadruplexes: G-quadruplexes with "U" turns. Curr Pharm Des 18(14):2102–2111
13. Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7(4):499–512
14. Woese CR, Winker S, Gutell RR (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". Proc Natl Acad Sci USA 87(21):8467–8471
15. Heus HA, Pardi A (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. Science 253(5016):191–194
16. Baeyens KJ, De Bondt HL, Pardi A, Holbrook SR (1996) A curved RNA helix incorporating an internal loop with G.A. and A.A. non-Watson-Crick base pairing. Proc Natl Acad Sci USA 93(23):12851–12855
17. Carter RJ, Baeyens KJ, SantaLucia J, Turner DH, Holbrook SR (1997) The crystal structure of an RNA oligomer incorporating tandem adenosine-inosine mismatches. Nucleic Acids Res 25(20): 4117–4122
18. Chen G, Znosko BM, Kennedy SD, Krugh TR, Turner DH (2005) Solution structure of an RNA internal loop with three consecutive sheared GA pairs. Biochemistry 44(8):2845–2856. doi:10.1021/bi048079y
19. Hammond NB, Tolbert BS, Kierzek R, Turner DH, Kennedy SD (2010) RNA internal loops with tandem AG pairs: the structure of the 5′GAGU/3′UGAG loop can be dramatically different from others, including 5′AAGU/3′UGAA. Biochemistry 49(27):5817–5827. doi: 10.1021/bi100332r
20. Klein DJ, Schmeing TM, Moore PB, Steitz TA (2001) The kink-turn: a new RNA secondary structure motif. EMBO J 20(15):4214–4221. doi:10.1093/emboj/20.15.4214
21. Szep S, Wang J, Moore PB (2003) The crystal structure of a 26-nucleotide RNA containing a hook-turn. RNA 9(1):44–51
22. Rietveld K, Van Poelgeest R, Pleij CW, Van Boom JH, Bosch L (1982) The tRNA-like structure at the 3′ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. Nucleic Acids Res 10(6):1929–1946
23. Giedroc DP, Cornish PV (2009) Frameshifting RNA pseudoknots: structure and mechanism. Virus Res 139(2):193–208. doi:10.1016/j.virusres.2008.06.008
24. Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinforma 5:140. doi:10.1186/1471-2105-5-140
25. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452(7183):51–55. doi:10.1038/nature06684
26. Jossinet F, Ludwig TE, Westhof E (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. Bioinformatics 26(16):2057–2059. doi:10.1093/bioinformatics/btq321
27. Martinez HM, Maizel JV Jr, Shapiro BA (2008) RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. J Biomol Struct Dyn 25(6):669–683. doi:10.1080/07391102.2008.10531240
28. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods 7(4):291–294. doi:10.1038/nmeth.1433
29. Tan RK, Petrov AS, Harvey SC (2006) YUP: a molecular simulation program for coarse-grained and multi-scaled models. J Chem Theory Comput 2(3):529–540. doi:10.1021/ct050323r
30. Sharma S, Ding F, Dokholyan NV (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics 24(17):1951–1952. doi:10.1093/bioinformatics/btn328
31. Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. RNA 14(6):1164–1173. doi:10.1261/rna.894608
32. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA 15(2):189–199. doi:10.1261/rna.1270809

33. Sijenyi F, Saro P, Ouyang Z, Damm-Ganamet K, Wood M, Jiang J, SantaLucia J Jr (2011) The RNA folding problems: different levels of sRNA structure prediction. In: Leontis N, Westhof E (eds) RNA 3D structure analysis and prediction. Springer, Berlin, pp 91–117
34. Darty K, Denise A, Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 25(15): 1974–1975. doi:10.1093/bioinformatics/btp250
35. Krieger E, Darden T, Nabuurs SB, Finkelstein A, Vriend G (2004) Making optimal use of empirical energy functions: force-field parameterization in crystal space. Proteins 57(4):678–683. doi:10.1002/prot.20251
36. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang J, Kollman P (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24(16):1999–2012. doi:10.1002/jcc.10349
37. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103(19): 8577–8593
38. Dallas A, Moore PB (1997) The loop E-loop D region of *Escherichia coli* 5S rRNA: the solution structure reveals an unusual loop that may be important for binding ribosomal proteins. Structure 5(12):1639–1653
39. Kitamura A, Muto Y, Watanabe S, Kim I, Ito T, Nishiya Y, Sakamoto K, Ohtsuki T, Kawai G, Watanabe K, Hosono K, Takaku H, Katoh E, Yamazaki T, Inoue T, Yokoyama S (2002) Solution structure of an RNA fragment with the P7/P9.0 region and the 3′-terminal guanosine of the tetrahymena group I intron. RNA 8(4):440–451
40. Zhang L, Doudna JA (2002) Structural insights into group II intron catalysis and branch-site selection. Science 295(5562):2084–2088. doi:10.1126/science.1069268
41. Warren JJ, Moore PB (2001) Application of dipolar coupling data to the refinement of the solution structure of the sarcin-ricin loop RNA. J Biomol NMR 20(4):311–323
42. Warren JJ, Moore PB (2001) A maximum likelihood method for determining D(a)(PQ) and R for sets of dipolar coupling data. J Magn Reson 149(2):271–275. doi:10.1006/jmre.2001.2307
43. Vallurupalli P, Moore PB (2003) The solution structure of the loop E region of the 5S rRNA from spinach chloroplasts. J Mol Biol 325(5): 843–856
44. Lee JH, Canny MD, De Erkenez A, Krilleke D, Ng YS, Shima DT, Pardi A, Jucker F (2005) A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165. Proc Natl Acad Sci USA 102(52):18902–18907. doi:10.1073/pnas.0509069102
45. Lee JH, Jucker F, Pardi A (2008) Imino proton exchange rates imply an induced-fit binding mechanism for the VEGF165-targeting aptamer, Macugen. FEBS Lett 582(13):1835–1839. doi:10.1016/j.febslet.2008.05.003
46. Ruckman J, Green LS, Beeson J, Waugh S, Gillette WL, Henninger DD, Claesson-Welsh L, Janjic N (1998) 2′-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165). Inhibition of receptor binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain. J Biol Chem 273(32):20556–20567