The role of nucleobase interactions in RNA structure and dynamics

Sandro Bottaro[*], Francesco Di Palma and Giovanni Bussi[*]

Scuola Internazionale Superiore di Studi Avanzati, International School for Advanced Studies, 265, Via Bonomea I-34136 Trieste, Italy

ABSTRACT

The intricate network of interactions observed in RNA three-dimensional structures is often described in terms of a multitude of geometrical properties, including helical parameters, base pairing/stacking, hydrogen bonding and backbone conformation. We show that a simple molecular representation consisting in one oriented bead per nucleotide can account for the fundamental structural properties of RNA. In this framework, canonical Watson-Crick, non-Watson-Crick base-pairing and base-stacking interactions can be unambiguously identified within a well-defined interaction shell. We validate this representation by performing two independent, complementary tests. First, we use it to construct a sequence-independent, knowledge-based scoring function for RNA structural prediction, which compares favorably to fully atomistic, state-of-the-art techniques. Second, we define a metric to measure deviation between RNA structures that directly reports on the differences in the base–base interaction network. The effectiveness of this metric is tested with respect to the ability to discriminate between structurally and kinetically distant RNA conformations, performing better compared to standard techniques. Taken together, our results suggest that this minimalist, nucleobase-centric representation captures the main interactions that are relevant for describing RNA structure and dynamics.

INTRODUCTION

Ribonucleic acid (RNA) plays a fundamental role in a large variety of biological processes, such as enzymatic catalysis (1), protein synthesis (2) and gene regulation (3). RNA molecules fold in a well-defined three-dimensional structure dictated by their nucleotide sequence (4), that can be determined by means of X-ray crystallography or nuclear magnetic resonance experiments (5). Despite recent important advances in the field, RNA structure determination is still a complex and expensive procedure. Computational/theoretical models, and in particular structure prediction methods, ideally complement experiments, as they can in principle provide the tertiary, native structure for a given RNA sequence.

Although traditionally considered simpler than the protein folding problem (4), *de novo* prediction of RNA three-dimensional structure still represents a challenging task (6). Atomistic molecular dynamics methods have been so far limited to the predictive folding of very small systems (7,8), whereas coarse-grained, physics-based models have proven useful in folding larger RNA structures (9,10). However, the most popular and successful RNA structure prediction methods are the so-called knowledge-based techniques, that typically employ fragment libraries for constructing RNA structures. These fragments are either combined using secondary structure information (11,12) or used to generate a large number of plausible, putative structures that are subsequently ranked according to a scoring function (13–17). Knowledge-based methods often rely on two assumptions: first, that the main structural features of RNA molecules can be described in terms of few relevant observables. Second, that the distribution of these observables in a data set of experimental structures displays specific features, that can be used for generating and/or scoring three-dimensional structures (18). A crucial issue is the choice of the aforementioned observables. While backbone atoms and torsional angles are the most natural choice for representing protein structures (19), RNA molecules are more difficult to treat, due to the presence of a larger number of flexible backbone degrees of freedom and because of the important structural role of base–base interactions. For this reason RNA molecules are often described in terms of several structural observables, including backbone dihedrals (20,21), pseudo dihedrals (22), helical parameters (23), hydrogen-bond networks and stacking interactions (24,25).

The proper choice of simplified, insightful descriptors for RNA molecules is not limited to the development of structure prediction methods, but is a general and relevant is-

[*]To whom correspondence should be addressed. Tel: +39 040 3787 407; Fax: +39 040 3787 528; Email: bussi@sissa.it
Correspondence may also be addressed to Sandro Bottaro. Tel: +39 040 3787 301; Fax: +39 040 3787 528; Email: sbottaro@sissa.it

sue in RNA structural analysis, as in the definition of metrics for comparing three-dimensional structures. The well-known root mean square deviation (RMSD) after optimal superposition (26) is known to be suboptimal in the analysis of RNA structure, as it does not provide reliable information about the differences in the base interaction network (6,27).

This motivated the development of alternative RNA-specific deviation measures based on selected geometrical properties. These measures can be used for comparing observed structures with *a priori* known patterns, as is done for instance in annotation methods (28) and in motif recognition algorithms (29–31). Other RNA specific similarity measures were introduced with the purpose of assessing the quality of RNA structural predictions (6) or for clustering/identifying recurrent structural motifs in large databases of known structures (21,29). The definition of a proper structural deviation measure is also a central issue in the construction of Markov state models (32), that typically assume the geometric similarity to provide an adequate measure of kinetic proximity.

The analysis of RNA three-dimensional structure, the construction of a knowledge-based structure prediction method and the definition of a structural deviation share a common trait: all of them are based on the apparently arbitrary choice of the structural observables used to represent RNA. Is there a simple representation of an RNA molecule that recapitulates its key structural and dynamical properties?

In this Paper we introduce a minimalist representation for RNA molecules consisting in one oriented bead per nucleotide. We first show that this description captures the fundamental base-pair and base-stacking interactions observed in experimental structures. Using this representation we define a simple sequence-independent scoring function for RNA structure prediction, which performs on par or better compared to existing atomistic techniques. Based on the same molecular description, we then introduce a metric for calculating structural deviation. When compared with the standard RMSD measure, this metric correlates better with the physical time required to explore the conformational space in molecular dynamics simulations. As a further proof of the capability of this metric to distinguish relevant structural differences, we show that it can be successfully used for recognizing local structural motifs, reproducing results obtained with state-of-the-art techniques. Finally, in the Discussion section, we summarize the analogies between these applications and we proceed with an extensive comparison with existing techniques. The software for performing the calculations is freely available at https://github.com/srnas/barnaba.

METHODS

This section is structured as follows: first, we introduce a simple representation for RNA three-dimensional structures that has an intuitive interpretation in terms of base-stacking and base-pairing interactions. We then use this molecular description to define i) a scoring function for RNA structure prediction ($\mathcal{E}$SCORE) and ii) a metric

Table 1. Number of base–base interactions detected in the crystal structure of the *H. marismortui* large ribosomal subunit

Interaction Type	Occurrences[a]	$\tilde{r} < \sqrt{2.5}$[b]
Stacking	2328	2318
Watson-Crick	723	723
Non-Watson-Crick (base–base)	399	382
Base-sugar/phosphate interactions	781	398
Not annotated	—	1214

[a]Number of interactions detected by MC-Annotate (25).
[b]Base pair (j,k) is counted when both $\tilde{r}_{jk}$ and $\tilde{r}_{kj} < \sqrt{2.5}$.

for calculating deviations between RNA three-dimensional structures ($\mathcal{E}$RMSD).

Molecular representation

We construct a local coordinate system in the center of the six-membered rings, as shown in Figure 1a. Following this definition, the relative position and orientation between two nucleobases is described by a vector **r**, that is conveniently expressed in cylindrical coordinates ρ, θ and z (Figure 1b). Note that **r** is invariant for rotations around the axis connecting the six-membered rings. We highlight that this definition is similar to the local referentials introduced by Gendron and Major (25). The use of a nucleotide-independent centroid makes it straightforward to compare and combine collection of position vectors deriving from different combinations of nucleobases. This is of particular importance for constructing the knowledge-based scoring function (see below). The position vector **r** has an intuitive interpretation in terms of base-stacking and base-pairing interactions. This aspect is illustrated in Figure 1c, that shows the distribution of **r** vectors for all neighboring bases in the crystal structure of the *Haloarcula marismortui* large ribosomal subunit (PDB code 1S72) (2) projected on the ρ and z coordinates. In the figure, different colors correspond to different types of interactions detected by MC-annotate. Due to steric hindrance, no points are observed in a forbidden ellipsoidal region. Furthermore, almost all the base-stacking and base-pairing interactions ($\approx$99.6%) belong to a well-defined ellipsoidal shell. It is therefore useful to introduce the anisotropic position vector

$$\tilde{\mathbf{r}} = \left(\frac{r_x}{a}, \frac{r_y}{a}, \frac{r_z}{b}\right) = \left(\frac{\rho}{a}\cos\theta, \frac{\rho}{a}\sin\theta, \frac{z}{b}\right) \quad (1)$$

with $a = 5$ Å and $b = 3$ Å, so that pairs of bases in the interaction shell are such that $1 < \tilde{r} < \sqrt{2.5}$. The majority of base–base contacts lying in this interaction shell are annotated either as Watson-Crick/non-Watson-Crick or as base stacking, as detailed in Table 1. Within this region we distinguish a *pairing zone* and a *stacking zone*, according to the type of featured interactions. The tri-modal histogram in Figure 1c shows that these two zones can be defined without ambiguity considering pairs such that the projection of **r** along the z axis is larger (stacking) or smaller (pairing) than 2 Å.

It is well known that the strength and nature of pairing and stacking interactions depend on the base–base distance, on the angle θ as well as on other angular parameters (e.g.
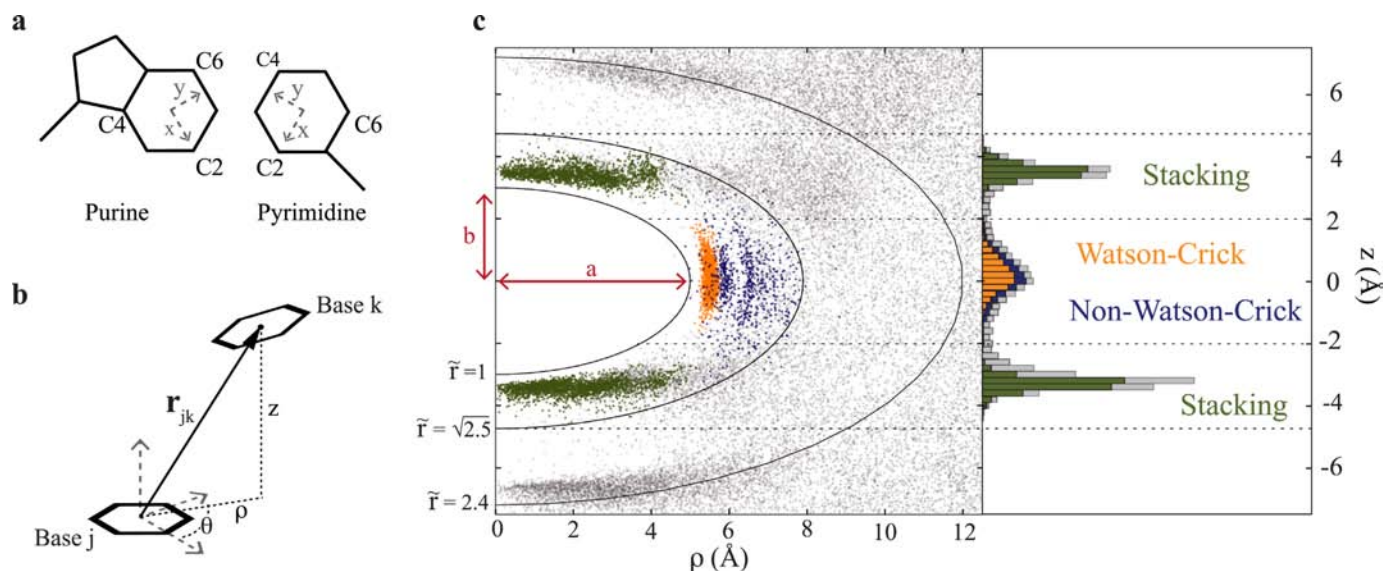
Figure 1. (a) Definition of the local coordinate system for purines and pyrimidines. (b) The vector $\mathbf{r}_{jk}$ describes the position of base ring k in the coordinate system constructed on base ring j and is expressed in cylindrical coordinates ρ, θ, z. (c) Observed positions of neighboring bases in the crystal structure of the *H. marismortui* large ribosomal subunit projected on the z–ρ plane. Isolines of the ellipsoidal distance $\tilde{r}$ are shown. Each point is colored according to the interaction type: canonical Watson-Crick pairs in orange, base–base non Watson-Crick pairs in blue and stacking in green. Base-sugar, base-phosphate and non-annotated pairs are shown in light gray. On the right, empirical distribution along the z coordinate obtained considering points with $\tilde{r} < \sqrt{2.5}$. Pairing and stacking regions can be unambiguously identified ($|z| \leq 2.0$ Å and $|z| > 2.0$ Å, respectively).

twist, roll, tilt) in a non-trivial manner (24,33). Such dependence can be observed in Figure 2, where the points belonging to the pairing and stacking zone of Figure 1c are projected on two separate ρ–θ planes. These distributions give an average picture containing contributions from different base pair types (purine–purine, purine–pyrimidine and pyrimidine–pyrimidine) and with weights dictated by the employed data set. Nevertheless, the observations below hold also when considering the 16 possible combinations of base pairs individually and different data sets (see Supporting Data (SD) Figrues S1–S3). In the pairing zone (Figure 2, left panel) we first observe a dominant peak centered around ($\rho = 5.6$ Å, $\theta = 60°$), corresponding to the position of canonical Watson-Crick base pairs as well as wobble (GU) base pairs. The two other peaks correspond instead to base pairs interacting through the Hoogsteen or sugar edge (13). One can also appreciate the absence of bases in the region occupied by the sugar ($190° < \theta < 290°$). The probability distribution in the stacking zone (Figure 2, right panel) shows a broad peak in the proximity of the origin and extending up to $\rho \approx 4$ Å, which can be compared to the typical radius of the six-membered ring (≈ 1.4 Å). This means that partial or negligible ring overlap is very frequent in RNA structures, as also observed in a seminal paper by Bugg *et al.* (34). This feature is more evident in pyrimidine–pyrimidine and purine–purine pairs, for which high overlap is the exception rather than the rule (see Supplementry Figure S3), whereas overlap is systematically observed in pyrimidine–purine pairs. The fact that bases in the stacking zone are very often 'imbricated,' similarly to roof tiles, rather than literally stacked one on top of the other, does not imply that they are not interacting. Indeed, base–base interaction is not limited to π–π stacking but also includes electrostatic

effects, London dispersion attraction, short range repulsion as well as backbone-induced effects (35).

Scoring function

The empirical distribution of all **r** vectors observed in experimental RNA structures displays very specific features, as described above. We use these geometrical propensities to construct a knowledge-based scoring function for RNA structure prediction. More precisely, we define a function of the atomic coordinates (in this case of the **r** vectors in a molecule) that quantifies the compatibility of a given RNA 3D conformation with respect to the expected distribution observed in native RNA structures. This scoring function, called $\mathcal{E}$SCORE, is defined as a weighted sum over all pairs of bases in a molecule:

$$\mathcal{E}\text{SCORE} = \sum_{j,k} p(\mathbf{r}_{jk}) \tag{2}$$

Notice that both permutations of the j, k indexes should be included in the sum since the vectors $\mathbf{r}_{jk}$ and $\mathbf{r}_{kj}$ provide two partially independent pieces of information. The weights $p(\mathbf{r})$ are given by the empirical probability distribution of nucleobases in the crystal structure of the *H. marismortui* large ribosomal subunit. With this definition, a lower weight is assigned to structures with pairs of bases in relative positions not compatible with the training set, including those with steric clashes. The probability distribution $p(\mathbf{r})$ is calculated considering only pairs of bases within the interaction shell ($\tilde{r} < \sqrt{2.5}$) and using a Gaussian kernel density estimation with bandwidth = 0.25 Å. Note that $p(\mathbf{r})$ does not depend on the identity of the nucleotides, and the RNA molecule is therefore treated as a homopolymer. The sum of the probabilities is used instead
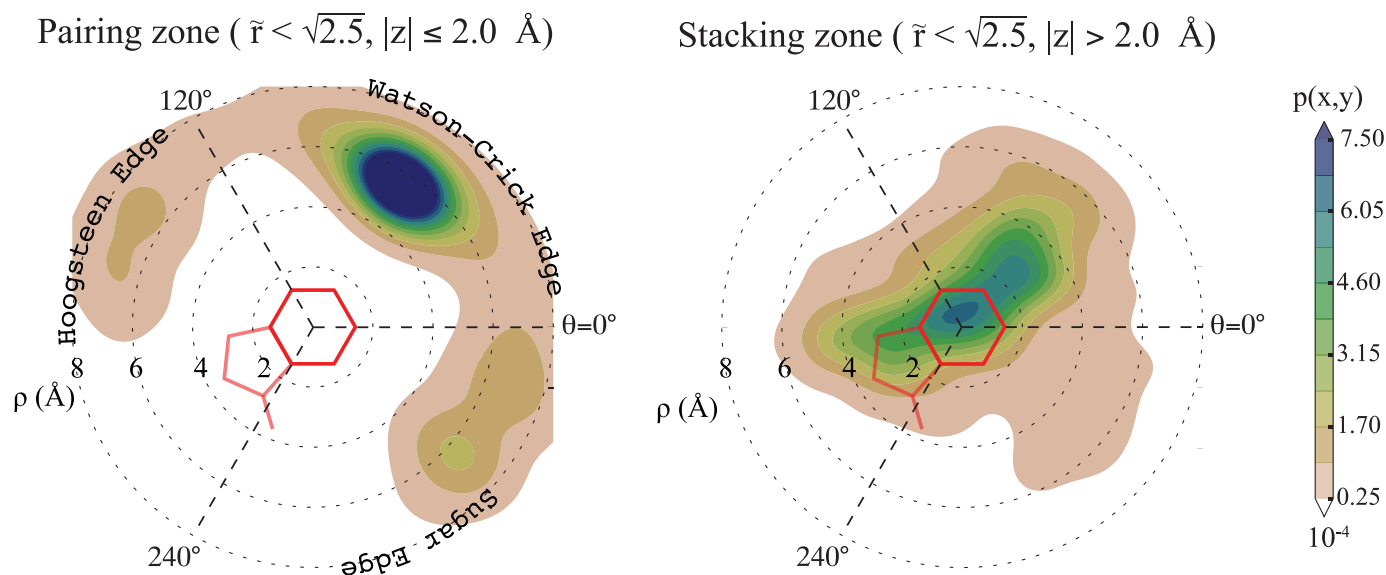
Figure 2. Empirical density distributions of neighboring nucleobases ($\tilde{r} < \sqrt{2.5}$), obtained by projecting points belonging to the pairing and stacking zone of Figure 1 on the θ–ρ plane. Densities are computed with a uniform binning in Cartesian coordinates and using a Gaussian kernel density estimation with bandwidth 0.25 Å. Hoogsteen, Watson-Crick and Sugar edges are indicated. The six-membered and five-membered (for purine only) rings are sketched in red.

of their product, in order to decrease the impact of low-count regions on the scoring function. This procedure has been first introduced in the field of hidden Markov modeling for speech recognition (36) and has a formal justification known as 'summation hack' (T. Minka, The 'summation hack' as an outlier model, http://research.microsoft.com/en-us/um/people/minka/papers/minka-summation.pdf). A non-redundant set of high-resolution structures (16) was also employed as a training data set for the $\mathcal{E}$SCORE, producing similar results.

Structural deviation

The collection of the scaled vectors $\tilde{r}$ (see Eq. 1) in an RNA molecule provides information about the relative base arrangement. One could thus define a metric for measuring the distance between two RNA structures, α and β, as

$$d = \sqrt{\frac{1}{N} \sum_{j,k} |\tilde{\mathbf{r}}_{jk}^{\alpha} - \tilde{\mathbf{r}}_{jk}^{\beta}|^2} \qquad (3)$$

Similarly to Eq. 2, also here both permutations of the j, k indexes should be included in the sum. The d quantity is highly non-local, because large differences in distant pairs give large contributions to the sum. To remove this effect it is convenient to introduce a cutoff distance $\tilde{r}_{\text{cutoff}}$ in the same spirit as we did for the scoring function discussed above. However, we notice that a cutoff would introduce discontinuities in the metric. A possible solution is to introduce a function that maps the vectors $\tilde{r}$ to new vectors $\mathbf{G}(\tilde{r})$ with the following properties:

(1) $|\mathbf{G}(\tilde{r}^{\alpha}) - \mathbf{G}(\tilde{r}^{\beta})| \approx |\tilde{r}^{\alpha} - \tilde{r}^{\beta}|$ if $\tilde{r}^{\alpha}, \tilde{r}^{\beta} \ll \tilde{r}_{\text{cutoff}}$.
(2) $|\mathbf{G}(\tilde{r}^{\alpha}) - \mathbf{G}(\tilde{r}^{\beta})| = 0$ if $\tilde{r}^{\alpha}, \tilde{r}^{\beta} \geq \tilde{r}_{\text{cutoff}}$.
(3) $\mathbf{G}(\tilde{r})$ is a continuous function.

There is no requirement here on the dimensionality of the $\mathbf{G}$ vector. The above properties are satisfied by the four-dimensional vectorial function

$$\mathbf{G}(\tilde{r}) = \begin{pmatrix} \sin(\gamma\tilde{r})\frac{\tilde{r}_x}{\tilde{r}} \\ \sin(\gamma\tilde{r})\frac{\tilde{r}_y}{\tilde{r}} \\ \sin(\gamma\tilde{r})\frac{\tilde{r}_z}{\tilde{r}} \\ 1 + \cos(\gamma\tilde{r}) \end{pmatrix} \times \frac{\Theta(\tilde{r}_{\text{cutoff}} - \tilde{r})}{\gamma} \qquad (4)$$

where $\gamma = \pi/\tilde{r}_{\text{cutoff}}$ and Θ is the Heaviside step function. We note that the dependence of $\mathbf{G}$ on the direction of $\tilde{r}$ vanishes as $\tilde{r}$ approaches the cutoff value $\tilde{r}_{\text{cutoff}}$. We thus set $\tilde{r}_{\text{cutoff}} = 2.4$, so that for the typical distances between stacked and paired bases ($\tilde{r} < \sqrt{2.5}$, see Figure 1c) the contribution of the $\tilde{r}$ directionality is significant. In-depth discussion on Eq. 4 can be found in SD Text 4. Having defined the mapping function $\mathbf{G}(\tilde{r})$, the $\mathcal{E}$RMSD distance reads

$$\mathcal{E}\text{RMSD} = \sqrt{\frac{1}{N} \sum_{j,k} |\mathbf{G}(\tilde{r}_{jk}^{\alpha}) - \mathbf{G}(\tilde{r}_{jk}^{\beta})|^2} \qquad (5)$$

Note that $\mathcal{E}$RMSD is proportional to the Euclidean distance between $\mathbf{G}$ vectors, which are non-linear functions of the atomic coordinates. As such, $\mathcal{E}$RMSD is positive semidefinite, symmetric, and satisfies triangular inequality. During the developmental stage we tested a non-vectorial version of the $\mathcal{E}$RMSD: instead of using the vectorial function $\mathbf{G}(\tilde{r})$, we considered the scalar, continuous function

$$G_s(\tilde{r}) = (\tilde{r}_{\text{cutoff}} - \tilde{r}) \times \Theta(\tilde{r}_{\text{cutoff}} - \tilde{r}) \qquad (6)$$

which is similar to a contact-map distance (37,38). This scalar version differs from $\mathcal{E}$RMSD when considering structures with four nucleotides or less, while the two quantities are highly correlated when analysing larger structures (Supplementry Figure S5).

RESULTS

Scoring function for RNA structure prediction

We assess the quality of the $\mathcal{E}$SCORE on its capability of recognizing the folded state among a set of decoys (39). We consider here a total of 39 different decoy sets previously used to benchmark other knowledge-based scoring functions: 20 decoy sets generated *de novo* using the FARNA algorithm (13) and 19 additional decoy sets from the work of Bernauer *et al.* (16) obtained by gradual deformation of the native structures. As a measure of performance we use the normalized rank (40), defined as the percentage of decoys scoring better than the native structure. In order to simulate a realistic structure prediction experiment, we additionally report the RMSD from native of the best scoring structure in the decoy set. Because the RMSD has been reported to be a suboptimal indicator of structural proximity for RNA, we also compute the interaction fidelity network (INF), which measures the overlap between annotations in two structures (27), as well as the $\mathcal{E}$RMSD introduced here. Equivalent analyses are performed using two state-of-the-art scoring functions, namely the FARFAR (15) and the all-atom knowledge-based scoring function RASP (17). Both FARFAR and RASP are defined at atomistic resolution. FARFAR combines the low-resolution FARNA potential with explicit terms for solvation and hydrogen bonding, while RASP is a statistical potential based on pairwise interatomic distances. These two scoring functions are among the best available (6) and the only ones for which it was possible to perform automatic scoring of a large set of decoys.

Figure 3 summarizes the results obtained on all the 39 decoy sets. For each method we report the number of decoy sets below the value indicated on the horizontal axis. Thus, the better is a scoring function, the faster the curve reaches the maximum number 39. In general, whereas FARFAR performs well in ranking and RASP in finding the best decoys, $\mathcal{E}$SCORE performs well in both tasks (See also Supplementry Table S6). This is a remarkable result considering that $\mathcal{E}$SCORE employs a coarse-grained representation (one bead per nucleotide) and is not sequence dependent. The scoring results on all RNA decoys, together with a short discussion of the cases for which $\mathcal{E}$SCORE fails to identify correctly the native structure can be found in Supplementry Figures S7 and S8.

RNA structural deviation

The definition of an accurate measure of structural deviation for comparing RNA three-dimensional structures is a relevant problem, as the standard RMSD approach does not report on differences of base-stacking and base-pairing patterns, that are the fundamental, key features of RNA molecules (25). As a striking example, the RMSD between an ideal A-form double helix $_{CCCC}^{GGGG}$ and the mismatched structure $_{CCCC}^{GGGG}$ is 1.9 Å, meaning that the threshold of 4Å that is sometimes employed to consider two structures as similar (8) may not be sufficiently strict. A similar issue arises in kinetic modeling of proteins, where the assumption that geometrically close configurations (in terms of the RMSD distance) are also kinetically close has been shown to be
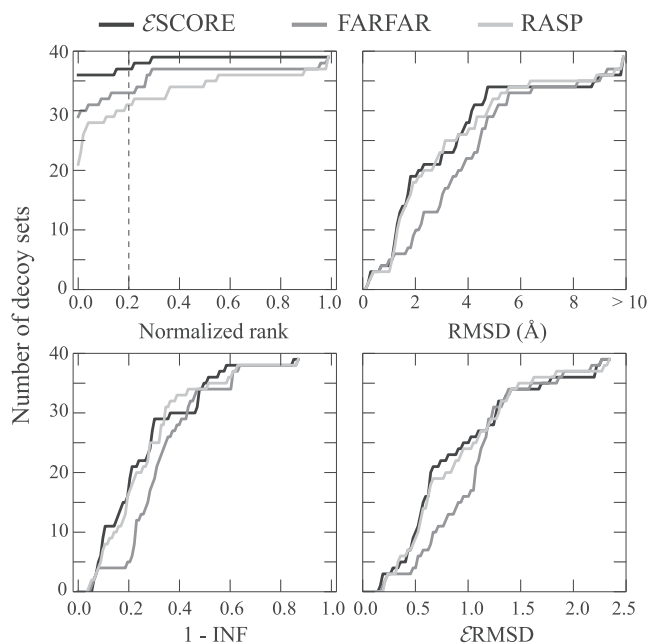

Figure 3. Quantitative comparison of our scoring function $\mathcal{E}$SCORE with FARFAR (15) and RASP (17) on 39 different decoy sets. In the upper left panel, number of decoy sets with normalized rank lower than the value on the horizontal axis. For example, the normalized rank is 0.2 or less for 37 out of 39 decoy sets using the $\mathcal{E}$SCORE, 33/39 using FARFAR and 31/39 using RASP. The faster the curve reaches 39, the better the performance of the scoring function. In the other panels, equivalent plots for RMSD, INF (27) and $\mathcal{E}$RMSD of the best scoring decoy to the native structure, as labeled. See Supplementry Table S6 for a complete list of results.

problematic in discretizing the register shift dynamics of β strands (32).

In order to elucidate the difference between the standard RMSD and the $\mathcal{E}$RMSD, we compute these two quantities on a series of snapshots taken from an atomistic molecular dynamics simulation of a short stem-loop. These structures have been obtained by repeatedly melting and folding (see SD Text 9), and thus include near-native structures, partially folded, misfolded as well as extended conformations. Figure 4 shows that structures with similar RMSD do not necessarily present the same base–base interaction network of the native state: in fact, the secondary structure can be substantially different. Conversely, the $\mathcal{E}$RMSD discriminates structures with near-native base–base contacts ($\mathcal{E}$RMSD <0.8) from structures with non-native base–base interactions ($\mathcal{E}$RMSD>1). Note that in general low $\mathcal{E}$RMSD correspond to high INF values, except for the region around $\mathcal{E}$RMSD=1.5, RMSD=2.7. In this region we observe structures with formed stem and distorted loop (high INF) but also structures with formed loop and incorrect stem (low INF). While Figure 4 serves as an illustrative example, the validity of the $\mathcal{E}$RMSD is more rigorously assessed by performing two tests that probe the accuracy of the method in the analysis of structural and dynamical properties of RNA molecules. More precisely, in the following two subsections we demonstrate i) that the $\mathcal{E}$RMSD is a good measure of kinetic proximity between configurations, performing considerably better compared to standard structural distances and ii) that the $\mathcal{E}$RMSD is highly spe-
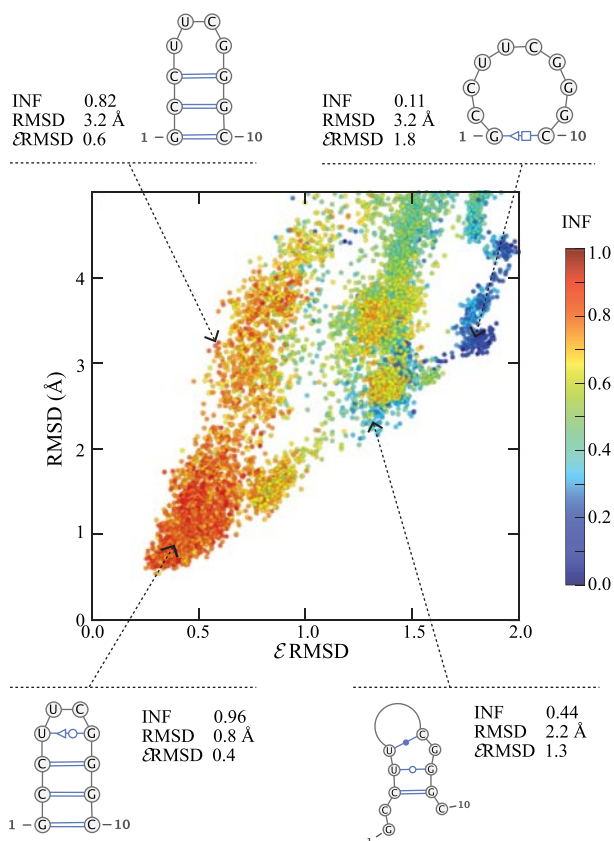
Figure 4. Heavy-atom RMSD from native versus $\mathcal{E}$RMSD of snapshots taken from an atomistic molecular dynamics simulation. The difference in the secondary structure between the native state and samples is calculated using the INF measure, which ranges from 1 (identical secondary structure) to 0 (completely different). Very different secondary structures can be found below 4 Å RMSD. Four selected structures with different values of RMSD, $\mathcal{E}$RMSD and INF are shown, together with their secondary structure. Figures were generated with Varna (48).

cific in recognizing known structural motifs within a large set of crystallographic structures.

Structural and kinetic proximity. The correspondence between a structural deviation (RMSD, $\mathcal{E}$RMSD, etc.) and the kinetic distance (defined as the time needed to evolve from one configuration to the other), can be quantitatively assessed by computing the coefficient of variation (CV).

The CV of a structural deviation d is defined as the ratio of the standard deviation to the mean calculated on structures separated by a time lag τ

$$CV_d(\tau) = \frac{\sigma_d(\tau)}{\langle d(\tau) \rangle} \qquad (7)$$

To a large CV corresponds a large standard deviation, and therefore a high uncertainty between geometric and kinetic distance (41). Note that the standard deviation is normalized with the mean, and the CV is thus scale-invariant. In the approximation that the deviations at fixed time lag are normally distributed, this analysis is equivalent to the histogram overlap discussed in Ref. (38). The connection between kinetic and structural distance is inevitably limited by the fact that the typical time necessary for a transition
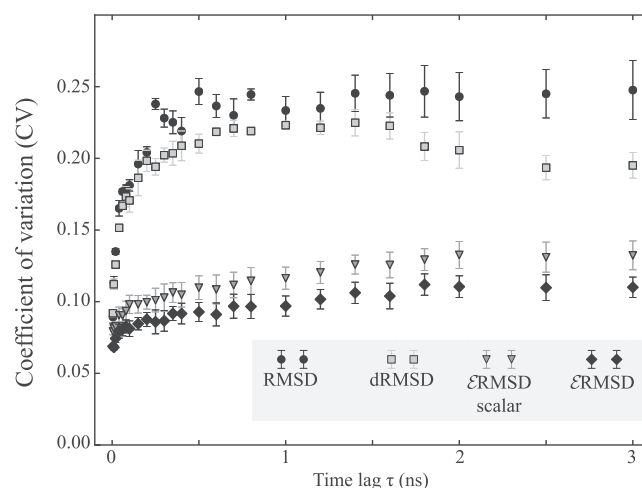

Figure 5. Coefficient of variation calculated on a 100 ns molecular dynamics simulation of the *add* riboswitch. The coefficient of variation for the $\mathcal{E}$RMSD is compared with heavy-atoms RMSD, with the distance RMSD (dRMSD) and with the scalar variant of the $\mathcal{E}$RMSD. Standard deviations are calculated using a blocking procedure.

is in general different from the time necessary for the reversed transition, in contrast with the symmetric definition of structural deviations. Nevertheless, this analysis provides insightful and intuitive information for the purpose of qualitatively comparing different metrics.

In Figure 5 we show the CV as a function of time separation in molecular dynamics simulation of the *add* riboswitch (42,43). The molecular dynamics simulation was initialized in the native state and remained stable throughout the 100 ns run, while experiencing non trivial dynamics such as base-pairing breakage/formation and double-helix breathing. The CV for $\mathcal{E}$RMSD is compared with RMSD and with dRMSD. Additionally, the scalar variant of $\mathcal{E}$RMSD, based on Eq. 6, is also presented.

We first observe that for short time separations the correspondence between geometric and kinetic distance is high (low CV) in all cases, and it diminishes at large time lag. The CV for the $\mathcal{E}$RMSD is consistently lower compared to both RMSD and dRMSD, suggesting this measure to better reflect the temporal separation between configurations. We also note that, in this test, the scalar and vectorial version of the $\mathcal{E}$RMSD behave very similarly.

Search of structural motifs. RNA molecules display a great variety of base–base interaction patterns (RNA motifs) that play a key role in RNA folding and that mediate important biological processes (44). A structural deviation should therefore accurately capture such RNA-specific features. This property can be assessed by searching for known RNA structural motifs in the database of experimentally solved structures. Here, we use the $\mathcal{E}$RMSD to search for five known internal loop motifs (K-Turn, C-loop, Sarcin Ricin, triple sheared, double sheared) and two hairpin loops (T-loop and GNRA) within the RNA structure atlas database (45) release 1.43, composed by 772 crystal structures. As a reference, we compared the results with 'find RNA 3D' (FR3D) (29), the most extensively used and carefully tested

Table 2. Structural search of known motifs within the RNA motif atlas. The table summarizes the number of found motifs compared with FR3D (http://rna.bgsu.edu/rna3dhub/motifs). See SD Text 11 for a full list.

Motif	Reference PDB	Found motifs	False positive
T-LOOP	3RG5	57/66	17
GNRA	4A1B	217/219	63
C-LOOP	3V2F	18 /22	7
D-SHEAR	1SAQ	23/26	3
T-SHEAR	3V2F	24/31	18
K-TURN	3GX5	10/22	15
SARCIN	1Q96	13/16	3

method for RNA motif recognition. FR3D employs a reduced representation with one centroid per nucleobase, and computes the geometrical deviation by combining fitting and orientation errors after optimal superposition. Reference and motif structures were obtained from the RNA 3D motif atlas release 1.13. Except for the K-turn motif, we were able to correctly identify almost all FR3D matches, including motifs with bulged bases (Table 2). By visual inspection, it can be seen that most of the false positive (i.e. motifs that were found using the $\mathcal{E}$RMSD and not by FR3D) are either *bona fide* matches or known motif variants, while the residual discrepancies are due to differences in the employed cutoff value (45). We refer the reader to SD Text 10 for a detailed description of the search strategy and to SD Text 11 for the full list of results.

DISCUSSION

A fully atomistic description of an RNA molecule requires on average $\approx$20 non-hydrogen atoms per nucleotide, and thus $\approx$60 degrees of freedom per nucleotide. Biologically relevant RNA structures are composed of hundreds or even thousands of nucleotides that display an intricate network of interactions. The complexity of the problem is further increased when considering that many RNA molecules are highly dynamic entities, and multiple chain configurations are needed for describing their mechanism and function (3). In order to provide a simpler representation, RNA structures are usually analysed in terms of sugar pucker conformations, backbone dihedrals (20,21) or pseudo dihedrals (22). Complementary approaches consider instead base-pair parameters (propeller, slide, roll, twist, etc.) and hydrogen bonding to classify base-pair interactions in terms of recurrent geometrical configurations observed in experimentally solved RNA structures (23–25).

A key result of the present work is that the main structural and dynamical properties of RNA can be described in terms of the spatial arrangement of the nucleobases only. Since nucleobases are substantially rigid, this implies 6 degrees of freedom per nucleotide. This observation is in full accordance with previous analysis of RNA crystal structures (25,46), where the position and orientation of nucleobases only were considered. This work complements the analysis by showing that nucleobase position and orientation are sufficient to correctly describe the RNA dynamics as obtained from explicit solvent molecular dynamics simulations.

As shown in Figure 1, RNA base–base interactions mainly occur in a restricted neighborhood of the nucleobase that is well approximated by an ellipsoidal shell. Since

specific regions of the ellipsoidal shell surrounding nucleobases correspond to different types of pairing and stacking interactions, it is in principle possible to infer secondary structure information directly from the pairing and stacking plots of Figure 2. This procedure is formally similar to the one employed in the analysis of protein structures, where the angular preferences of the backbone approximately correspond to different secondary structures (19). Our approach depends on nucleobase positions only and is thus complementary to that proposed in Ref. (22), where backbone pseudo dihedrals were used to classify RNA secondary structures.

By considering only the relative spatial arrangement of nucleobases, we introduce a knowledge-based scoring function for RNA structure prediction ($\mathcal{E}$SCORE). In contrast with several studies underlining the importance of an atomistic description for an accurate scoring (15–17), our work shows that a minimalist description of nucleobase arrangement is sufficient to produce equivalent or better results. Two aspects are worth highlighting. First, while the backbone can be ignored for scoring, one cannot avoid including the chain connectivity in the sampling algorithm used to generate the decoys. Second, the outcome of the scoring benchmarks is strongly dependent on the employed decoy sets. Some of the decoy sets generated using the FARNA algorithm are very challenging, as all scoring functions ($\mathcal{E}$SCORE, FARFAR and RASP) are highly degenerate, and assign good scores to structures that are very far from native (see e.g. Supplementry Figure S6, decoy set 1A4D and Ref. (16)). These decoys can be therefore used to improve the current methodologies for *de novo* predictions of RNA structure.

A second aspect intimately connected with the choice of molecular representation is the definition of a proper structural proximity measure. Typically, structural deviations are used either to analyse and compare three-dimensional structures (e.g. RMSD, dRMSD, INF (27)) or to recognize and cluster RNA structural motifs (29–31). The $\mathcal{E}$RMSD introduced here is suitable for both tasks. A distinctive feature of the $\mathcal{E}$RMSD is the use of the ellipsoidal metric $\tilde{r}$, that naturally stems from the empirical data and that takes into account both relative distance and orientation of nucleobases. Interestingly, the intrinsic base anisotropy has been also discussed in the context of the electronic polarizability of nucleobases (35) and anisotropic interactions have been used in coarse-grain models for proteins (47). We note that the $\mathcal{E}$RMSD is conceptually reminiscent of the INF measure (27), that also considers the differences in the network of base-pairing and base-stacking interactions. With

respect to INF, however, the $\mathcal{E}$RMSD presents important advantages: i) it does not depend on the annotation scheme employed and therefore on the quality of the structure, ii) it is a continuous function of the atomic coordinates and iii) it is well defined even when the annotation is not informative (e.g. for unfolded states, unstructured, or single-stranded RNAs). The last two items make the $\mathcal{E}$RMSD particularly well-suited for studying dynamical processes, such as RNA folding or binding events. The $\mathcal{E}$RMSD also shares a number of similarities with the structural distance used in FR3D for recognizing three-dimensional motifs (29). However, in the $\mathcal{E}$RMSD calculation we introduce important simplifications: i) the definition of the local coordinate system is not nucleobase-specific and requires the knowledge of three atoms per nucleobase only (Figure 1a), ii) the $\mathcal{E}$RMSD calculation does not require least square fitting and iii) the problem of combining the positional and orientational base–base distance is circumvented by introducing an ellipsoidal metric. The fact that the two methods produce similar results (Table 2) strongly suggests that for most practical uses the $\mathcal{E}$RMSD is accurate enough to capture the specific pattern of base–base interactions.

In this study we introduce a minimalist representation of RNA three-dimensional structures that solely considers the relative orientation and position of nucleobases. Based on this representation, we construct a scoring function for RNA structure prediction and we define a metric for calculating deviation between RNA three-dimensional structures. In both cases, the results are on par or better compared with state-of-the-art techniques. Taken together, the presented results suggest that this minimalist representation captures the main interactions that are relevant for describing RNA structure and dynamics. As a final remark, we note that the methodology presented here can be readily applied to DNA molecules as well.

AVAILABILITY

Python scripts to compute $\mathcal{E}$SCORE and $\mathcal{E}$RMSD from PDB structures are available at https://github.com/srnas/barnaba.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. Kruger,K., Grabowski,P.J., Zaug,A.J., Sands,J., Gottschling,D.E. and Cech,T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, **31**, 147–157.
2. Klein,D., Moore,P. and Steitz,T. (2004) The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.*, **340**, 141–177.
3. Serganov,A. and Nudler,E. (2013) A decade of riboswitches. *Cell*, **152**, 17–24.
4. Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
5. Tinoco,I. (1996) Nucleic acid structures, energetics, and dynamics. *J. Phys. Chem.*, **100**, 13311–13322.
6. Cruz,J.A., Blanchet,M.-F., Boniecki,M., Bujnicki,J.M., Chen,S.-J., Cao,S., Das,R., Ding,F., Dokholyan,N.V., Flores,S.C. *et al.* (2012) RNA-puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, **18**, 610–625.
7. Kuhrova,P., Banas,P., Best,R.B., Sponer,J. and Otyepka,M. (2013) Computer folding of RNA tetraloops? Are we there yet? *J. Chem. Theory Comput.*, **9**, 2115–2125.
8. Chen,A.A. and Garcia,A.E. (2013) High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 16820–16825.
9. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
10. Denesyuk,N.A. and Thirumalai,D. (2013) Coarse-grained model for predicting RNA folding thermodynamics. *J. Phys. Chem. B*, **117**, 4901–4911.
11. Parisien,M. and Major,F. (2008) The MC-fold and MC-sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
12. Cao,S. and Chen,S.-J. (2011) Physics-based de novo prediction of RNA 3D structures. *J. Phys. Chem. B*, **115**, 4216–4226.
13. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 14664–14669.
14. Frellsen,J., Moltke,I., Thiim,M., Mardia,K.V., Ferkinghoff-Borg,J. and Hamelryck,T. (2009) A probabilistic model of RNA conformational space. *PLoS Comput. Biol.*, **5**, e1000406.
15. Das,R., Karanicolas,J. and Baker,D. (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat. Methods*, **7**, 291–294.
16. Bernauer,J., Huang,X., Sim,A.Y. and Levitt,M. (2011) Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA*, **17**, 1066–1075.
17. Capriotti,E., Norambuena,T., Marti-Renom,M.A. and Melo,F. (2011) All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics*, **27**, 1086–1093.
18. Tanaka,S. and Scheraga,H.A. (1976) Medium-and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945–950.
19. Ramachandran,G., Ramakrishnan,C.T. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
20. Murray,L.J., Arendall,W.B., Richardson,D.C. and Richardson,J.S. (2003) RNA backbone is rotameric. *Proc. Natl Acad. Sci. U.S.A.*, **100**, 13904–13909.
21. Hershkovitz,E., Sapiro,G., Tannenbaum,A. and Williams,L.D. (2006) Statistical analysis of RNA backbone. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **3**, 33–46.
22. Duarte,C.M. and Pyle,A.M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.
23. Saenger,W. (1984) *Principles of nucleic acid structure*. Springer-Verlag, NY, **7**.
24. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
25. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.

26. Kabsch,W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, **32**, 922–923.
27. Parisien,M., Cruz,J.A., Westhof,E. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
28. Yang,H., Jossinet,F., Leontis,N., Chen,L., Westbrook,J., Berman,H. and Westhof,E. (2003) Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.*, **31**, 3450–3460.
29. Sarver,M., Zirbel,C.L., Stombaugh,J., Mokdad,A. and Leontis,N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215–252.
30. Apostolico,A., Ciriello,G., Guerra,C., Heitsch,C.E., Hsiao,C. and Williams,L.D. (2009) Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res.*, **37**, e29.
31. Zhong,C., Tang,H. and Zhang,S. (2010) Rnamotifscan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic Acids Res.*, **38**, e176.
32. Beauchamp,K.A., McGibbon,R., Lin,Y.-S. and Pande,V.S. (2012) Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 17807–17813.
33. Sponer,J., Leszczynski,J. and Hobza,P. (2001) Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers*, **61**, 3–31.
34. Bugg,C., Thomas,J., Sundaralingam,M.t. and Rao,S. (1971) Stereochemistry of nucleic acids and their constituents. X. Solid-slate base-slacking patterns in nucleic acid constituents and polynucleotides. *Biopolymers*, **10**, 175–219.
35. Sponer,J., Riley,K.E. and Hobza,P. (2013) Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.*, **10**, 2595–2610.
36. Juang,B.-H., Hou,W. and Lee,C.-H. (1997) Minimum classification error rate methods for speech recognition. *IEEE Trans. Audio Speech Lang. Processing*, **5**, 257–265.
37. Bonomi,M., Branduardi,D., Gervasio,F.L. and Parrinello,M. (2008) The unfolded ensemble and folding mechanism of the C-terminal GB1 β-hairpin. *J. Am. Chem. Soc.*, **130**, 13938–13944.
38. Cossio,P., Laio,A. and Pietrucci,F. (2011) Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Phys. Chem. Chem. Phys.*, **13**, 10421–10425.
39. Simons,K.T., Kooperberg,C., Huang,E. and Baker,D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.*, **268**, 209–225.
40. Cossio,P., Granata,D., Laio,A., Seno,F. and Trovato,A. (2012) A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci. Rep.*, **2**, 351.
41. Zhou,T. and Caflisch,A. (2012) Distribution of reciprocal of interatomic distances: a fast structural metric. *J. Chem. Theory Comput.*, **8**, 2930–2937.
42. Serganov,A., Yuan,Y.-R., Pikovskaya,O., Polonskaia,A., Malinina,L., Phan,A.T., Hobartner,C., Micura,R., Breaker,R.R. and Patel,D.J. (2004) Structural basis for discriminative regulation of gene expression by adenine-and guanine-sensing mRNAs. *Chem. Biol.*, **11**, 1729–1741.
43. Di Palma,F., Colizzi,F. and Bussi,G. (2013) Ligand-induced stabilization of the aptamer terminal helix in the add adenine riboswitch. *RNA*, **19**, 1517–1524.
44. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
45. Petrov,A.I., Zirbel,C.L. and Leontis,N.B. (2013) Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA*, **19**, 1327–1340.
46. Lemieux,S. and Major,F. (2006) Automated extraction and classification of rna tertiary structure cyclic motifs. *Nucleic Acids Res.*, **34**, 2340–2346.
47. Fogolari,F., Esposito,G., Viglino,P. and Cattarinussi,S. (1996) Modeling of polypeptide chains as Cα chains, Cα chains with Cβ, and Cα chains with ellipsoidal lateral chains. *Biophys. J.*, **70**, 1183–1197.
48. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974.