What Can Human-Guided Simulations Bring to RNA Folding?

Liuba Mazzanti, Sébastien Doutreligne, Cédric Gageat, Philippe Derreumaux, Antoine Taly, Marc Baaden, Samuela Pasquali

What can human-guided simulations bring to RNA folding?

L. Mazzanti[a], S. Doutreligne[a*], C. Gageat[b], P. Derreumaux[a] ,
A. Taly[a], M. Baaden[a], S. Pasquali[a,c]

[a] Laboratoire de Biochimie Théorique, IBPC, CNRS UPR 9080,
Université Paris Diderot, Sorbonne Paris Cité,
13 rue Pierre et Marie Curie, Paris
[b] Ecole normale supérieure, Département de Chimie,
PSL Research University, Université Pierre et Marie Curie, Sorbonne Paris Cité
24, rue Lhomond, 75005 Paris
[c] Laboratoire de Cristallographie et RMN Biologiques, CNRS UMR 8015,
Université Paris Descartes, Sorbonne Paris Cité
4 avenue de l'Observatoire, Paris

Abstract

Inspired by the recent success of scientific-discovery games for predicting protein tertiary and RNA secondary structures, we have developed an open software for coarse-grained RNA folding simulations guided by human intuition. To determine to what extent interactive simulations can accurately predict three-dimensional RNA structures of increasing complexity and lengths (4 RNAs with 22-47 nucleotides), a participative experiment was conducted on 141 participants who had very little knowledge of nucleic acids systems and computer simulations and just received a brief description of the important forces stabilizing RNA structures. Their structures and full trajectories have been analyzed statistically and compared to standard replica exchange molecular dynamics simulations. Our analyses show that participants gain easily chemical intelligence to fold simple and non-trivial topologies with little computer time, and this result opens the door for the use of human-guided simulations to RNA folding. Our experiment shows that interactive simulations have better chances of success when the user widely explores the conformational space. Interestingly, the on-the-fly feedback of the Root Mean Square Deviation with respect to the experimental structure, did not improve the quality of the proposed models.

1 Introduction

It is recognized that the function of many RNA molecules depends crucially on their three-dimensional structures. These structures exhibit a wide diversity of architectures, often including non-canonical pairs as well as triplets and quartets with 145 different base-pairs, according to Leontis classification, found experimentally (RNA Basepair Catalog of the Nucleic Acids Database) [1, 2]. Compared to proteins, the number of experimentally resolved RNA structures, is still very limited. In silico predictions can therefore help fill the gap between sequences and structures. In recent years three series of RNA structure predictions competitions (RNA-puzzles [3, 4, 5]) have highlighted how computer predictions are best when homology reconstruction is a viable route, when experimental information is available on local base pairing from chemical

*L.M. and S.D. equally contributed to this work

probing, and when the structure itself is mainly driven by Watson-Crick base-pairings. Predictions of structures stabilized by non Watson-Crick base pairs are still challenging, even when the sequence information is complemented by chemical probing data [4].

The best prediction methods currently available are those based on fragment reconstructions [6] and those including predictions of secondary structures first, followed by 3D motif assembly [7]. Methods based on secondary structure predictions start by considering canonical base pairs, since they are the most abundant and stacks of canonical base pairs make up A-RNA 3D stems. Canonical base pairs are also the most well characterized for $\Delta\Delta$G's, and can therefore be used for accurate thermodynamic predictions of duplex formations [8]. However, in a significant percentage of experimental RNA structures, non-canonical base pairs, triplets, quartlets, as well as pseudoknots, increase substantially the complexity of RNA 3D structures (in 28S rRNA 15% of in-stem pairs are non-canonical and $\sim$20% are long-range pairs or triplets). As a result, the combinatorial complexity of RNA increases sharply with sequence length: $O(N^3)$ for secondary structures without pseudoknots [9, 10], and between $O(N^4)$ and $O(N^6)$ for secondary structures with pseudoknots [11, 12]. All taken into account, RNA secondary structure prediction including pseudoknots has been shown to be NP-complete [13].

A complementary strategy to bioinformatics approaches is that of building physical models by simulating the molecule's folding according to a force field. Physical models have the advantage that the base pairing space is naturally restricted by physically accessible conformations, allowing to consider an arbitrarily large set of possible base pairs and to generate all topologies with the same computational complexity. The limitation of physical models resides in the sampling of the conformational space even with the most advanced enhanced simulation techniques. In order to investigate large structural rearangements, like the ones involved in folding, a simplification of the system through coarse-graining is needed [14, 15, 16, 17, 18]. Despite the fact that coarse-grained force-fields are still in their infancy, simulations can complement bioinformatic predictions by giving access to the dynamical and thermodynamical behavior of the molecule, and also by identifying possible alternative conformations, metastable states and kinetic traps [19, 20, 21].

Though more work is certainly necessary to achieve reliable RNA force fields, we present here an application of coarse-grained modeling coupled to interactive molecular dynamics (MD) simulations as a proof of principle of what can be accomplished when a user is given the opportunity to steer the system based on a reasonable force field. For most biomolecular systems, for which it is difficult to identify a limited set of descriptors able to capture the specificity of a given state, justifying why dihedral angle principal component analysis is often used to describe the energy landscape [22], interactive simulations offer the possibility of exploiting the human ability to recognize patterns.

Inspired by the excellent results of Foldit [23] for predicting protein 3D structures and EteRNA [24] for predicting RNA secondary structures, which pioneered the coupling between the powers of computer predictions and the intuition of the human intellect, we have developed an open software combining interactive non-equilibrium molecular dynamics simulations with the HiRE-RNA force field for folding, unfolding or deforming structural models. Interactive simulations are performed with the in-house software UnityMol [25], which allows for the visualization of a MD trajectory in real time, allows the user to change the temperature, and to apply forces to selected particles through a variety of hardware devices, including the ubiquitous computer mouse.

As a first test of the effectiveness of our approach, we set up a participative experiment where 141 participants were asked to make RNA folding predictions using interactive simulations for 4 molecules of increasing length (22-47 nucleotides) and 3D complexity. The experiment was carried out in two successive rounds, with slight variations as detailed below. In this manuscript we present the basic ideas of the HiRE-RNA model and of interactive simulations, the setup of the experiment, and the prediction results comparing also with the performance of fully automatic computer simulations. The software and benchmark molecules used in the experiment are freely

available on the HiRE-RNA contest page (https://hirerna.galaxy.ibpc.fr/).

2 Materials and Methods

We carry out interactive simulations by coupling UnityMol, a molecular visualization software for chemistry and biology, and the simulator MD engine implementing the HiRE-RNA force field [26, 27, 28, 29].

2.1 The HiRE-RNA coarse-grained RNA model

This description of the HiRE-RNA model corresponds to the explanations that all participants received prior to carrying out the experiment. The full presentation of the model can be found in [17, 30].
HiRE-RNA is an implicit solvent, implicit ion model where each nucleotide is represented by 6 or 7 beads (see fig. 1 of SI) corresponding to the backbone heavy atoms P, O5', C5' and C4, C1' of the sugar, and to the center of mass of each of the aromatic rings of the bases (G1, G2, A1, A2, C1, U1). The force field of the model is composed of local interactions accounting for the local stereochemistry and an excluded volume interaction giving a physical size to the beads, and non-local interactions accounting for base pairing, base stacking and electrostatics. Local interactions are composed of an harmonic potential for bond lengths and for angles amplitudes, and a sinusoidal potential for dihedral angles. A fast-decreasing exponential function describes the excluded volume potential. Phosphate beads carry one negative charge each and have a repulsive interaction with each other.
Both base pairing and stacking crucially depend on the relative positions and orientation of the bases. In order to recover the anisotropy of a base, from the model's isotropic particles, base planes are identified by the particles C1'-B1-B2 for purines and C4-C1'-B1 for pyrimidines. Both stacking and base-pairing can occur between any two bases of the system. The stacking potential is minimized when the distance between bases is close to an equilibrium distance, and when the planes are parallel and vertically aligned (see fig. 2 of SI). Base pairing occurs when two bases are side-by-side on the same plane and depends on the relative distance and orientation. To account for the multiple pairing possibilities of each base, equilibrium values depend on the bases' species and on their orientation. In the current model we account for 22 different possible pairs, including the two canonical pairs A-U and G-C, 8 pairs occurring between Watson-Crick sides of any two bases (all possibles with the exception of G-G), and 12 other pairs representing interactions involving also the Hoogsteen and Sugar edges of the base. The energy of each base pair is proportional to the number of hydrogen bonds forming the pair, that is 3 for G-C, and two or one for the other pairs according to the table in [18].
The HiRE-RNA force field, as any coarse-grained force field for RNAs, is still evolving and suffers from the limitations of not having an explicit description for ions, of a parametrization yet to be extended to include thermodynamical and dynamical quantities and it has to be put to the test on larger and more complex systems than benchmark molecules. However, for the the experiment we present here, the goal was to have a plausible physical coarse-grained model, to which HiRE-RNA seemed adequate. Given the modular set-up of interactive simulations, the molecule's representation and force-field can be easily changed.

2.2 Visualization and user interaction through the UnityMol application

UnityMol is a molecular visualization software based on the Unity3D game engine [25]. It features molecule representations commonly found in the discipline and serves as an experimental platform to propose specialized methods (i.e. custom polysaccharides rendering [31]).

As coarse-grained models are not easily rendered on standard software, UnityMol was modified to generate appropriate and visually appealing representations. For HiRE-RNA, bases can be rendered through ellipsoids whose orientations correspond to those of bases' planes, as explained in the previous section. This makes it easier to visually detect stacking and possible base pairing. When connected to an interactive HiRE-RNA simulation, plots of selected energy terms over time yield a quantitative insight into the molecule's stability (Figure 5 SI). Using the computer mouse direct action on the simulation is possible. Force vectors are computed based on the selected atom and the current cursor displacement. These forces are transmitted to the simulation engine and added to the force field. This scheme offers a direct, almost instantaneous, visual feedback. More details about the functionalities of UnityMol and the web application are given in Supplementary Material.

2.3 Setting up an RNA folding challenge as participative experiment

Participants for this study involve two classes of third year college students majoring in biology. During the course of 2015 and 2016 interactive nucleic acids simulations have been integrated as a mandatory lab for the bioinformatics course, at Paris Diderot University. The course was the very first introduction to numerical tools for the study of biomolecules. During the semester, students received a two-hours lecture on the analysis of bimolecular structures including a light overview of structure prediction methods as well as a one-hour lecture on modeling of biomolecules and basic principles of molecular dynamics. All participants were therefore novice users of molecular simulation techniques and RNA structures. Since users were familiar only with the DNA double helix, but ignored the folding capabilities of single stranded nucleic acids, an overview of nucleic acids structures was given as introduction to the lab.

Users learned how to use UnityMol and interactive RNA simulations through two exercises of pulling open a double helix and reforming it, making observations on the different energy terms, with the local harmonic potential governing the response as the molecule is being pulled by an external force, while base pairing and stacking driving and stabilizing folding.
Users were then given 3 hours to work on the HiRE-RNA folding challenge, having to fold four molecules of increasing complexity. The starting point of each exercise was a completely stretched conformation. Users could launch an interactive molecular dynamics simulation with Langevin dynamics for friction. The launching applet allowed to choose the temperature, which could then be changed by pausing the simulation and relaunching it with a different T value.
Users were given the instruction to select up to five conformations that could correspond to the native structure for the molecule. Their selection was submitted to a server and entered in the competition. In 2015, evaluation of the RMSD structures with respect to the experimental ones was given to the users at the end of the competition, while in 2016 the server was indicating the score of the structure and RMSD, immediately upon submission, giving users a real-time assessment of the validity of their structures while performing the experiment. We will refer to the 2015 as the "non-feedback experiment" and to the 2016 as the "feedback experiment". Since the strategies adopted for the non-feedback experiment and the feedback experiment were different, we have analyzed both rounds separately.
For our subsequent analysis, all submitted structures were recovered from the server as well as the full trajectories, that were physically recovered from each machine.

2.3.1 Four RNA molecules of increasing complexity

The four molecules consist of a simple hairpin, a hairpin with an asymmetric bulge, an H-pseudoknot, and a triple helix pseudoknot (fig. 1).
1F9L is a hairpin of 22 nucleotides including 6 canonical base-pairs, one Hoogsteen G-U pair, and

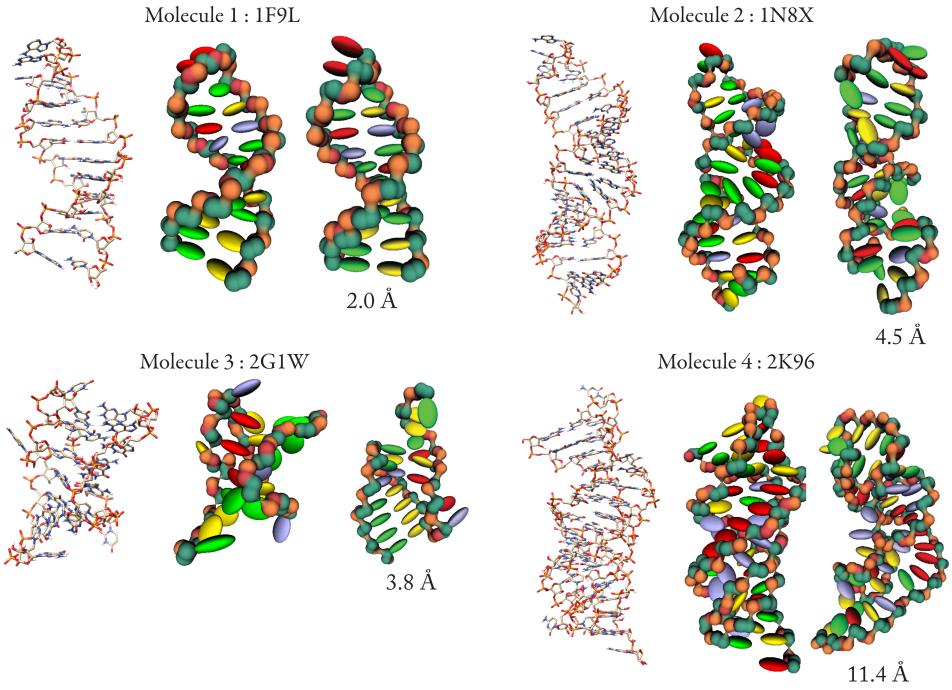

Figure 1: The four molecules proposed to the participants for the folding challenge represented atomistically (left) and in the coarse-grained representation used in the actual experiment with UnityMol (center). The image on the right represents the lowest RMSD prediction made by the participants in the first round. 1F9L and 2G1W correspond well to the native structure with a low RMSD and the correct base pairing organization. 1N8X exhibits some non native base pairs. 2K98 has the correct overall shape, forming a triple helix pseudoknot, but deviates significantly in the local organization. The complete list of base pairs for the 4 molecules (native and predictions) is given in SI.

two A-G pairs [32]. According to Leontis classifications all pairs are cis Watson-Crick/Watson-Crick (see SI for detailed list). 1N8X is a 36 nucleotides hairpin including one asymmetric bulge [33]. The native conformation is stabilized by 14 cis Watson-Crick/Watson-Crick base pairs, including one A-G pair adjacent to the bulge and one G-U pair in proximity of the hairpin loop. 2G1W is a 22 nucleotides simple pseudoknot composed of 7 canonical G-C base pairs [34]. 2K96 is a 47 nucleotides triple helix pseudoknot [35]. It consists of a Watson-Crick double helix and of an A-rich dangling strand inserting into the WC helix groove and forming several stacked triplets. Overall, 21 base pairs, canonical and non-canonical, stabilize the native structure, including six triplets (five A-U-A and one C-G-A).
These four sequences, starting from fully elongated states, were previously folded [17, 18] by long non-biased simulated tempering (ST) and by replica exchange molecular dynamics (REMD) simulations coupled to HiRE-RNA. The simulation on 1F9L found the global free energy minimum at 3.2 Å from the experimental state, for 1N8X the deviation is 3.8 Å for 2G1W the deviation is 4.3 Å and for 2K96, the deviation is 4.3 Å.

2.3.2 Analysis of the participants' performance

We carried out two separate analyses to study the usefulness of interactive simulations in addressing the question of RNA folding. The first focuses only on the structures submitted by the users to the online server and wants to answer the question if a "naive" user is able to produce correctly

folded structures and recognize them as such. The second analysis focuses on the full trajectories generated by each participant and wants to investigate how the molecule's conformational space is explored.

The two quantities used to compare submitted structures and trajectories to the native conformation are the RMSD, computed on all beads of the coarse-grained representation, and base-pairing. For RMSD we have used a cutoff of 6Å to detect structures corresponding to the native state. This value comes from our experience of previous simulations with HiRE-RNA at physiological temperatures where the RMSD can fluctuate of about 6Å while preserving all correct base-pairs and fold. This criterion gives only a rough estimate of the correspondence between two structures, as even lower RMSD values do not necessarily imply correct stack or pairs. Base-pairs were considered formed if they have at least 10% of the maximal interaction energy between the two bases.

For all trajectories we analyzed structures from frames taken every $4ps$. We monitored the total internal energy given by the HiRE-RNA potential, which we then normalize with respect of the energy of the native structure, the overall number of base pairs as well as native base pairs. In order to better detect base pairs, we smoothed fluctuation using a moving window over several subsequent frames as described in [19]. To give a more accurate, yet concise, description of the molecule's architecture we also looked at its topology starting from the list of detected base pairs, as defined in [36, 37]. Details of the analysis procedure are given in Supplementary Material.

3 Results and Discussion

3.1 Participants predict a significant proportion of native folds

Overall participants submitted between 80 and 200 structures, depending on the molecule and on the year. Not all participants used all the 5 attempts at their disposal. A summary of the results of submitted structures is reported in Table 1 in which we present also, as reference, the predictions by two commonly used programs McSym [7] and Vfold [38]. For each molecule, in SI we report the base pairing of the lowest RMSD structure proposed by the students next to the details of the base paring of the experimental structure.

The ratio of predicted structures of RMDS lower than 6Å ρ, varies significantly with the molecule, with the best predictions, as expected, for the simple hairpin (1F9L), where almost half of the submissions correspond to the native state in the non-feedback experiment and one quarter in the feedback experiment, and for which the lowest RMSD structures in both experiment have base pairing identical to native.

Molecule 2 (1N8X) was harder to predict than molecule 1 because of the asymmetric bulge in its middle region. Most submitted structures include 7 correctly paired bases (lower stem). Some structures predicted the correct base pairs but resulted in distorted overall shapes, bringing the RMSD to about 10Å, and are therefore not included in ρ. Other high RMSD structures exhibit also a high number of non-native base pairs, and were folded into low-energy structures alternatives to the experimental configuration. The lowest RMSD structures have 9/14 native base pairs for the non-feedback experiment, and 10/14 for the feedback experiment.

Molecule 3, 2G1W, has a markedly doubly-peaked distribution. The lowest peak corresponds to the formation of the two stems in the pseudoknot configuration, while the higher peak corresponds to only one of the stems being formed. This is in agreement with results from REMD simulations. Most structures predicted the formation of one of the stems and include also some non-native pairs, achieving alternative compact structures. They are mainly mismatched (non-canonical) hairpins. The lowest RMSD structures have 7/7 native base pairs for the non-feedback experiment and 4/7 for the feedback experiment.

Given the size and complexity of molecule 4, we did not expect users to be able to fully predict

Table 1: Submitted structures' statistics. For each round we report the total number of structures submitted by the participants (str.), the percentage ρ of structures with RMDS below 6Å, the lowest RMSD among all submissions, and the approximate values of the first two peaks of the distribution in RMSD of all structures (the full distribution is visible in gray in fig. 2, horizontal histograms). *For 2K96 we give a looser definition of the percentage of success and we consider the number of structures exhibiting the native topology. As a reference we report also RMSD values of structures folded with the two bioinformatics programs McSym and Vfold accessible on-line. Values are averages computed over the 10 best structures according to the programs. For the two pseudoknots, McFold/MCSym was not able to predict the correct topology despite having allowed the search for H-shaped pseudoknots. Instead it proposed hairpins. VFold found the correct secondary structure but gave errors when attempting to build a 3D structures based on the pseudoknotted secondary structure, not finding suitable motifs.

	non-feedback experiment				feedback experiment				McSym	Vfold
Molecule	str.	ρ	lowest RMSD (Å)	RMSD peaks (Å)	str.	ρ	lowest RMSD (Å)	RMSD peaks (Å)	RMSD (Å)	RMSD (Å)
1F9L	88	50%	2.0	4; 7	203	25%	2.1	6; 12	4.0	3.75
1N8X	90	10%	4.5	5; 10	207	5%	2.6	8; 12	5.5	3.7
2G1W	96	13%	3.8	6; 12	168	3%	5.7	8; 12	12.9	na
2K96	74	13%*	11.4	11; 15	119	8%*	11.2	11; 15	33	na

Table 2: For each molecule we analyze the 15 lowest energy structures submitted by the users and we report the number of structures with low RMSD values and the number of structures with a high percentage of native base pairs. The choice of 15 lowest energy structures is arbitrary but is in the range of what is typically analyzed by prediction methods.

	non-feedback experiment		feedback experiment	
Molecule	best rmsd	native BP	best rmsd	native BP
1F9L	$10/15 \leq 5$Å	$6/15 \geq 0.75$	$12/15 \leq 5$Å	$7/15 \geq 0.75$
1N8X	$7/15 \leq 6$Å	$5/15 \geq 0.75$	$4/15 \leq 6$Å	$3/15 \geq 0.75$
2G1W	$3/15 \leq 6$Å	$7/15 \geq 0.75$	$2/15 \leq 8$Å	$8/15 \geq 0.75$
2K96	$6/15 \leq 11$Å	$5/15 \geq 0.40$	$5/15 \leq 12$Å	$2/15 \geq 0.40$

its structure. We were however interested in testing how far they could come in proposing a plausible structure with the correct topology. Both years 10 structures were submitted with RMSD between 11 and 12 Å and corresponded to the topology of the pseudoknot. Other structures included the WC helix but did not reach the folding into a pseudoknot, leaving a dangling end. The distribution in RMSD exhibits a small peak at 12Å and is for the rest rather flat, showing how there wasn't an alternative structure found by the users, but all other proposed structures sampled widely the more or less unfolded states. The lowest RMSD structures are also the ones with the most native base pairs, with 9/21 native base pairs for the non-feedback experiment and 12/21 for the feedback experiment.
When we analyze submitted structures based on their internal energy, we systematically find some of the lowest RMSD and highest native base pairs structures among the 15 lowest energy submissions (Table 2). This is an encouraging results as in a blind prediction one would usually focus on the lowest energy structures.

The combined results for both experiments show that simple molecules could be folded quickly and easily by a large portion of users, while molecules with more articulate structures are clearly harder to predict. Still, a significant portion of users were able to generate the native conformation and recognize it as such in about 30 minutes of interactive simulation. In addition, it was possible to generate alternative conformations and test them for stability. Even for very complex

architectures such as the triple helix, some users were led to the intuition of the correct topology of the molecule. This is particularly remarkable as none of them had any prior experience with RNA structures other than double helices and hairpins.

The users' strategy in the feedback experiment was different than in the non-feedback experiment. Indeed the number of submissions in the feedback experiment is roughly twice as much as in the non-feedback experiment. The number of users being comparable in the two years, one can immediately observe that users in the feedback experiment submitted more structures than their non-feedback experiment colleagues. In the feedback experiment a significant percentage submitted structures have a high RMSD suggesting that users submitted one or two randomly chosen structures just to test how far they were from the correct solution and used this information for the pursuit of the challenge. It is interesting to notice how this real-time feedback does not seem to give any particular advantage in the prediction of the folded structure as it can be observed by the comparison of all statistical quantities in Table 1 between the non-feedback experiment and the feedback experiment. On the contrary, one can argue that results in the non-feedback experiment are better than those of the feedback experiment. This observation is reminiscent of the observation made with Foldit that players could move from one basin to another through their ability to ignore a quantitative score [23].

3.2 Humans explore phase space more broadly than automated approaches

Having retrieved single trajectories from each user's machine, we have analyzed the full exploration of the conformational space of each simulation with the goal of understanding the contribution of interactive simulations over regular, enhanced sampling, simulations. Results were assessed after merging all trajectories together, keeping the distinction between the non-feedback experiment and the feedback experiment and comparing them to the results from REMD simulation performed on a computer cluster using 32 replicas, spanning from 250 to 500K. For REMD simulation we have analyzed the structures of one low temperature replica, corresponding to a temperature below melting where the native state is present, if not dominant. An example of a participant's single trajectory is presented in Supplementary Material (Figures 5 and 6).

Sampling is focused on low energy conformational space

Figure 2 illustrates the distributions of internal energy vs. RMSD. Interactive simulations focus sampling on low energy conformational space. Most of the structures in the full trajectories are well above the native internal energy, however, structures picked by participants have lower energies as shown by the height of the gray peaks in population density (both for RMSD and energy) compared to the blue peaks extracted from full trajectories. For low RMSD structures these energies are close to native. Selected structures' energies are generally low because users spontaneously proceeded in a sort of simulated tempering by restarting the simulation at different temperatures. When they thought a structure was close to native, they stopped the simulation and relaunched it with a lower temperature to reduce fluctuations and perform small adjustments to the structure, with temperature lowered to as much as 10K. They then raised it back to room temperature on the refined structure to test for its stability.

In the non-feedback experiment, users sampled extensively different basins, including the exact native state for all molecules except the triple helix (2K96). In the feedback experiment, users sampled more uniformly the conformational space in RMSD and internal energy. This can be observed by the presence of several, well separated, population peaks in the plots of the non-feedback experiment, while a more uniform diagonal shade is observed for the feedback experiment. It appears that the instantaneous assessment provided in the feedback experiment led to a gradual decrease in RMSD, but prevented from exploring disconnected basins. This can explain why in the feedback experiment users were less successful at folding than in the

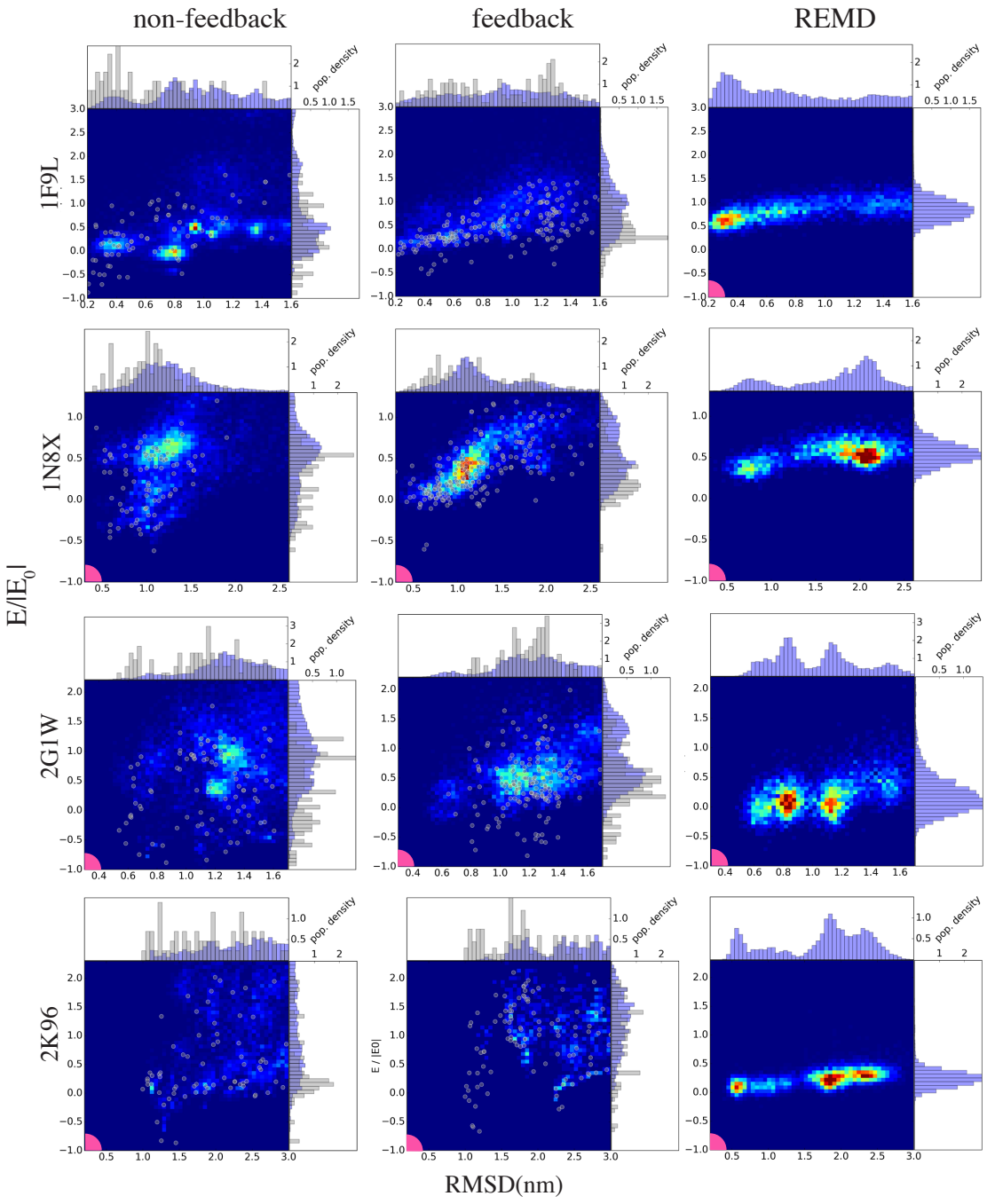

Figure 2: Internal energy vs. RMSD distributions for all interactive simulations as well as for one REMD simulation. The population color coding for full trajectories goes from blue (low) to red (high), while values from individual submitted structures are superposed as gray circles. Internal energy at finite temperature is normalized with respect to the absolute value of the energy of the minimized native structure $|E_0|$. RMSD distributions for both full trajectories (blue bars) and submitted structures (gray bars) are presented on the horizontal histogram, while energy distributions are presented on vertical histograms. The pink wedge in each PMF indicates the position of the native structure (RMSD = 0, $E/|E_0| = -1$).

non-feedback experiment.
The details of the results vary from molecule to molecule. For 1F9L, in the non-feedback experiment users sampled extensively at least 3 different basins, as it appears from the three distinct peaks in population density, while they sampled more connected basins in feedback experiment, remaining further away from the native state. For 1N8X, the full trajectories of the non-feedback experiment remain globally at a higher internal energy than those of the feedback experiment, however in the non-feedback experiment users were able to reach lower energy states with a better correspondence to the native structure and select them as candidates for native. The same is true for 2G1W. Interestingly for this molecule, in the feedback experiment users did sample a basin at 6Å RMSD corresponding to the native state, but they did not select these structures as possible native candidates. In the non-feedback experiment this region was less explored, but recognized as native by a dozen users. As a general trend, in the non-feedback experiment users explored more widely in energy. They seem to have sampled lower energy states than in the feedback experiment and have picked these states for their submission.

For comparison, trajectories from REMD simulations spend most of their time exploring the unfolded states and, despite the presence of low temperatures, do not minimize the energy as effectively as interactive simulations. Still, a peak corresponding to the native structure is clearly visible, even though it represents only a small fraction of the overall population and its internal energy is similar to other states.

Base pairing and topology measure native fold propensity

To assess whether an RNA structure is correctly folded it is important to consider also the base pairing network, and not simply at the RMSD. For the non-feedback experiment, which based on the previous analysis and discussion we consider the most interesting, we have analyzed the details of base pairing. Results are reported in fig. 3. For each molecule we looked at the overall number of base pairs, at the number of native base pairs and at two topological parameters allowing to compare the general features of the base pairing network to that of the native structure.

For all molecules we can observe that trajectories focused on configurations with a relatively high number of base pairs. This is particularly clear for 1F9L and 1N8X where we can observe a peak of the distribution of base pairs at values close to the native number of pairs (first column, vertical histogram in blue). The number of native base pairs however is low. Only a negligible percentage of all trajectories explore conformations with exactly the same base pairs as the native structure (second column), but, interestingly, these structures were chosen for submission and indeed correspond to the best predictions also in terms of RMSD. A possible explanation for the choice of the users comes from the observation of the stability of the molecule, which is not captured by the instantaneous structure they submitted. Indeed, native states are generally more stable than other states, as observed by our previous computer simulation studies for these same molecules, and users had the tendency to submit structures that remained stable for a while in the simulation.
We can observe that for 1N8X and 2G1W one native stem is clearly explored in the trajectories. This corresponds to the lower stem for 1N8X, formed by 7 pairs, and one or the other of the stems of 2G1W, which can both be composed of 4 pairs. Trajectories of 2K96 explore configurations with a wide range of base pairs, with the number of native base pairs not exceeding 40%. There is not a clear peak of the distribution, but all options seemed to be explored rather uniformly.

The comparison of topological values (column 3) gives a measure of the extent to which the base pairing organization of generated structures (trajectories or submitted) corresponds to the native secondary structure. For 1F9L and 1N8X the topology explored in the trajectories, and even more so that of selected structures, corresponds well to the native topology, suggesting that most users focused on the hairpin as their prediction for the molecule's architecture. Indeed

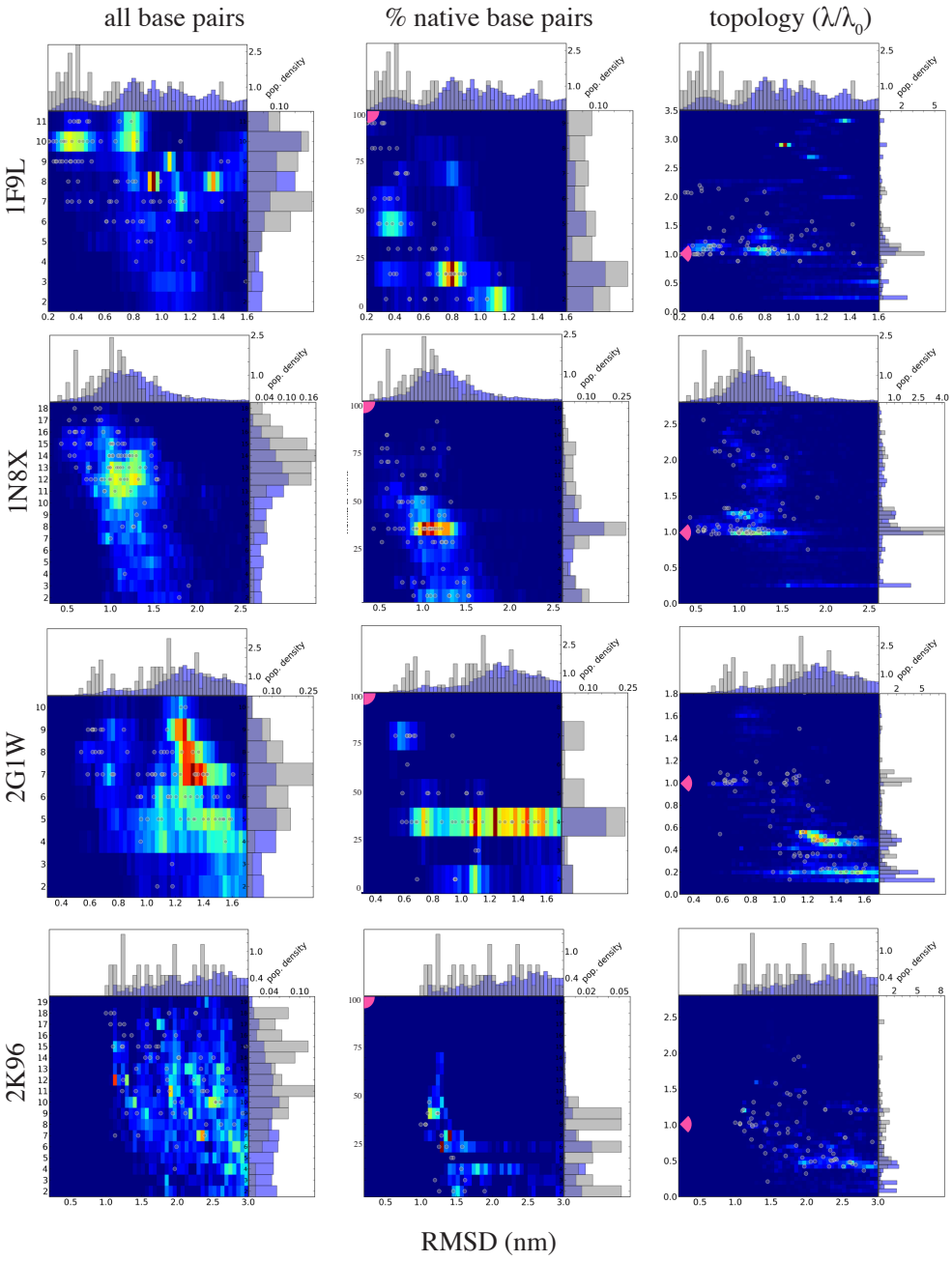

Figure 3: Base pairs vs. RMSD analysis for the non-feedback experiment: number of detected base pairs (left), percentage of native base pairs (center), molecule's topology as defined by the second eigenvalue of the Laplacian matrix. Eigenvalues are normalized with respect to the second eigenvalue of the Laplacian matrix λ_0 for the native structure. In the central and right columns the pink wedge corresponds to the position of the native structure (RMSD=0, % native pairs = 100, $\lambda_0/\lambda_{0,native} = 1$).

most trajectories focus on a topological parameter (λ, see SI) equal or close to native, and selected structures are very strongly peaked at the correct eigenvalue. If we consider dual graph topological parameters [37], for 1N8X 17% of all trajectories and 33% of selected structures share the native values of number of vertices and second eigenvalue of the Laplacian matrix, indicating that the overall base pair organization and stem-loop organization of the explored configurations correspond to native. For 2G1W analysis of topological parameters shows that full trajectories focused on configurations of topologies different than the pseudoknot (indeed most users tried to form hairpins), however submitted structures were chosen also from conformations of the correct topology, as it is shown by a peak of the distribution for $\lambda/\lambda_0 \sim 1$. For 2K96 the best predictions have the native value $\lambda \sim \lambda_0$, supporting the observation that even though the details of the structures are not predicted correctly the overall organization of submitted structures corresponds to native.

3.3 Interactive RNA folding open new opportunities

The fact that participants were quite successful in folding the four molecules and exploring phase space in a broad manner opens the prospect for applications. In research, such interactive simulations on unknown targets may provide complementary means to generate a pool of plausible structures. In combination with experimental data this can be a powerful tool to refine structural models. Teaching is another promising application area, as we noticed that many of the complex concepts associated to RNA conformational flexibility were quite naturally made aware to the participants. The interactive approach is also a wonderful tool for outreach activities.
A key question for research applications is whether one is able to select the correctly folded structures from the pool of all submissions. The current dataset suggests that structural clustering of the solutions combined with a low energy filter should lead to a good selection of candidate structures. In that context it should also be recalled that in some sense our experiment setup was not ideal, because the participants were only allowed quite basic tools, without 3D visualization of the structures nor use of 3D input devices that would facilitate the manipulation in space. Furthermore, available time for the experiment was limited. In particular for the more complex 2K96 molecule this limitation had an impact on what could be achieved. Another promising avenue for future extension would be to implement collaborative strategies whereby users would not only be able to work individually but also collectively. This is the route successfully taken by Foldit through the use of a scripting tool that allowed players to share their strategies [39]. Furthermore, one could imagine several participants working on distinct parts of one molecule at the same time, or cross-checking each others' solutions.
These preliminary results on molecular explorations by interactive simulations are encouraging especially if projected onto the direction of the use by the research community, where the average user would already have some prior knowledge of the possible motifs of nucleic acids systems and possibly modeling. One of the main directions of our continuing development of HiRE-RNA and UnityMol is to include different sources of experimental information. In the current simulation software, it is possible to include local restraints such as base-pairs, including information from secondary structure predictions, crystallographic data of subparts of the molecule, and preliminary NMR data. In a new version soon to be released, it will be possible to include the on-the-fly calculation of theoretical SAXS curves and compare them to a target curve as the simulation proceeds. The introduction of indirect experimental data to guide simulations has to be done with the awareness that a successful strategy is to explore very different regions of the conformational space, and not to focus on a restricted region of one single parameter. Indeed, the knowledge of a score with respect to a target structure, such as the RMSD in the feedback experiment appears to limit the conformational space that is explored by the users, which have then the tendency to remain a single region of good scores, instead of exploring more widely.

4 Conclusions and perspectives

The approach presented here fundamentally builds upon the interactive molecular dynamics family of approaches [40], yet provides many significant improvements, in particular the introduction of crowd sourcing aspects for harvesting user contributions. To our knowledge, this is the first large-scale participative experiment at a coarse-grained level of representation, whereas alternative approaches such as FoldIt focus on all-atom models. We previously demonstrated that the coarse grained level provides particular opportunities for interactive simulations [40]. In our approach, the physically sound simulation of the conformational dynamics is at the center of the experiment and it is guided by the user; in other folding challenges the user 3D puzzle is at the center of interest with limited contribution from modeling, mostly through instantaneous minimization. For our purpose, we extend the existing Interactive Molecular Dynamics (IMD) protocol with the possibility to steer simulation parameters such as temperature or to exchange experimental data used as additional constraint on the simulation. We provide several adapted real time analyses, such as live plots of relevant quantities to monitor the simulation, on-the-fly topology and secondary structure graphs as well as generation of experimentally relevant information, for instance a SAXS scattered intensity profile. Overall we propose an open design that others can build on for similar experiments, providing among others a convenient web application to harvest and manage participants' contributions.

The main result of our experiment was that through interactive simulations and a simplified representation of the molecule, "naive" users were able to successfully predict native RNA folds. The use of interactive non-equilibrium MD simulations, with the possibility of monitoring in real time certain features such as internal energies, not only allows the participants to explore the conformational space more widely and in different regions with respect to what is done by standard REMD computer simulations, but lead them to identify native-like structures and to explore more thoroughly their basins. The plurality of proposed structures is an advantage in folding predictions. Given the variability of experimental conditions that cannot be accounted for in simulations, the ability to quickly produce plausible alternative structures is indeed a valuable feature in the context of a real scientific research, in which the target structure is unknown and where possible conformations have to be selected based on their agreement with indirect experimental information. Submission of several different structures is also a winning strategy in RNA and protein folding competitions. Looking at all energy and base pairing plots together from non-feedback experiment, we observe that there is no straightforward correlation between energy and RMSD, nor between the number of base pairs and RMSD, yet the participants could reach a high success rate. Comparing the results from non-feedback experiment and feedback experiment, there is no increase of the success rate. This is surprising but expected since a single parameter is not sufficient to detect the native state. Participants were able, however to guide their molecules to native basins and to select native-like conformations, by acquiring chemical and physical intelligence that standard computer simulations based on the equations of motion, and energy calculations, do not have yet. This observation makes a strong argument for the pursuit of hybrid methods where the power of computers is combined with the creativity of humans.

With the amazing variety of RNA structures, are many more RNA folding challenges than those we presented here, biologically more interesting and more intriguing for predictions. However, our goal here was to demonstrate what naive users, with little background in nucleic acids structures and modeling, in just one day could go from learning to use the software to the proposing solutions in good agreement with experimental structures. Our ultimate goal is to provide this open software to experts in RNA structures and function, who are aware of the complexity of the structural details of single stranded nucleic acids, have a good knowledge of the NDB, and by intuition can test their ideas very rapidly without having to enumerate an excessively large

number of conformations.

Authors contributions

L.M. analyzed data and contributed to writing the paper, S.D. contributed the interactive simulation and visualization tools, developed the website, managed data collection and contributed to writing the paper, C.G. helped setting up the lab, P.D. contributed developing the coarse-grained model, wrote the MD simulation code and contributed to writing the paper, A.T. contributed in designing the in-class experiment, M.B. designed the simulation tools, contributed in designing the research set up and in writing the paper, S.P. designed the coarse-grained model, designed the research set-up and wrote the paper.

Acknowledgements

We wish to thank all third year biology majors of 2015 and 2016 at Paris Diderot University for enthusiastically taking part in this experiment. We also wish to thank the first year students of "Frontières du Vivant" curriculum at Paris Descartes University as well as the MOOC students "Bases Moléculaires de la Vie" for also testing our interactive simulations software and folding challenge, helping us properly set up the experiment. L.M. was supported by the "Initiative d'Excellence" program from the French government, project "DYNAMO", ANR-11-LABX-0011-01. S.D. was supported by the French National Agency for Research, proposals "ExaViz" (ANR-11-MONU-003) and "GRAL" (ANR-12-BS07-0017).

References

[1] Berman, H. M., J. Westbrook, Z. Feng, L. Iype, B. Schneider, and C. Zardecki. 2002. The nucleic acid database. *Acta Crystallogr D Biol Crystallogr.* 58:889–898.

[2] Leontis, N. B., J. Stombaugh, and E. Westhof. 2002. The non-watson-crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* 30:3497–3531.

[3] Cruz, J. A., and et al. 2012. Rna-puzzles: a casp-like evaluation of rna three-dimensional structure prediction. *RNA.* 18:610–625.

[4] Miao, Z., and et al. 2015. Rna-puzzles round ii: assessment of rna structure prediction programs applied to three large rna structures. *RNA.* 21:1066–1084.

[5] Miao, Z., and et al. 2017. Rna-puzzles round iii: 3d rna structure prediction of five riboswitches and one ribozyme. *RNA.* 23:655–672.

[6] Cheng, C. Y., F.-C. Chou, and R. Das. 2015. Modeling complex rna tertiary folds with rosetta. *Methods Enzymol.* 553:35–64.

[7] Parisien, M., and F. Major. 2008. The mc-fold and mc-sym pipeline infers rna structure from sequence data. *Nature.* 452:51–55.

[8] Turner, D. H., and D. H. Mathews. 2010. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.* 38:D280–D282.

[9] Zuker, M., and P. Stiegler. 1981. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9:133–148.

[10] Nussinov, R., and A. B. Jacobson. 1980. Fast algorithm for predicting the secondary structure of single-stranded rna. *Proc Natl Acad Sci U S A.* 77:6309–6313.

[11] Ruan, J., G. D. Stormo, and W. Zhang. 2004. An iterated loop matching approach to the prediction of rna secondary structures with pseudoknots. *Bioinformatics.* 20:58–66.

[12] Rivas, E., and S. R. Eddy. 1999. A dynamic programming algorithm for rna structure prediction including pseudoknots. *J Mol Biol.* 285:2053–2068.

[13] Lyngsoe, R. B., and C. N. Pedersen. 2000. Rna pseudoknot prediction in energy-based models. *J Comput Biol.* 7:409–427.

[14] Xia, Z., D. R. Bell, Y. Shi, and P. Ren. 2013. Rna 3d structure prediction by using a coarse-grained model and experimental data. *J Phys Chem B*. 117:3135–3144.

[15] Ding, F., S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan. 2008. Ab initio rna folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*. 14:1164–1173.

[16] Sulc, P., F. Romano, T. E. Ouldridge, J. P. K. Doye, and A. A. Louis. 2014. A nucleotide-level coarse-grained model of rna. *J Chem Phys*. 140:235102.

[17] Pasquali, S., and P. Derreumaux. 2010. Hire-rna: a high resolution coarse-grained energy model for rna. *J Phys Chem B*. 114:11957–11966.

[18] Cragnolini, T., P. Derreumaux, and S. Pasquali. 2015. Ab initio rna folding. *J Phys : Condens Matter*. 27:233102.

[19] Stadlbauer, P., L. Mazzanti, T. Cragnolini, D. J. Wales, P. Derreumaux, S. Pasquali, and J. Sponer. 2016. Coarse-grained simulations complemented by atomistic molecular dynamics provide new insights into folding and unfolding of human telomeric g-quadruplexes. *J Chem Theory Comput*. 12:6077–6097.

[20] Cho, S. S., D. L. Pincus, and D. Thirumalai. 2009. Assembly mechanisms of rna pseudoknots are determined by the stabilities of constituent secondary structures. *Proc. Natl. Acad. Sci. U.S.A.* 106:17349–17354.

[21] Sulc, P., T. E. Ouldridge, F. Romano, J. P. K. Doye, and A. A. Louis. 2015. Modelling toehold-mediated rna strand displacement. *Biophys J*. 108:1238–1247.

[22] Zhang, T., J. Zhang, P. Derreumaux, and Y. Mu. 2013. Molecular mechanism of the inhibition of egcg on the alzheimer aβ1–42 dimer. *The Journal of Physical Chemistry B*. 117:3993–4002.

[23] Cooper, S., F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and F. Players. 2010. Predicting protein structures with a multiplayer online game. *Nature*. 466:756–760.

[24] Lee, J., W. Kladwang, M. Lee, D. Cantu, M. Azizyan, H. Kim, A. Limpaecher, S. Yoon, A. Treuille, R. Das, and E. Participants. 2014. Rna design rules from a massive open laboratory. *Proc Natl Acad Sci U S A*. 111:2122–2127.

[25] Lv, Z., A. Tek, F. Da Silva, C. Empereur-mot, M. Chavent, and M. Baaden. 2013. Game on, science - how video game technology may help biologists tackle visualization challenges. *PLoS One*. 8:e57990.

[26] Doutreligne, S., C. Gageat, T. Cragnolini, A. Taly, S. Pasquali, P. Derreumaux, and M. Baaden. 2015. Unitymol: interactive and ludic visual manipulation of coarse-grained rna and other biomolecules. *In* Virtual and Augmented Reality for Molecular Science (VARMS@ IEEEVR), 2015 IEEE 1st International Workshop on. IEEE. 1–6.

[27] Sterpone, F., S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. H. Nguyen, and P. Derreumaux. 2014. The opep protein model: from single molecules, amyloid formation, crowding and hydrodynamics to dna/rna systems. *Chem Soc Rev*. 43:4871–4893.

[28] Chebaro, Y., S. Pasquali, and P. Derreumaux. 2012. The coarse-grained OPEP force field for non-amyloid and amyloid proteins. *J Phys Chem B*. 116:8741–8752.

[29] Nguyen, P. H., Y. Okamoto, and P. Derreumaux. 2013. Communication: Simulated tempering with fast on-the-fly weight determination. *Journal Chem Phys*. 138:061102.

[30] Cragnolini, T., Y. Laurin, P. Derreumaux, and S. Pasquali. 2015. Coarse-grained hire-rna model for ab initio rna folding beyond simple molecules, including noncanonical and multiple base pairings. *J Chem Theory Comput*. 11:3510–3522.

[31] Pérez, S., T. Tubiana, A. Imberty, and M. Baaden. 2015. Three-dimensional representations of complex carbohydrates and polysaccharides-SweetUnityMol: A video game-based computer graphic software. *Glycobiology*. 25:483–491.

[32] Rudisser, S., and I. Tinoco. 2000. Solution structure of cobalt(iii)hexammine complexed to the gaaa tetraloop, and metal-ion binding to g.a mismatches. *J Mol Biol*. 295:1211–1223.

[33] Lawrence, D. C., C. C. Stover, J. Noznitsky, Z. Wu, and M. F. Summers. 2003. Structure of the intact stem and bulge of hiv-1 psi-rna stem-loop sl1. *J Mol Biol*. 326:529–542.

[34] Nonin-Lecomte, S., B. Felden, and F. Dardel. 2006. Nmr structure of the aquifex aeolicus tmrna pseudoknot pk1: new insights into the recoding event of the ribosomal trans-translation. *Nucleic Acids Res*. 34:1847–1853.

[35] Kim, N.-K., Q. Zhang, J. Zhou, C. A. Theimer, R. D. Peterson, and J. Feigon. 2008. Solution structure and dynamics of the wild-type pseudoknot of human telomerase rna. *J Mol Biol*. 384:1249–1261.

[36] Gan, H. H., S. Pasquali, and T. Schlick. 2003. Exploring the repertoire of rna secondary motifs using graph theory; implications for rna design. *Nucleic Acids Res*. 31:2926–2943.

[37] Fera, D., N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick. 2004. Rag: Rna-as-graphs web resource. *BMC Bioinf.* 5:88.

[38] Xu, X., P. Zhao, and S.-J. Chen. 2014. Vfold: a web server for rna structure and folding thermodynamics prediction. *PLoS One.* 9:e107504.

[39] Khatib, F., S. Cooper, M. D. Tyka, K. Xu, I. Makedon, Z. Popovic, D. Baker, and F. Players. 2011. Algorithm discovery by protein folding game players. *Proc Natl Acad Sci U S A.* 108:18949–18953.

[40] Delalande, O., N. F'erey, G. Grasseau, and M. Baaden. 2009. Complex molecular assemblies at hand via interactive simulations. *J Comput Chem.* 30:2375–2387.