Published online 29 May 2015

Nucleic Acids Research, 2015, Vol. 43, No. 17 ***e114***
doi: 10.1093/nar/gkv559

MINT: software to identify motifs and short-range interactions in trajectories of nucleic acids

Anna Górska[1,2], Maciej Jasiński[1,3] and Joanna Trylska[1,*]

[1]Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland, [2]Master studies at the Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, Warsaw, Poland and [3]College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Al. Żwirki i Wigury 93, 02-089 Warsaw, Poland

Received January 15, 2015; Revised April 30, 2015; Accepted May 15, 2015

ABSTRACT

Structural biology experiments and structure prediction tools have provided many high-resolution three-dimensional structures of nucleic acids. Also, molecular dynamics force field parameters have been adapted to simulating charged and flexible nucleic acid structures on microsecond time scales. Therefore, we can generate the dynamics of DNA or RNA molecules, but we still lack adequate tools for the analysis of the resulting huge amounts of data. We present MINT (Motif Identifier for Nucleic acids Trajectory) — an automatic tool for analyzing three-dimensional structures of RNA and DNA, and their full-atom molecular dynamics trajectories or other conformation sets (e.g. X-ray or nuclear magnetic resonance-derived structures). For each RNA or DNA conformation MINT determines the hydrogen bonding network resolving the base pairing patterns, identifies secondary structure motifs (helices, junctions, loops, etc.) and pseudoknots. MINT also estimates the energy of stacking and phosphate anion-base interactions. For many conformations, as in a molecular dynamics trajectory, MINT provides averages of the above structural and energetic features and their evolution. We show MINT functionality based on all-atom explicit solvent molecular dynamics trajectory of the 30S ribosomal subunit.

INTRODUCTION

Nucleic acids, especially RNA, acquire many complicated tertiary structures to perform cellular functions (1). Provided that this tertiary structure is known, one of the common tools to investigate the structural and dynamical properties of nucleic acids and their complexes on atomic scale is molecular dynamics (MD) (2). With this technique riboswitches (3), protein–RNA complexes (4,5) and even the entire ribosome (6) have been studied. Other methods to sample the conformational space are the stochastic-based Monte Carlo techniques. Their applications to RNA molecules include the investigation of folding kinetics (7,8). Most importantly, all these simulation methods generate large data sets, i.e. many molecule conformations, which have to be post-processed.

Many computational tools have been designed to analyze single RNA conformations (9). One of the most comprehensive is Assemble2 (10), which reads the RNA secondary structure, constructs structural alignments of several RNAs, and overall facilitates RNA structure prediction and modeling. For detailed geometric analyses of RNA, especially its helical fragments, the Curves+ (11) or 3DNA (12) can be used. The programs apply standard reference frame (13) and describe the mutual position of two nucleotides and the conformation of the backbone using torsional angles. Many tools can analyze RNA structures based on the contacts and interactions between nucleotides. For basic identification of base pairs, stacking interactions and structural elements, such as helices, bulges and pseudoknots the MC-Annotate (14) can be used. More detailed description, together with the two-dimensional (2D) representations of RNA, can be obtained with RNAView (15) or RNAmap2D. The latter can also analyze complexes of nucleic acids and other molecules such as proteins or ligands and metal ions (16). The ClaRNA program additionally classifies each contact based on the similarity to a reference obtained from a large number of experimentally determined RNA structures (17). Recently announced, DSSR, a new component of the 3DNA, can define the secondary structures of RNA from three-dimensional (3D) coordinates, recognize motifs and non-pairing interactions (18). However, it does not analyze energetics of the recognized interactions.

On the other hand, there are only few programs for the analysis of RNA conformation sets obtained from full-atom MD simulations. One is Canal, which applies Curves+ (11) to every trajectory frame and computes statistics, histograms and correlations for various measures such as groove widths and depths, backbone dihedrals and base

*To whom correspondence should be addressed. Tel: +48 22 5543600; Fax: +48 22 5540801; Email: joanna@cent.uw.edu.pl

pairing parameters. 3DNA includes scripts facilitating the analysis of MD data for nucleic acids (12), and do_x3dna extends 3DNA applications to GROMACS trajectories (19).

In fact, there is no complex tool to help analyze the dynamics of both the secondary and tertiary RNA structures.Therefore, we have designed Motif Identifier for Nucleic acids Trajectory (MINT) to characterize multiple RNA structures and the changes in hydrogen bonds, base pairing patterns, stacking and secondary structure motifs. The conformation sets can be from MD trajectories, Monte Carlo simulations, X-ray or nuclear magnetic resonance-derived conformations of the same molecule. MINT works for both RNA and DNA. However, since it is mainly RNA that acquires complicated 3D folds, we describe the software based on the RNA example.

MATERIALS AND METHODS

MINT works in a single and multiple conformation mode. For a single RNA or DNA conformation MINT outputs:

- nucleotides forming helices, hairpin loops, internal loops, junctions, pseudoknots and other motifs, together with their classification,
- all Watson–Crick (WC) edge and non-Watson–Crick (non-WC) edge pairs, along with their *cis* or *trans* configuration and edge-to-edge classification (20).
- the number of WC-edge and non-WC-edge hydrogen bonds (and their sum) per nucleotide,
- the stacking energy: van der Waals (VDW) and electrostatic interaction terms (and their sum) per nucleotide,
- all phosphate anion–π interacting nucleotides,
- files necessary for the visualization of the above properties.

The multiple conformation mode works as a standalone package to analyze many conformations of one molecule, e.g. from a trajectory. MINT computes the above listed properties for every frame/conformation and, in addition, outputs the statistics:

- nucleotides forming helices, loops, pseudoknots and other motifs together with their occurrence (i.e. the frame numbers in which these motifs were detected and the percentage of trajectory time they lasted),
- clusters of secondary structure motifs and average motifs along with 2D and 3D contacts,
- all WC-edge, non-WC-edge pairs and triples, stacking and anion–π interacting nucleotides with their occurrence,
- for each nucleotide MINT lists the nucleotides with which it formed hydrogen bonds (giving the number of hydrogen bonds and their occurrence),
- the average secondary structure,
- correlations in the breaking and forming of the WC-edge pairs,
- the average number of WC-edge and non-WC-edge hydrogen bonds (and their sum) per nucleotide,
- the average stacking energy – VDW and electrostatic terms (and their sum) per nucleotide,

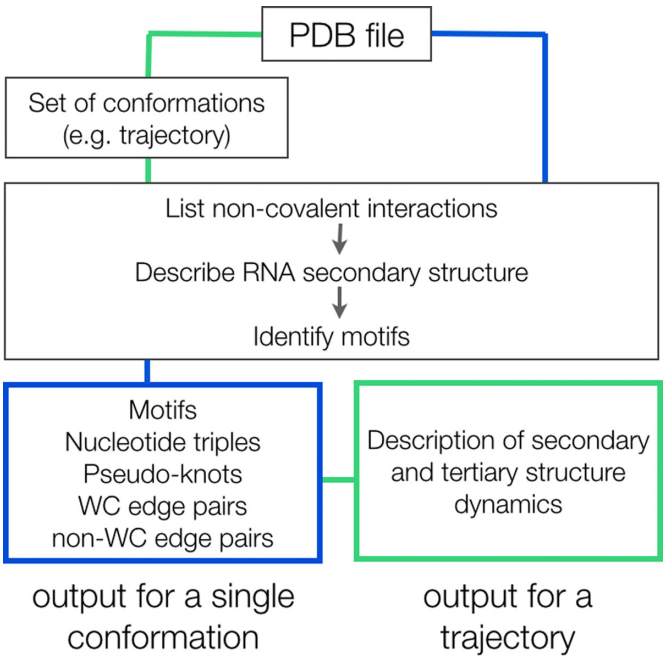

Figure 1. MINT workflow. The main function implements the analysis of a single frame. For a trajectory, the function first creates a table of all atoms from a .pdb file with their coordinates from the entire trajectory. While analyzing trajectory frames, the coordinates are read from the created table.

- visualizations of the outputs and files that can be used in VARNA (21), Visual Molecular Dynamics (VMD) (22), Chimera (23) and RNAMovies (24).

MINT is written in Python language. BioPython (25) is used to read files in the Protein Data Bank format (.pdb) and MDAnalysis (26) to process trajectories. The analysis of many frames/conformations can be run in parallel on any number of CPUs and is limited only by the amount of memory. MINT splits the trajectory into pieces of equal lengths and analyzes each sub-trajectory on a separate core but at the same time. Finally, the program computes statistics for all frames. The software, manual and server are available at http://mint.cent.uw.edu.pl.

Implementation

First, MINT reads a .pdb file with coordinates and then a file with many conformations of the same molecule. The supported formats for the latter are .dcd, .xyz, .trr and .crd. For every inputted frame MINT maps the hydrogen bonds, recognizes base pairs and writes the dot-bracket representation of the secondary structure. Next, it runs the algorithm to classify structural motifs and at the same time searches for stacking and anion–π interactions (Figure 1).

Hydrogen bond definition. A hydrogen bond is the basic term of the program. It is defined as a non-covalent interaction in which a hydrogen atom of a donor is placed close to the acceptor. The hydrogen bond criteria are defined by an angle between the acceptor, hydrogen atom and the donor (default minimal angle is 140 degrees) and a distance be-

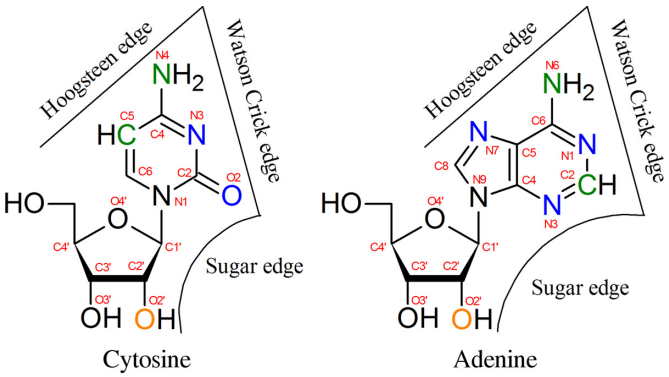

Figure 2. Nucleotide edges with hydrogen donors in green, acceptors in blue and atoms that serve as both hydrogen donor and acceptor in orange (27).

tween the donor and acceptor (default: 3.25 Å) or a distance between the acceptor and hydrogen atom (default: 2.8 Å).

Donors and acceptors. To analyze nucleic acids we defined a list of possible acceptors and donors for all standard nucleotides (A, U, T, G, C). Following the classification by Leontis and Westhof (20) the acceptors and donors are assigned to the nucleotide edges shown in Figure 2.

`MINT` checks if there is a hydrogen bond created by any of the donors or acceptors of all possible pairs placed within a defined distance cutoff. Knowing the atoms participating in hydrogen bonds, the program determines the interacting edges and using the edge information classifies a pair. Note that we use the edge-to-edge classification (20,27), instead of the concept of the canonical base pairs, to unambiguously and consistently describe all possible geometric pairs that may, even transiently, occur in a trajectory.

Base pair geometric isomerism. For detected nucleotide pairs geometric isomerism of their glycosidic bonds is computed. The program measures the torsion angle formed by four atoms (C1', N1 in pyrimidines and C1', N9 in purines) and depending on its value a *cis* or *trans* conformation is denoted.

Aromatic stacking. For almost 300 geometries of stacked base dimers, Šponer *et al.* compared *ab initio* stacking energies with energies obtained using simple pairwise-additive empirical potentials (28–31). They found that calculations applying the Lennard-Jones potential and Coulombic terms with atom-centered point charges reproduce the *ab initio* stacking energies of base dimers within +/−1.5 kcal/mol (30). This agreement suggests that calculations based on empirical potentials approximate well the stacking interaction energy between nucleobases. Therefore, we estimate the energy of stacking between two nucleobases as the sum of electrostatic (U_{el}) and van der Waals (U_{VDW}) interaction terms:

$$U_{el} = k \sum \frac{q_i q_j}{r_{ij}} \qquad (1)$$

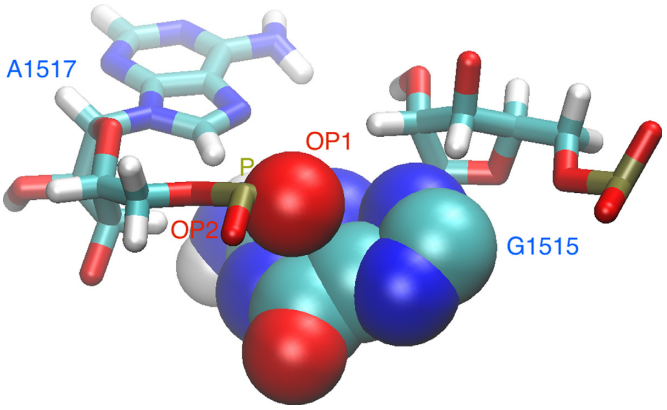

Figure 3. The phosphate group oxygen atom 'stacking' over the guanine base. The oxygen and base atoms are shown as spheres of sizes corresponding to their VDW radii. A fragment is from the `4GD2.pdb` file.

$$U_{\mathrm{VDW}} = 4\epsilon \sum \left[\frac{1}{4} \left(\frac{r_0}{r_{ij}} \right)^{12} - \frac{1}{2} \left(\frac{r_0}{r_{ij}} \right)^{6} \right], \qquad (2)$$

where k is the Coulomb constant ($k = \frac{1}{4\pi\epsilon_0}$), q_i and q_j are atom-centered point charges, r_{ij} is the distance between the atoms i, j, ϵ is the depth of the Lennard-Jones potential well for atoms i, j, and r_0 is the sum of VDW radii of atoms i and j. The sum runs over all atoms of both interacting nucleobases.

The planar shape of nucleobases ensures that the lowest VDW energies are obtained for parallel orientations of bases with the largest geometric overlap. With this in mind, we assume that two nucleobases are stacked, if their VDW energy is lower than a given threshold. By default it is set to −0.5 kcal/mol, and was estimated by trial and error but seems appropriate for non-modified nucleobases. `MINT` provides the VDW parameters and partial atomic charges for RNA and DNA nucleotides from the Amber (32) and Charmm (33,34) force fields. Users may also supply their own parameters.

Anion–π contacts. Following recent study on the types of non-covalent contacts in RNA, `MINT` also analyzes the phosphate oxygen contacts with nucleobases, the so-called anion–π contacts (P. Auffinger 2013, personal communication). An example of such contact is shown in Figure 3. These contacts are detected based on the distance between the oxygen atom of a phosphate group of one nucleotide and the center of mass of another nucleobase ring. The energy of interaction for these non-covalent contacts is estimated in the same way as for aromatic stacking. Only systems with the distance lower than a given threshold (default is 5 Å) are considered as anion–π interacting. Since there is no guarantee that the current force fields, based on empirical potentials, describe well these types of contacts, the user has to verify the energetics of the recognized complexes.

Modified nucleotides. Modified nucleotides are found in many functional non-coding RNAs, e.g. tRNAs (35,36). However, the standard Amber (32) and Charmm (33,34) force fields provide parameters only for the non-modified

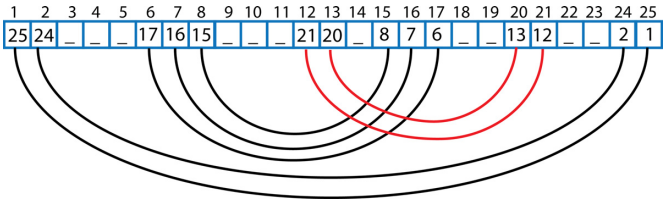

Figure 4. An example of a list-representation of RNA secondary structure. Every cell of the matrix contains the nucleotide number that is WC-edge paired with the nucleotide indicated by the matrix index (above the cell). Therefore, nucleotide no 1 pairs with nucleotide no 25, nucleotide no 2 with 24 and so on. Base pairing is marked by black curved lines and the WC-edge interactions creating a pseudoknot by red curves.

nucleotides. Also, the edge-to-edge classification of hydrogen bonds between RNA bases was proposed only for non-modified nucleotides A, U, G and C (20).

For modified nucleotides we provide the VDW parameters and partial atomic charges developed by Aduri *et al.* (37) for 107 naturally occurring modifications. For the edge-to-edge classification of base pairs formed by these nucleotides, we assign their atoms to four distinct edges: the WC edge, the Hoogsteen edge, the sugar edge and to the edge termed modification. Atoms common for the modified and non-modified nucleobase are assigned as in the non-modified one (Figure 2). If an atom existing in a non-modified nucleotide is substituted by one other atom or if one atom is added, we assign such atoms to the same edge as in the non-modified case. The 2'O methyl carbon is classified into the sugar edge. All other atoms of the modified nucleotide are classified into the modification edge.

If MINT detects a modified nucleotide, for which there are no parameters, it automatically assigns its atoms to edges. For atoms assigned to the modification edge the stacking interaction energy is set to 0. However, the user may add parameters for modified nucleotides. The force field parameters prepared for all-atom MD simulation of modified nucleotides can be further adopted to MINT parameter format.

Representation of RNA secondary structure. After detecting the WC-edge pairs, we create a list representation of the RNA secondary structure. Most nucleotides have only one WC-edge partner but in MD trajectory it may transiently happen that a second WC-edge partner is encountered, and such a triple is not considered in the secondary structure analysis. The index in the list represents the nucleotide number; the stored value is the index of its WC-edge partner. The list is easy interpretable if the arcs connecting the pairs are drawn as in Figure 4.

Pseudoknots. The list-representation contains also pseudoknots–non-secondary motifs formed by WC-edge pairing. MINT detects a pseudoknot fold if the arcs intersect. A pseudoknot is a symmetric structure so in Figure 4 both the three pairs 6–17, 7–16 and 8–15 and the two pairs 12–21, 13–20 form a pseudoknot.

To erase pseudoknots from the list representation of the secondary structure, so they do not disturb the motif-search algorithm, we use a conflict elimination method (38) leading to a nested structure containing the maximum number

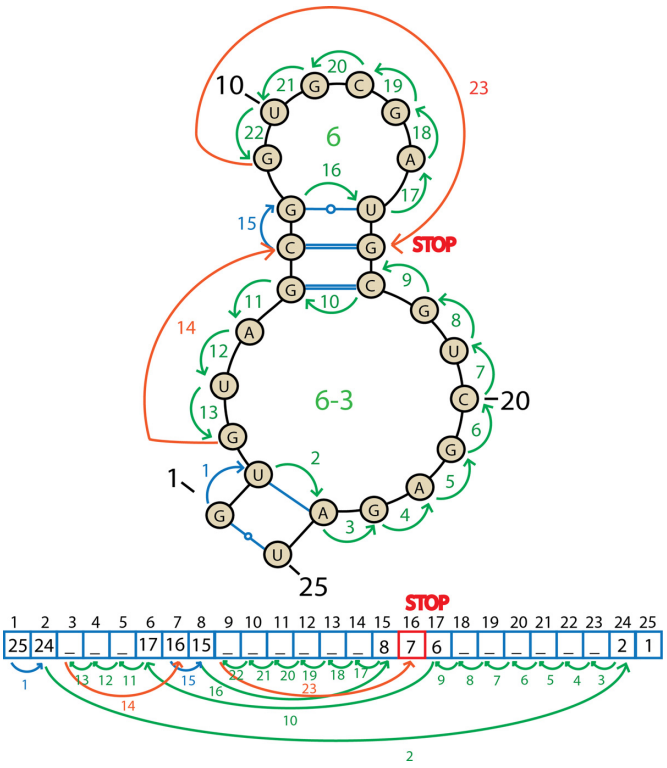

Figure 5. A scheme of the algorithm traveling around the secondary structure of RNA in the graph and list representation. The arrows and their numbers indicate sequential steps of the algorithm. Blue arrows mark helices and green arrows other structural motifs. Orange lines show the jumps that the algorithm takes after distinguishing a motif. The STOP sign indicates the position where the algorithm terminates.

of base pairs. In this case the pairs 12–21 and 13–20 are removed and classified as a pseudoknot.

RNA motif description. MINT describes motifs by numbering unpaired nucleotides detected between base pairs on the edges of the motif. For examples of the codes describing motifs see Supplementary Figure S1. A four nucleotide loop of a hairpin is assigned a single number 4. An asymmetric internal loop, with three unpaired nucleotides in one strand, is represented by two numbers 0–3. A symmetric internal loop, with three unpaired nucleotides in each strand, is coded as 3–3, and a three-way junction, without unpaired nucleotides: 0–0–0.

Motif-search algorithm. The algorithm uses a list representation of the RNA secondary structure. To detect helices and other motifs it walks through the list and stores the information about the structure (Figure 5).

For two sequential WC-edge paired nucleotides, the algorithm stores that the nucleotide is a part of a helix (step 1 in Figure 5). If a helix ends, i.e. an unpaired nucleotide is encountered ahead of a pair, the algorithm starts to travel around the motif; it remembers the first pair, the border of the motif, and goes to the index stored in the list (step 2). The algorithm moves back until it encounters an unpaired nucleotide—with decreasing indexes (steps 3–9); if a pair is found, the algorithm goes to the indicated position in the

list (step 10) and again moves back (steps 11–13); the motif ends once the algorithm finds itself one step ahead of the previously remembered starting index of the motif (step 13). After a motif is found and classified, the algorithm jumps one step further from the last seen pair (step 14), and behaves identically: stores the helix (step 15), jumps to a pair of the last nucleotides (step 16), travels around the motif (steps 17–22), finds itself one step ahead of the starting motif (step 22) and jumps to the last seen pair plus one index (step 23). The algorithm stops searching for the motifs once the index is larger than the value stored in the list (step 23).

Single-conformation analysis mode

To analyze a single `.pdb` file, we also provide the `MINT` web server at `http://mint.cent.uw.edu.pl`. Either an all-atom `.pdb` file or a PDB code can be submitted. In the latter case, the file with a specified PDB code will be automatically downloaded and protonated using `Reduce` (39). The user can further download the output files and visualize the secondary and tertiary structures colored by the computed descriptors analogous to the ones shown in Figure 6.

Trajectory analysis mode

For many RNA or DNA conformations, e.g. from the trajectory files, every frame is characterized as previously described. The main output is an `.xls` file listing all base pairs, helices, loops, junctions, nucleotide triples, pseudoknots, etc., with their topologies and participating nucleotides, as well as the frame numbers in which these motifs were detected.

Clustering. To describe the changes in the RNA secondary structure during dynamics, the detected motifs are clustered. Clustering is parameterized with two user-defined parameters: the minimal percentage of frames in which the motif has to be present to be classified and the minimal percentage of similarities between the two motifs to belong to the common cluster.

In the first step rare motifs are removed by filtering them according to the frequency of occurrence. Second, the motifs' distance matrix is computed. The distance between motifs is defined as the number of their common nucleotides. The order of the nucleotides is not taken into account. Third, the motif with the longest list of partners is incorporated to the first cluster. Next, the second longest is chosen and so on. The motifs used in the first cluster and the sequential created clusters are deleted from the list—a motif can be present only in one cluster.

Dynamical propagation of the secondary structure. To detect correlations in the propagation of the secondary structure for every nucleotide, we compute a ϕ correlation coefficient defined as

$$\phi = \frac{n_{11}n_{00}}{\sqrt{n_{\bullet 1}n_{\bullet 0}n_{0\bullet}n_{1\bullet}}} \quad (3)$$

where

- n_{11} is the number of frames in which both nucleotides form a WC-edge pair, analogously n_{00} is the number of frames in which none of the nucleotides forms a WC-edge pair.
- in the denominator $n_{\bullet 1} = n_{11} + n_{01}$, $n_{1\bullet} = n_{11} + n_{10}$, $n_{\bullet 0} = n_{00} + n_{10}$, $n_{0\bullet} = n_{00} + n_{01}$.
- n_{01} is the number of frames in which the first nucleotide forms a WC-edge pair and the second nucleotide does not, analogously n_{10} is the number of frames in which the first nucleotide is WC-edge-paired and the second one is not. Note, the numbering in `Python` starts from 0 index.

The ϕ coefficient ranges from -1 to 1 so ϕ close to 0 suggests no correlation. A symmetric matrix with ϕ values is outputted both as a text file and heat map.

MD simulations of small ribosomal subunit

System preparation. The crystal structure of the *Escherichia coli* 30S subunit resolved with the 3.0 Å resolution (PDB code 4V9D) was taken as the starting conformation (40). This structure had the best resolution and longest 16S rRNA among the ribosome structures deposited in PDB (as of June 2012). The tRNAs were removed but the crystal waters and divalent ions were kept. Hydrogen atoms were added and the system was solvated with explicit waters extending at least 15 Å from any atom of the solute. Counterions and excess ions were added to achieve 0.15 M concentration of NaCl. These preparatory steps were performed with the `VMD` (22) program and CHARMM36 force field (41). The simulated system consisted of almost 866 000 atoms and trajectories were generated using `NAMD` (42).

Particle Mesh Ewald method, SHAKE algorithm with a time step of 2 fs and periodic boundary conditions were used. The cutoff parameter for the VDW and electrostatic interactions was set to 12 Å, the switching distance to 15 Å and pair list distance to 18 Å.

Simulation protocol. First, the solvent was energy minimized, second it was gradually heated, with the solute constrained. The temperature was increased from 30 K, increasing 10 K every 100 steps up to 300 K. Third, the system was equilibrated at 300 K in two phases. The constraints on solute atoms were gradually decreased in the following steps of 0.1 ns runs using the force constants: (i) 25 → 1 kcal/mol, (ii) 1 → 0.0076 kcal/mol, (iii) 0.0075 → 0.0042 kcal/mol, (iv) 0.0042 → 0.00167 kcal/mol. Next, unconstrained equilibration was performed for ∼35 ns at 300 K. It followed by 30 ns of the production, which was analyzed with `MINT`.

In the production phase the average root-mean-square deviation from the starting structure for heavy atoms was 4.2 ± 0.8 Å and the radius of gyration increased from the starting value of 66.2 Å to an average of 67.7 ± 0.1 Å (Supplementary Figure S2).

RESULTS

We present `MINT` functionality based on the analysis of a 16S rRNA fragment (nucleotides 500–545, encompassing helix 18) from the 30 ns MD production trajectory of the

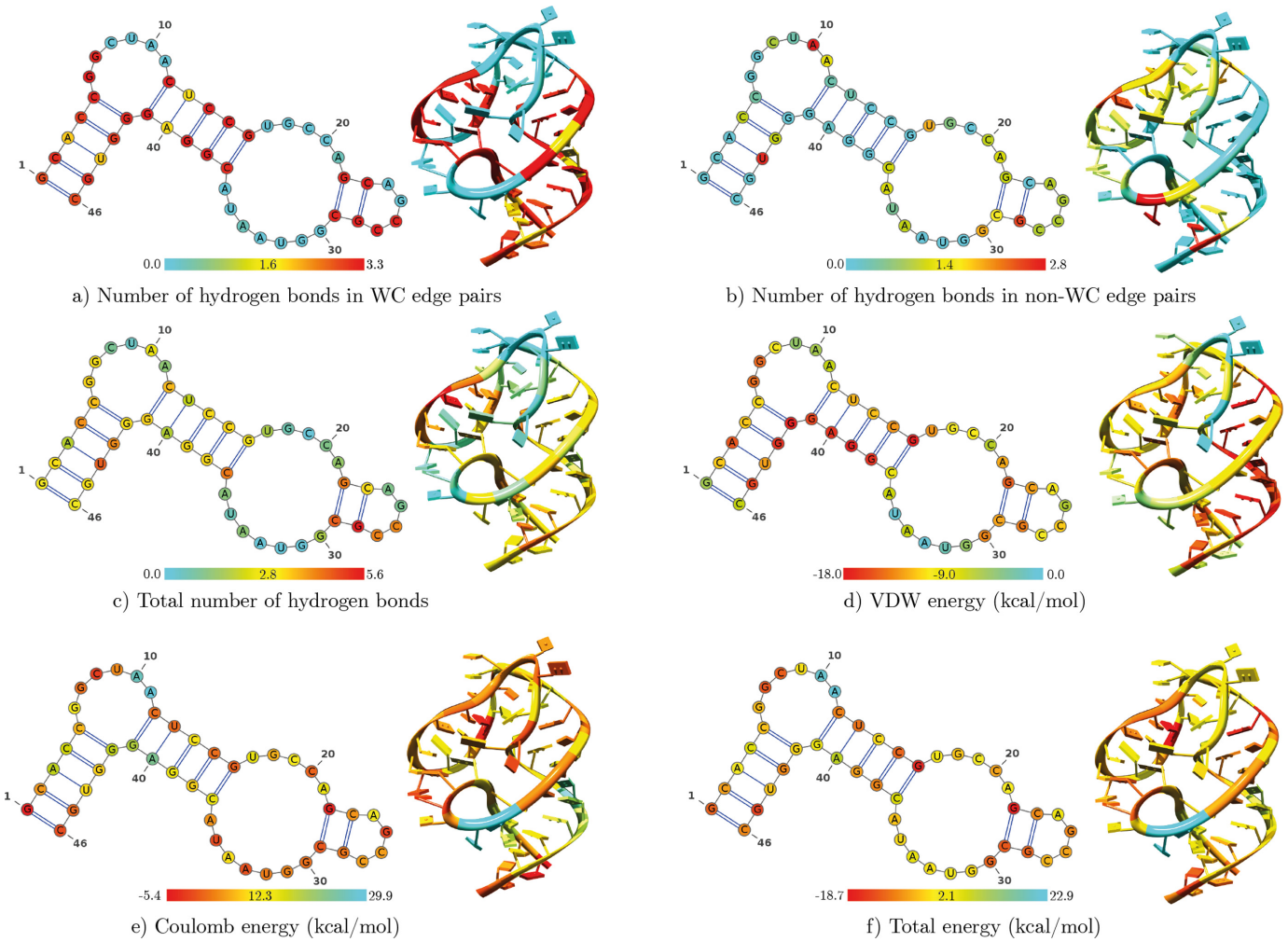

Figure 6. Secondary and tertiary structures of a 16S rRNA fragment (nucleotides 500–545) colored based on various descriptors calculated per nucleotide and averaged over the trajectory.

30S subunit (Supplementary Figure S3). The MINT output files comprehensively describe RNA conformations at both secondary and tertiary levels.

Nucleotide surrounding and interactions

The main MINT output provides a general overview of the contacts in the conformation set. It lists all nucleotides with the number of WC- and/or non-WC-edge hydrogen bond pairs formed by a particular nucleotide, as well as stacking interactions: Coulomb, VDW, their sum and averages over all frames. This output, whose fragment is listed in Table 1, helps detect unusual hydrogen bonding patterns, e.g. with the high average number of non-WC-edge bonds. Whole output listing also the average VDW and Coulomb energies is shown in Supplementary Table S1.

MINT allows projecting the computed descriptors on tertiary and secondary structures. It creates six .pdb files in which the temperature factor column for each nucleotide is replaced with, respectively, the number of WC-edge, non-WC-edge hydrogen bonds, and their sum, Coulomb, VDW energy and stacking energy. Therefore, one may view the structure in, e.g. VMD (22) or Chimera (23) and produce images analogous to the ones shown in Figure 6. To annotate average secondary structures and create images, MINT uses the VARNA (21) visualization applet. Examples of secondary structure graphics for helix 18 are shown in Figure 6 and for the entire 3′ major domain of 16S RNA in Supplementary Figure S4.

MINT also lists statistics of nucleotide contacts giving the percentage of frames in which a given pair, triple or other pattern occurred. An example is listed in Table 2 and shown in Figure 7.

Classification of pairs

The nucleic acid structure is characterized with a list of pairs described according to the edge-to-edge (Figure 2) classification (20), geometric isomerism and percentage of frames a certain pair occurred in the trajectory along with frame numbers. Table 3 shows a fragment of the output listing nucleotide pairs.

Table 1. Average numbers of hydrogen bonds (hbonds) with standard deviations observed in a trajectory for WC-edge and non-WC-edge pairs

Nucleotide no.	WC hbonds	Non-WC hbonds	Nucleotide no.	WC hbonds	Non-WC hbonds
G527	3.2±0.6	2.4±0.7	A532	0.0±0.0	0.5±0.5
C528	3.1±0.5	1.5±0.5	A533	0.2±0.7	1.1±0.2
G529	0.0±0.0	1.8±0.4	U534	0.0±0.0	0.4±0.5
G530	0.0±0.1	0.0±0.1	A535	0.0±0.0	1.2±0.4
U531	0.0±0.0	0.2±0.5	C536	3.1±0.5	1.2±0.5

Table 2. Nucleotide hydrogen contacts observed in a trajectory

Nucleotide	Hydrogen bond contacts	
G527	C522 A535: 87%	
C528	G521 A535: 74%	
G529	G517 C519 A520: 39%	C519 A520: 37%
G530	no contatcs: 100%	
U531	no contatcs: 83%	
A532	no contatcs: 51%	U516 A520: 15%
A533	U516 A520 A535: 22%	
U534	no contacts: 57%	C511: 20% U512: 16%
A535	G527 C528 A533: 48%	G527 C528: 18%
C536	G515 G521: 43%	

Each row lists: the nucleotide type with its number, the nucleotides it hydrogen bonds with, the percentage of trajectory time these contacts were formed. The 'no contacts' denote the percentage of frames the nucleotide did not create any hydrogen bonds. Only the contacts present in more than 10% of the trajectory are listed so the contacts in a row do not have to sum up to 100%. For example, A535 for 48% of trajectory time contacts G527, C528 and A533, but 18% of time only G527 and C528 (Figure 7). This dynamic interaction is revealed in MD—in the crystal A535 pairs only with U516. For full output which includes also the number of hydrogen bonds see Supplementary Table S2.

Table 3. The `MINT` output listing nucleotide pairs

Nucleotide pair	Interacting edges	Configuration	% of frames
G527/A535	Sugar/WC	*trans*	81%
G527/A535	Sugar/Sugar	*trans*	15%
C528/A535	Sugar/Sugar	*trans*	97%
C528/A533	Sugar/Sugar	*cis*	13%
U531/G1207	Sugar/Sugar	*cis*	16%
A532/U1056	WC*Hoogsteen/WC*Sugar	*cis*	20%
A532/A1055	WC*Hoogsteen/Sugar	*trans*	17%
A532/A1055	WC*Hoogsteen/Sugar	*cis*	16%
A533/A535	Sugar/Sugar	*trans*	55%
A533/C536	Sugar/WC*Hoogsteen	*trans*	22%

Each pair is classified by interacting nucleotide edges, configuration, and the percentage of trajectory frames in which the pair was detected. The * notation is used if only one hydrogen bond between nucleotides is found, involving corner atoms, and it is impossible to assign the edge uniquely.

Table 4. List of stacked bases along with their trajectory averaged stacking energy and the percentage of frames the interaction was present

Stacked bases	Average Coulomb energy (kcal/mol)	Average VDW energy (kcal/mol)	Average stacking energy (kcal/mol)	% of frames
G527/C528	−6.8±1.3	−5.5±0.5	−12.2±1.4	100%
C528/G529	−2.8±1.1	−3.9±0.5	−6.7±1.2	99%
C528/A533	6.8±0.9	−0.6±0.1	6.2±0.9	28%
G529/A533	2.3±0.3	−0.6±0.1	1.8±0.3	22%
A532/G1206	1.9 ±0.8	−1.1±0.5	0.8±0.9	85%
A532/A1055	15.1±2.6	−1.1±0.4	14.1±2.6	40%
A532/U1056	5.5±2.5	−0.9±0.3	4.7±2.5	22%
A533/C536	6.7±1.9	−5.7±0.6	0.9±2.0	77%

A nucleotide can show up several times because it may create alternative stacking interactions with different nucleotides.

Stacking interactions

The stacking energy is presented as the sum of the Coulomb and VDW energies between the bases of two nucleotides with an energy threshold for stacked bases on the VDW energy (see Materials and Methods). An example of the output is shown in Table 4.

An analogous table is constructed for the anion–π contacts — their energy is calculated as the sum of the Coulomb and VDW energies between a base and oxygen atom (Supplementary Table S3).

Table 5. A fragment of the MINT output with a list of secondary structure motifs and their occurrence

Cluster	Motif		Motif forming nucleotides	% of frames
No.	No.	Code		
0	0	4	C522-G527-C526-C525-G524-A523-C522-	100%
1	1	0–6	C504-G541-G540-C511-A510-A509-U508-C507-G506-G505-C504-	99%
2	2	7–5	G515-C536-A535-U534-A533-A532-U531-G530-G529-C528-G521-A520-C519-C518-G517-U516-G515-	88%
	3	4–0	A520-A533-A532-U531-G530-G529-C528-G521-A520-	12%
	4	2–4	G515-C536-A535-U534-A533-A520-C519-C518-G517-U516-G515-	12%

A motif is labeled according to classification shown in Supplementary Figure S1.

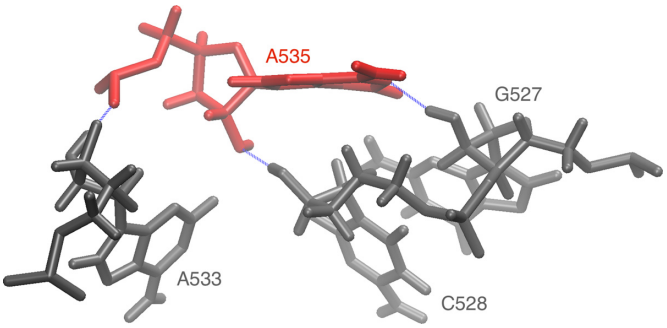

Figure 7. Prevalent hydrogen bonding pattern of A535. The conformation is from the 9.49 ns frame with hydrogen bonds marked as blue dashed lines. Nucleotides are listed in Table 2. The structural context of this arrangement is shown and explained in Supplementary Figure S5.

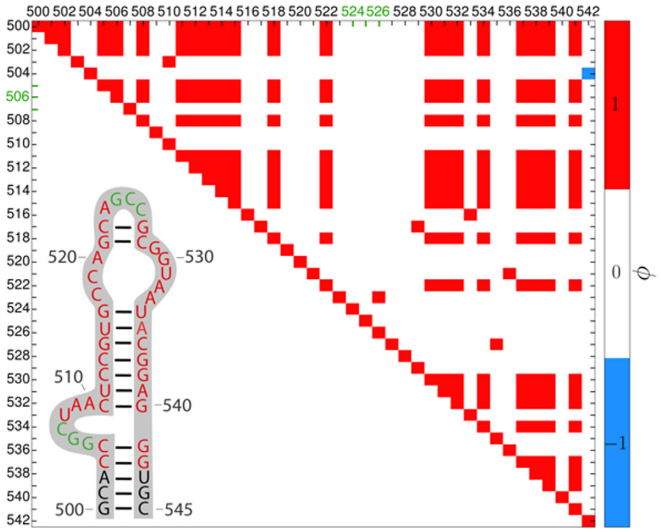

Figure 8. Heat map of the ϕ correlation coefficient for the 503–542 16S rRNA region from the MD simulation. The inset shows the secondary structure (43) with nucleotides creating a pseudoknot in green. Axes labels stand for nucleotide numbers. The ϕ coefficients larger than +0.4 (the `cutoff` for the color scale is defined by the user) are in red, lower than −0.4 in blue and the rest is in white. Nucleotides correlate with themselves so the diagonal is red. Every paired nucleotide and its neighboring pairs have ϕ above 0.4, indicating positive correlations in accord with their synchronous movement.

Dynamical correlations within hydrogen bonding patterns

Based on the contacts that each nucleotide forms in each frame, we computed the ϕ coefficient. Figure 8 shows its illustration as a heat map. The uncorrelated regions (ϕ close to 0) characterize nucleotides which either do not form pairs or form a strong pair that never opens in the course of the dynamics. Therefore, the heat map does not characterize the secondary structure (43) but rather the mobile parts of the structure. The synchronous movement of two nucleotides is indicated by positive ϕ (the same pair opens and closes). This heat map should be analyzed taking into account the pseudoknot in this fragment (not seen in a simple 2D representation) whose 3D structure is described in Supplementary Figure S6. A negative correlation occurs for an asynchronous movement, i.e. if a nucleotide is paired ('closed') while the other is open and conversely. A trajectory-based example for nucleotides 504 and 542, in which the G542 base 'jumps' between two others, is shown in Supplementary Figure S7. Overall, the heat map is useful while searching for the non-obvious structural blocks.

Dynamics of structural motifs

For every frame MINT creates a dot-bracket representation of the secondary structure. Next, it generates an `.xml` file, which can be further imported in RNAMovies (24) to visualize the evolution of the secondary structure in a trajectory (Supplementary Figure S8).

All detected motifs and their classification are written to a separate sheet whose fragment is listed in Table 5. The motifs are further clustered. A cluster contains various secondary structure motifs appearing in the same RNA region so it describes the structural flexibility of the given fragment (Figure 9).

Next, for a given cluster MINT computes an average motif by taking the longest motif from the cluster and annotating it based on the average WC- and non-WC-edge hydrogen bonds. Such a list is used to identify active nucleotides (which form and break hydrogen bonds during the simulation) as presented in Supplementary Table S4.

In summary, MINT output consists of an `.xls` file with separate sheets (and, if requested, the `.csv` files), `.pdb` files for structure visualization, `.png` files with secondary structures colored by the hydrogen bonding and stacking energy, an `.xml` input file to be readily used in RNAMovies and a simple text file with a detailed description of the RNA or DNA structure in each conformation frame. Supplementary Figure S9 presents MINT benchmarks and performance depending on the RNA chain length and CPUs.

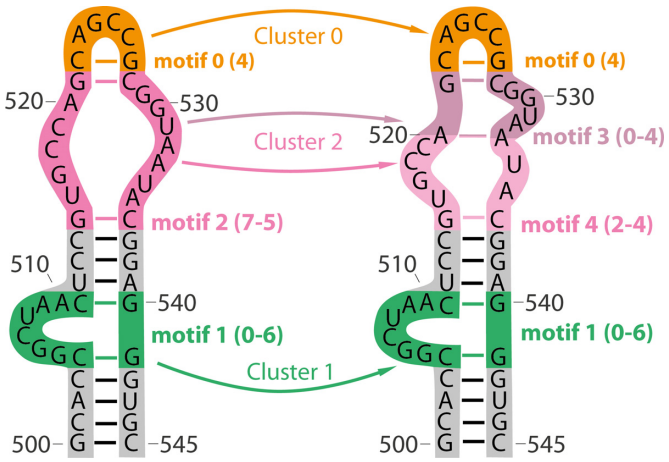

Figure 9. Schematic representation of clustering of motifs listed in Table 5. Cluster 0 consists of motif 0 with code 4, cluster 1 of motif 1 with codes 0–6, and cluster 2 of motifs no. 2, 3 and 4. For motif codes see Supplementary Figure S1. The scheme presents two secondary structures of the same RNA fragment that was simulated. Motifs engaging the same nucleotides fall in the same clusters.

CONCLUSION

We characterized a tool for post-processing MD-derived trajectories or other large conformation sets of nucleic acid molecules. `MINT` calculates various descriptors for nucleic acid structures, provides their time evolution and facilitates their visualization. `MINT` lists the statistics of these descriptors to enable relating them with function, e.g. experimental mutational analyses listed in Supplementary Table S5 for ribosomal RNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

University of Warsaw [CeNT/BST and ICM/KDM/G31-4]; National Science Centre [DEC-2012/05/B/NZ1/00035 to J.T., M.J. and DEC-2011/03/N/NZ2/02482 to M.J.]; European Social Fund [contract no UDA-POKL.04.01.01-00-072/09-00 to M.J.]; Foundation for Polish Science (Team project co-financed by European Regional Development Fund operated within Innovative Economy Operational Programme).
Conflict of interest statement. None declared.

REFERENCES

1. Bloomfield,V.A., Crothres,D.M. and Tinoco,I.J. (1999) *Nucleic acids: structures, properties, and functions*. University Science Books, Sausalito, CA.
2. Šponer,J. and Lankas,F. (2006) *Computational studies of RNA and DNA*. Springer, Dordrecht.
3. Onuchic,J.N. and Sanbonmatsu,K.Y. (2012) Magnesium fluctuations modulate RNA dynamics in the SAM-I riboswitch. *J. Am. Chem. Soc.*, **134**, 12043–12053.
4. Krepl,M., Réblová,K., Koča,J. and Šponer,J. (2013) Bioinformatics and molecular dynamics simulation study of L1 stalk non-canonical rRNA elements: kink-turns, loops, and tetraloops. *J. Phys. Chem. B*, **117**, 5540–5555.
5. Panecka,J., Mura,C. and Trylska,J. (2014) Interplay of the bacterial ribosomal A-site, S12 protein mutations and paromomycin binding: a molecular dynamics study. *PLoS ONE*, **9**, e111811.
6. Whitford,P.C., Onuchic,J.N. and Sanbonmatsu,K.Y. (2010) Connecting energy landscapes with experimental rates for aminoacyl-tRNA accommodation in the ribosome. *J. Am. Chem. Soc.*, **132**, 13170–13171.
7. Nivón,L.G. and Shakhnovich,E.I. (2004) All-atom Monte Carlo simulation of GCAA RNA folding. *J. Mol. Biol.*, **344**, 29–45.
8. Tang,X., Thomas,S., Tapia,L., Giedroc,D.P. and Amato,N.M. (2008) Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, **381**, 1055–1067.
9. Laing,C. and Schlick,T. (2010) Computational approaches to 3D modeling of RNA. *J. Phys. Condens. Matter.*, **22**, 283–101.
10. Jossinet,F., Ludwig,T.E. and Westhof,E. (2010) Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. *Bioinformatics*, **26**, 2057–2059.
11. Lavery,R., Moakher,M., Maddocks,J.H., Petkeviciute,D. and Zakrzewska,K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917–5929.
12. Lu,X.J. and Olson,W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding, and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.*, **3**, 1213–1227.
13. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.J., Neidle,S., Shakked,Z. *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
14. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
15. Yang,H. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
16. Pietal,M.J., Szostak,N., Rother,K.M. and Bujnicki,J.M. (2012) RNAmap2D - calculation, visualization and analysis of contact and distance maps for RNA and protein-RNA complex structures. *BMC Bioinformatics*, **13**, 333.
17. Wale,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.
18. Lu,X.J., Olson,W.K. and Bussemaker,H.J. (2010) The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex. *Nucleic Acids Res.*, **38**, 4868–4876.
19. Kumar,R. and Grubmüller,H. (2015) do_x3dna: a tool to analyze structural fluctuations of dsDNA or dsRNA from molecular dynamics simulations. *Bioinformatics*, doi:10.1093/bioinformatics/btv190.
20. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
21. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: interactive drawing and editing of the secondary structure. *Bioinformatics*, **25**, 1974–1975.
22. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.
23. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.
24. Evers,D. and Giegerich,R. (1999) RNA Movies: visualizing RNA secondary structure spaces. *Bioinformatics*, **15**, 32–37.
25. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F. and Wilczynski,B.E.A. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
26. Michaud-Agrawal,N., Denning,E.J., Woolf,T.B. and Beckstein,O. (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, **32**, 2319–2327.
27. Lescoute,A. and Westhof,E. (2006) The interaction networks of structured RNAs. *Nucleic Acids Res.*, **34**, 6587–6604.
28. Carter,A.P., Clemons,W.M., Brodersen,D.E., Morgan-Warren,R.J., Wimberly,B.T. and Ramakrishnan,V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340–348.

29. Šponer,J., Gabb,H.A., Leszczynski,J. and Hobza,P. (1997) Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study. *Biophys. J.*, **73**, 76–87.
30. Šponer,J., Leszczynski,J. and Hobza,P. (1996) Nature of nucleic acid–base stacking: nonempirical ab initio and empirical potential characterization of 10 stacked base dimers. Comparison of stacked and H-bonded base pairs. *J. Phys. Chem.*, **100**, 5590–5596.
31. Šponer,J., Leszczynski,J. and Hobza,P. (1995) Base stacking in cytosine dimer. A comparison of correlated ab initio calculations with three empirical potential models and density functional theory calculations. *J. Comput. Chem.*, **17**, 841–850.
32. Wang,J., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
33. Mackerell,A.D. and Banavali,N.K. (2000) All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.*, **21**, 105–120.
34. Foloppe,N. and Mackerell,A.D. (2000) All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.*, **21**, 86–104.
35. Machnicka,M.A., Milanowska,K., Osman Oglou,O., Purta,E., Kurkowska,M., Olchowik,A., Januszewski,W., Kalinowski,S., Dunin-Horkawicz,S. and Rother,K.M.E.A. (2013) MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Res.*, **41**, D262–D267.
36. Cantara,W.A., Crain,P.F., Rozenski,J., McCloskey,J.A., Harris,K.A., Zhang,X., Vendeix,F.A., Fabris,D. and Agris,P.F. (2011) The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res.*, **39**, 195–201.
37. Aduri,R., Psciuk,B.T., Saro,P., Taniga,H., Schlegel,H.B. and SantaLucia,J. (2007) AMBER force field parameters for the naturally occurring modified nucleosides in RNA. *J. Chem. Theory Comput.*, **3**, 1464–1475.
38. Smit,S., Rother,K., Heringa,J. and Knight,R. (2008) From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal. *RNA*, **14**, 410–416.
39. Word,J.M., Lovell,S.C., Richardson,J.S. and Richardson,D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
40. Dunkle,J.A., Wang,L., Feldman,M.B., Pulk,A., Chen,V.B., Kapral,G.J., Noeske,J., Richardson,J.S., Blanchard,S.C. and Cate,J.H. (2011) Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. *Science*, **332**, 981–984.
41. Denning,E.J., Priyakumar,U.D., Nilsson,L. and MacKerell,A.D. Jr (2011) Impact of 2-hydroxyl sampling on the conformational properties of RNA: update of the CHARMM all-atom additive force field for RNA. *J. Comput. Chem.*, **32**, 1929–1943.
42. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L. and Schulten,K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
43. Noller,H.F. and Woese,C.R. (1981) Secondary structure of 16S ribosomal RNA. *Science*, **212**, 403–411.