RNA fast-folding path for the prediction of secondary structures and folding dynamics

Vaitea Opuu[1,‡], Nono S. C. Merleau[1], and Matteo Smerlak[1]

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[‡]Contact email: vopuu@mis.mpg.de

May 15, 2021

Abstract

The biological roles of non-coding RNA are better understood with their structural features. The static structure predictions of RNAs have seen tremendous progress in the thermodynamic and machine learning approaches. But an understanding of dynamical aspects can provide for complementary biological insights. Here, we propose a method to predict RNA structures and folding dynamics. This method has been inspired by the RNA fast-folding paths principle. For this, We developed an efficient algorithm exploiting the fast Fourier transform and the stem rate model. It predicts multiple parallel folding paths by following the thermodynamic energy landscape. When only a single prediction per RNA was considered, this method's performance for the folding task was only fair. However, when all structures found were analyzed, we found near-native predictions (79% PPV and 81% sensitivity) for small RNA (200 nucleotides). On average, those predictions were found to be of similar quality to recent deep-learning-based methods. Furthermore, from these trajectories, we built up a folding kinetic ansatz that allows extracting even more dynamical information. Although only as few as $\sim$ 60 structures allowed us to produce relevant folding kinetic trajectories while known methods may require millions of them. Because of its simple foundations, this work can help develop new insights into RNA structure roles.

INTRODUCTION

Natural RNAs have various essential functions in many cellular contexts such as in the protein translation machinery (mRNA, tRNA...) [24]. In addition to its important place in the central dogma of molecular biology, it also plays roles in gene regulations (microRNAs) [1]. In biotechnology, RNAs also found its way with, for example, the design of biosensors or ribozymes [13]. With ribozymes, RNAs entered the realm of proteins where RNAs can perform enzymes functions. These functions are generally understood through the lens of one static tri-dimensional structure. This structure to function paradigm is used by most of the current frameworks to understand relationships between RNAs. Since experimental determination of their structures by X-ray or nuclear magnetic resonance are usually out of reach, these analysis

are based on predictions. Unfortunately, some important RNAs like riboswitches have their biological function tightly bound to their dynamical behavior [33]. A more complete description of RNAs static and dynamical structural features is therefore desirable. And given the ever so increasing number of RNA sequences, fast methods with such scopes could help in tasks like refining data bases annotations. On the biotechnological side, such a method could help develop new design strategies for synthetic RNAs [7].

Three main structure levels are generally considered to describe RNA molecules: the primary structure is the nucleotide sequence itself. The secondary structure is defined by interacting pairs of nucleobases called base pairs. The tertiary structure involve other weaker non-trivial interactions within the same sequence. Unlike proteins, RNA structures are usually hierarchically formed. The secondary structure is formed first, followed by the tertiary structure [31]. Moreover, the secondary structure provides an accurate enough description of the thermodynamics and kinetics of RNA molecules. Although base pairs can be formed with various configurations [17], we considered here only the canonical base pairs edge-to-edge interactions: G-C, A-U, and G-U. Many formal subtleties can be used to define the secondary structure, but we used here the formal definition called pseudoknot-free. The important consequence is that each base pair formed separates the RNA into two completely independent sections, the interior and the exterior sections. In the rest of this work, structure refers to the RNA secondary structure.

The structure space of an RNA molecule is described by the stability of all possible structures. The stability ΔG_s of a structure s is the free energy changes with the completely unfolded state. To predict biologically relevant structures, most of the current methods rely on free energy minimization or stability maximization. The nearest-neighbor loop energy model is the most used model to compute stabilities [32]. By assuming the additivity principle [6], the stability of a structure is a sum of independent loop contributions. This model consists of a set of tabulated parameter values associating free energy values to loop types and nucleotide compositions. The Turner2004 [22] is one of the widely used set of parameters. Its functional form allows for general energy parameters and the use of an efficient dynamic programming algorithm. This algorithm can determine the minimum free energy (MFE) structure of a sequence in the structure space. The MFE is considered a gold standard for free-energy-based predictions, however, it represents only one structural estimate among other. Other estimates exist such as the maximum expected accuracy (MEA), however, it was not found to be significantly better than the MFE [21].

Several tools implement the MFE search algorithm, namely Zucker algorithm [38], such as `RNAfold` [14], `Mfold` [37], or `RNAstructure` [25]. Although those methods were found to be consistently accurate at predicting RNA secondary structures as shown in recent benchmarks [28, 15], the additivity foundation is expected to be doomed when sequences get larger and structures complexify. Moreover, like tertiary interactions, pseudoknots loop are not defined in the main parameters sets like the Turner2004 model. The discrepancy for larger RNAs could be explained by the fact that tertiary interactions and pseudoknots are neglected. Machine learning (ML) approaches were investigated to overcome some of these shortcomings. The ML structure estimate provides substantial improvements in structure prediction according to recent benchmarks [29, 28]. But, in addition to some over-fitting concerns [26], these approaches cannot give dynamical information since too few data are available, to this date, on structure dynamics. This structure estimate represents a complex view of the RNA folding. Indeed, structural data are largely obtained through phylogenetic analyses,

From a dynamical standpoint, the RNA molecule navigates the structure space by following the landscape drawn by the stability. To follow the dynamic of individual RNAs, three rate models describing elementary steps in the structure space are currently used. First, the base stack model uses base stacks formations and breaking as elementary moves [35]. The second model uses base pair as elementary steps. `kinfold` [9] uses the base pair rate and a continuous-time Monte Carlo simulation to follow the RNA folding dynamic. It gives the finest resolution in the secondary structure folding landscape, but at the cost of computation time. The third model uses the creation or deletion of stems to construct the folding dynamics. It is the first strategy explored [20], and provides a coarse-grained description of the dynamics. The folding rates are determined by the free energy changes when stems are added or removed. Although none of these models were definitively rejected nor accepted, this one makes a notable assumption. Indeed, transition states (or saddle points) hidden in the formation of a given stem are not considered [36]. An alternative approach, implemented in `kinwalker` [11], used the observation that folded intermediates are generally locally optimal conformations. Therefore, locally optimal structures are formed using the standard dynamic programming algorithm and aggregated together along with the folding procedure.

From folding experiments, Pan and coworkers found parallel pathways for a ribozyme which involve two types of path to reach the native structure [23]. One population of sequences was found to fold rapidly, and one quickly reached metastable misfolded structures that slowly fold into the native structure. However, in some cases, the metastable states are functional, this is a direct consequence of the rugged nature of the RNA folding landscape [30]. Russell and coworkers revealed experimentally the presence of multiple deep channels separated by large energy barriers on the folding landscape which lead to the fast and slow folding paths observed [27]. The formal description of this mechanism, called kinetic partitioning mechanism, was introduced by Guo and Thirumalai on proteins first [12]. In the free energy landscape, those metastable conformations are competing attraction basins from which RNA molecules are temporarily trapped.

Here, we propose a novel approach for RNA structure predictions and dynamics. This method has been inspired by the fast-folding path idea and built upon intuitive folding rules. The basic idea is to use the stem rate model to create multiple parallel folding paths. It sequentially forms stems along the folding trajectory if the stability is improved. Once a stem is formed, it cannot be removed. To speed up the search of stems, RNA sequences are encoded in a numerical fashion we called mirror encoding. This encoding combined with the fast Fourrier transform allowed for a quick search of stems. This algorithm is inspired by MAFFT [16], a well-known multiple-sequence-alignment tool. The use of signal processing techniques to analyze nucleotide sequences has been investigated since the early 80's [8, 4], however, to our knowledge, this its first time use in an RNA folding algorithm.

To assess the reliability of the paths predicted, we compared its performance on the folding task for a well-curated dataset, archive II [21]. This dataset is composed of The algorithm predictions were compared to two structure estimates: the MFE structure computed by `RNAfold` and the ML structure computed with `MxFold2` [28].

The low energy structure may not be the biological structure, as shown in the accuracy assessment of the MFE. This can be explain by the energy model limits but not only. Some RNAs may have their active structure into kinetics traps far from the MFE. In some cases, the MFE may not even be reachable in biologically relevant time. With the ensemble of paths produced for each sequence, we also derived a folding kinetic ansatz. Next, we applied the

algorithm to a simple test case, the Coronavirus frameshifting stimulation element [3], and a classic bi-stable sequence. These experiments allowed to find structures closer to the native one for the biological first example. From the folding intermediates obtained from RAFFT, we built a kinetic model from which we recovered trajectories qualitatively similar to some trajectories obtained from the barrier tree kinetic [10].

MATERIALS AND METHODS

The folding algorithm

We now describe the heuristic starting from one sequence of nucleotides $S = (S_1 \dots S_L)$ of length L, and its associated unfolded structure. We first create a numerical representation of S where each nucleotide of S is replaced by of one unit vector of 4 components:

$$A \to \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, U \to \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, C \to \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, G \to \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}. \tag{1}$$

This gives us a $(4 \times L)$-matrix we call X where each row corresponds to a nucleotide type as shown below

$$X = \begin{pmatrix} X^A \\ X^C \\ X^G \\ X^U \end{pmatrix} = \begin{pmatrix} X^A(1) & X^A(2) & \dots & X^A(L) \\ X^C(1) & X^C(2) & \dots & X^C(L) \\ X^G(1) & X^G(2) & \dots & X^G(L) \\ X^U(1) & X^U(2) & \dots & X^U(L) \end{pmatrix} \tag{2}$$

where, for example, $X^A(i) = 1$ if $S_i = A$. Next, we create a second copy $\bar{S} = (S_L \dots S_1)$ for which we reverted the sequence order. Then, each nucleotide of $\bar{S}$ is replaced by one of the following unit vectors:

$$\bar{A} \to \begin{pmatrix} 0 \\ 0 \\ 0 \\ w_{\text{AU}} \end{pmatrix}, \bar{U} \to \begin{pmatrix} w_{\text{AU}} \\ w_{\text{GU}} \\ 0 \\ 0 \end{pmatrix}, \bar{C} \to \begin{pmatrix} 0 \\ 0 \\ w_{\text{GC}} \\ 0 \end{pmatrix}, \bar{G} \to \begin{pmatrix} 0 \\ w_{\text{GC}} \\ 0 \\ w_{\text{GU}} \end{pmatrix}. \tag{3}$$

$\bar{A}$ (respectively $\bar{U}, \bar{C}, \bar{G}$) is the complementary of A (respectively U, C, G). w_{AU}, w_{GC}, w_{GU} are tunable parameters that represent the weight associated with each canonical base pair. These parameters are chosen empirically. We call this complementary copy $\bar{X}$, the mirror of X.

To search for stems, we use the complementary relation between X and $\bar{X}$ with the correlation function $\text{cor}(k)$. This correlation is defined as the sum of individual X and $\bar{X}$ row correlations

$$\text{cor}(k) = c_{X^A,\bar{X}^A}(k) + c_{X^U,\bar{X}^U}(k) + c_{X^G,\bar{X}^G}(k) + c_{X^C,\bar{X}^C}(k) \tag{4}$$

where one row correlation between X and $\bar{X}$ is given by

$$c_{X^\alpha,\bar{X}^\alpha}(k) = \frac{1}{\min(k, 2 \times L - k)} \sum_{\substack{1 \le i \le L \\ 1 \le i+k \le L}} X^\alpha(i) \times \bar{X}^\alpha(i+k). \tag{5}$$

For each $\alpha \in \{A, U, C, G\}$, $X^{\alpha}(i) \times \bar{X}^{\alpha}(i+k)$ is non zero if sites i and $i+k$ can form a base pair, and will have the value of the chosen weight as described above. If all the weights are set to one, $\mathrm{cor}(k)$ gives the frequency of base pairs for a positional lag k. Although the correlation requires $O(L^2)$ operations, it can take advantage of the FFT which reduces drastically its complexity to $O(L \log(L))$.

The large $\mathrm{cor}(k)$ values between the two copies indicate the positional lag k at which the frequency of base pair is high. Indeed, this does not allow to determine the exact stem positions. Hence, we use a sliding window strategy to search for the largest stem within the positional lag. Since the copies are symmetrical, we only need to slide over one-half of the positional lag. Once the largest stem is identified, we compute the free energy change associated with the formation of the stem. We perform the same search for the n highest correlation values, which gives us n potential stems. Then, we fix into the current structure the stem that give the best change of free energy. Here, free energies were computed using Turner2004 energy parameters through Vienna RNA package API [18].

We are now left with two independent parts, the interior, and the exterior of the stem formed. If the exterior part is composed of two fragments, they are concatenated into one. Then, we simply apply recursively the same procedure on the two segments independently in a "Breadth First" fashion to form new consecutive base pairs. The procedure stops when no base pair formation can improve the energy. Given this simple recursive scheme, it is straightforward to consider pseudoknots by simply concatenating both parts. When multiple stems can be formed in these independent fragments, we combine all the possible independent stems and pick the composition that has the best overall stability. If too many composition can be formed, we restrict this to the 10^4 bests in term of energy. Figure 2 shows an example of execution to illustrate the procedure. The complexity of this algorithm depends strongly on the number and the size of the stems formed. The best case is the trivial structure composed of one large stem where the complexity correspond to the correlation evaluation for the whole sequence. The worst case is an idealized case where at most $L/2$ base pairs can be formed (assuming L is even). The rough complexity depends on $\sum_{1 \leq i \leq L/2} 2i \times \log(2i) \approx \frac{L^2}{2.} \times log(L) + \delta$ where δ is small compared to L^2. Therefore, the procedure has a worst rough complexity of $O(L^2 \log(L))$.

The algorithm described so far tends to be stuck in the first local minima found along the folding trajectory. To alleviate this, we implemented a stacking procedure where the N best trajectories are stored in a stack and evolved in parallel. Figure 1 illustrates this modified procedure. Like the initial version, the procedure starts with the unfolded structure. Then, the $N = 5$ best potential stems are stored at the first stack. From these N structures, the procedure tries to add stems in the unpaired regions left and save the N best structures formed. Once no stem can be formed, the algorithm stops and output the structure with the best energy found among the structures saved in the last stack. This procedure leads to the construction of a graph we call fast-folding graph. In this graph, two structures are connected if the transition from one to the other correspond to the formation of a stem.

Kinetic ansatz analyses

The folding kinetic ansatz used here is derived from the fast-folding graph. As described in figure 1, transitions can occur from left to right (and right to left) but not vertically. Two adjacent structures x and y in this graph are connected if the transformation from x to y (or y to x) only requires the formation of a stem or if x and y are the same structure. The fast-

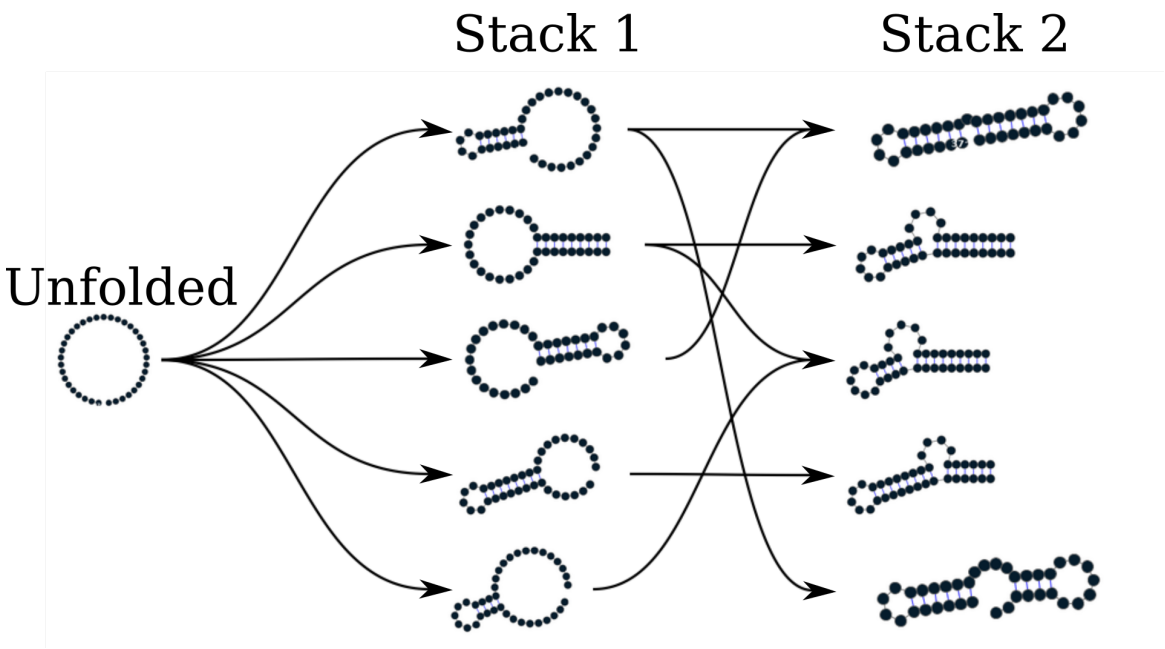

Figure 1: **Fast folding graph derived from parallel folding paths.** In this example, the sequence is folded in two steps. The procedure starts with the unfolded structure in the left. $N = 5$ best stem formed are saved in the stack 1. From stack 1, multiple stems formation are considered, but only the N best are stored in the stack 2. Structures are ordered (from top to bottom) by energy in each stack. All the secondary structure visualization were obtained using VARNA [5].

folding graph follows the idea that parallel paths quickly reach their end points. If the end points are non-native states, those structures will slowly fold back into the native state [23]. To simulate this behavior, we use the population kinetics. As usually done, the kinetic is modeled as a continuous time Markov chain [19], where populations of structure evolve according to a network of structures and the transition rates between structures. The fast-folding graph is used here as network of structures. The Arrhenius formulation is commonly invoked to derive the transition rates $r(x \to y) \propto \exp(-\beta E^{\ddagger})$ where $E^{\ddagger}$ is the activation energy separating x from y, and β is the inverse thermal energy (mol/kcal). However, here, we chose the transition rates $r(x \to y)$ to be based on the Metropolis scheme defined as follow

$$r(x \to y) = k_0 \times \min(1, \exp(-\beta \Delta\Delta G(x \to y))) \tag{6}$$

where $\Delta\Delta G(x \to y)$ is the stability change between structure x and y. Therefore, this does not yield the traditional kinetic analysis but an ansatz. k_0 is a conversion constant that we set to 1 for the sake of simplicity. This rate is non-zero if y is connected to x in the graph (or y is in the neighborhood of x, $y \in \mathcal{X}$). Here, we initialize the population p_0 with only unfolded structures, therefore, this represents a complete folding mechanism. The population change of a structure x is given by:

$$\frac{\mathrm{d}p_x}{\mathrm{d}t} = \sum_{y \in \mathcal{X}} r(y \to x) p_y(t) - r(x \to y) p_x(t) \tag{7}$$

where the sum is running over the neighborhood $\mathcal{X}$ of x. $p_x(t)$ is the population of x at time t.

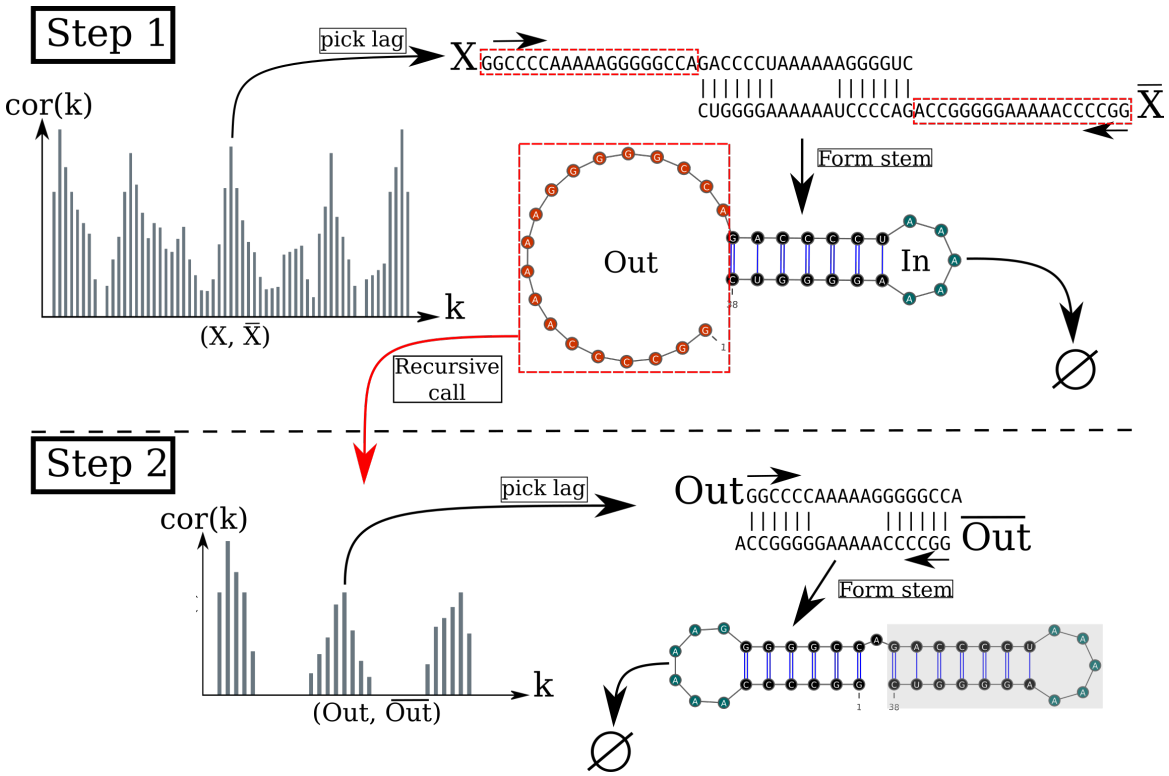

Figure 2: **Algorithm execution for one example sequence which requires two steps.** (Step 1) From correlation $(X, \bar{X})$, we pick one peak which corresponds to a position lag. We search for the largest stem and form it. Two fragments, "In" and "Out", are obtained, but only the "Out" may contain a new stem to add. (Step 2) The procedure is call recursively on the "Out" sequence fragment only. It produces a new positional lag from which we form a new stem. The fragment left (colored in blue) do not contain any additional stem, so the procedure stops.

Benchmark dataset

To build the dataset for the folding task application, we started from the ArchiveII dataset. We first removed all the structures with pseudoknots since all tools considered here don't handle pseudoknots. Next, we removed all the structures which were evaluated with positive or null energy with the Turner2004 energy parameters. Since positive energies mean that the completely unfolded structure is more stable than the native one. Those structures are assumed not well modeled by the energy function used here and therefore would blur the interpretation of the kinetic we try to extract. This dataset is composed of 2698 structures. 240 sequences were found multiple times (from 2 to 8 times). 19 of them were found with different structures. We discarded all duplication and picked the structure with the lowest energy for each. We obtained a dataset of 2296 sequences.

Structure prediction protocols for benchmarks

To evaluate the structure prediction power of the proposed method, we compared it to two structure estimates: the MFE structure, and one ML structure. To compute the MFE structure, we used `RNAfold` with the default parameters and the Turner2004 set of energy parameters. For the ML structure, we computed the prediction using `Mxfold2` with the default parameters. Therefore, only one structure prediction per sequence for those two methods were used for the statistics.

Two parameters are critical for RAFFT, the number of positional lags in which stems are searched and the number of saved configurations in the stack. For the experiments, we search for stems in the 100 best positional lags and stored 50 conformations. The $\mathrm{cor}(k)$ which allow to choose the positional lags is computed using the weights $w_{GC} = 3$, $w_{AU} = 2$, and $w_{GU} = 1$.

To compute the performance of RAFFT, we analyzed the output in two ways. First, we only displayed the structure with the lowest energy found for each sequence. Second, we compute the accuracy of all structures in the last stack predicted, and displayed the best structure.

To measure the prediction accuracy, we used two metrics from epidemiology. The positive predictive value (PPV) is the fraction of correct base pairs predictions in the predicted structure. The sensitivity is the fraction of correctly predicted base pairs in the true structure. Both metrics are defined as follow

$$PPV = \frac{TP}{TP + FN}, \quad \text{Sensitivity} = \frac{TP}{TP + FP} \tag{8}$$

where TP, FN, and FP stand respectively for the number of correctly predicted base pairs (true positives), the number of base pairs not detected (false negatives), and the number of wrongly predicted base pairs (false positives). To maintain consistency with previous and future studies, we computed these metrics using the implementation in the `scorer` tool provided by Matthews and coworkers [21], which provide also a more flexible estimate where shifts are allowed.

Structure space visualization

To visualize the structure loop diversity in the ensembles of structures considered here, we used the Principal component analysis (PCA). For one ensemble of structure. We first extracted the loop compositions in percent for each structure of the ensemble. To obtain the loop composition, we first convert the structures into Shapiro notation using Vienna Package API. From the

notation, we extracted the sizes of interior, exterior, bulge, stacking, hairpins, and multibranch loops. We obtain a table of 6 features and n entries. This allows us to compute a 6×6 correlation matrix that we diagonalize using the `eigen` routine implemented in the *scipy* package. For visual conveniences, the structure compositions were projected onto the first two principal components (PC). The direction of each feature in the PC space are represented with arrows.

RESULTS

Application to the folding task

To evaluate the relevance of the folding method, we assessed its performance for the folding task. Also, to measure the effect of sequence length, we analyzed their performance length-wise. We compared the method with two structure estimates: the MFE structure computed by `RNAfold` and the ML-based structure computed by `MxFold2`. For RAFFT, we saved 50 structures for each sequence.

Figure 3 A shows the performance in predicted positive values (PPV) and sensitivity for the three methods. It shows that the ML method is consistently better than RAFFT and MFE predictions, the thermodynamic methods. The length-wise T-test between the MFE and ML predictions showed that this difference is significant (p-value $\approx 10^{-12}$) with a substantial improvement of about 10%. Although RAFFT predictions were found to be comparable to the MFE predictions, they are significantly less accurate (p-value ≈ 0.0002), with a drastic loss of performance for sequences of length greater than 300 nucleotides.

Among the 50 structures saved per sequence with RAFFT, we found on average at least one prediction with 59% of PPV and 63% of sensitivity as shown figure 3 A. The overall gain of performances is not significantly different from the MFE predictions. However, for the sequences of length below 200 nucleotides, this gain was found to be substantial and significant ($\approx$ 16 % better than the MFE) with PVV $\approx$ 79% and sensitivity $\approx$ 81%. The accuracy for these predictions is equivalent to ML performances. For sequence lengths greater than 300 nucleotides, we observed the same drastic loss of accuracy, although we extracted the best prediction among the 50 saved configurations for each sequence. We investigated the dependency to the base pair spanning, however, we did not find any striking effect (see supp. mat.).

Two regions of lack of performance were observed for all methods. A group of 28 sequences of length shorter than 80 nucleotides have their known structures at on average 9.8 kcal/mol greater than the MFE structures. Some of them involve large unpaired loops such as displayed in figure 3 B. The second region is around 200 nucleotides in length. The known structure of these sequences also displayed large unpaired regions (figure 3 B).

To investigate the region of the structure space where the thermodynamic model tends to fail, we extracted a view of the different structure space produced by each method and the known structures. Figure 3 C shows principal component analysis (PCA) of the structure compositions in term of percent of loops. From the PCA, we observed that the known structures are distributed in the structure space toward interior loops. Also, some natural structures, as shown in figure 3 C, have large unpaired loops. The center of mass in the principal component space is located in between the high-density stacking and interior loops.

Next, we investigated the structure space produced by the three methods. The thermodynamic approach seems to produce a more diverse structure space as shown in figure 3 D. Loop

contents were extracted from the predicted structures of each method and projected onto their respective two first principal components space. Both RAFFT and MFE predictions seemed to produce similar structure spaces. The ML method allowed for long unpaired regions such as hairpins which tend to be closer to the structure space produced by the dataset.

Selected applications of the kinetic ansatz

The active structure observed is, in some cases, in kinetic trap far from the MFE on the free energy landscape. To illustrate those phenomenon, we applied folding kinetic ansatz proposed to two RNAs: the Coronavirus frameshifting stimulation element and one classic bi-stable sequence. The Coronavirus frameshifting stimulation element is an RNA sequence of about 82 nucleotides with a secondary structure determined by sequence analysis and obtained from the RFAM database. The known structure has a pseudoknot but was not taken into account here. Figure 4 panel A and B show the fast-folding graph, the MFE prediction, and the known structure. The folding mechanism is predicted in at most four steps where 20 structures were stored and 100 positional lags were searched for stems. RAFFT was able to recover near-native structures, found to be closer than the MFE. Nevertheless, The greediness effect can be easily spotted at step two in the fast-folding graph. One intermediate leading to the MFE structure is ranked 9. Hence, if less than 9 structure are stored, the MFE structure cannot be obtained.

To visualize the folding landscape drawn by RAFFT, we mapped all 68 unique structures found onto a plan using the multidimensional scaling (MDS) algorithm. On this landscape, the MDS optimized the mapping in such a way that the structure base pair distances are mostly preserved. Figure 4 panel D shows the landscape interpolated with the unique structures found. It illustrates the two states folding landscape where all trajectories started from the high peak in the center, and smoothly roll down to the good stability area (blue area).

Figure 4 shows the kinetic obtained from the fast-folding graph. The near-native structure 44 dominates the kinetic between t=10^2 to t=10^{13}. Its stability was evaluated at -23.2 kcal/mol. Then, one of the MFE structures dominated the population from t=10^{13} to the t=10^{15} and has been evaluated with a stability of -25.8 kcal/mol. As shown in figure 4, long-live intermediates are structures that couldn't be folded any further, such as structures 25, 27, 28, or 30. One speculative folding scenario that may support this kinetic model is the formation of the pseudoknot when the structure 44 dominates the population, which may have fixed the observed secondary structure.

Next, we compared the fast-folding kinetics with the widely used barrier tree kinetics [10]. First, we generated the 1.5×10^6 suboptimal structures at 15 kcal/mol from the MFE structure using `RNAsubopt` [18]. Since the MFE structure is evaluated at -25.8 kcal/mol, the unfolded structure couldn't be sampled. Next, the ensemble has been coarse-grained into 40 basins and presented as a barrier tree in figure 5 A. Here, the kinetics represent the competition between basins. The coarsening was performed by the tool `barrier` [10]. The barrier tree displayed a bi-stable topology. The near-native structure is the deepest basin in the lower part of the tree (its structure ID is 4). From the barrier tree and the Arrhernius formula [34], we used the tool `Treekin` to derive transition rates and compute folding kinetics. Since the unfolded structure couldn't be reached, we chose two other structures to initialize two kinetic trajectories (structures I1 and I2). Figure 5 C shows the folding kinetic where the near-native and MFE structures are dominating the population at different interval such as displayed by fast-folding kinetic model in figure 4. When the trajectory was initialized in the same branch than the

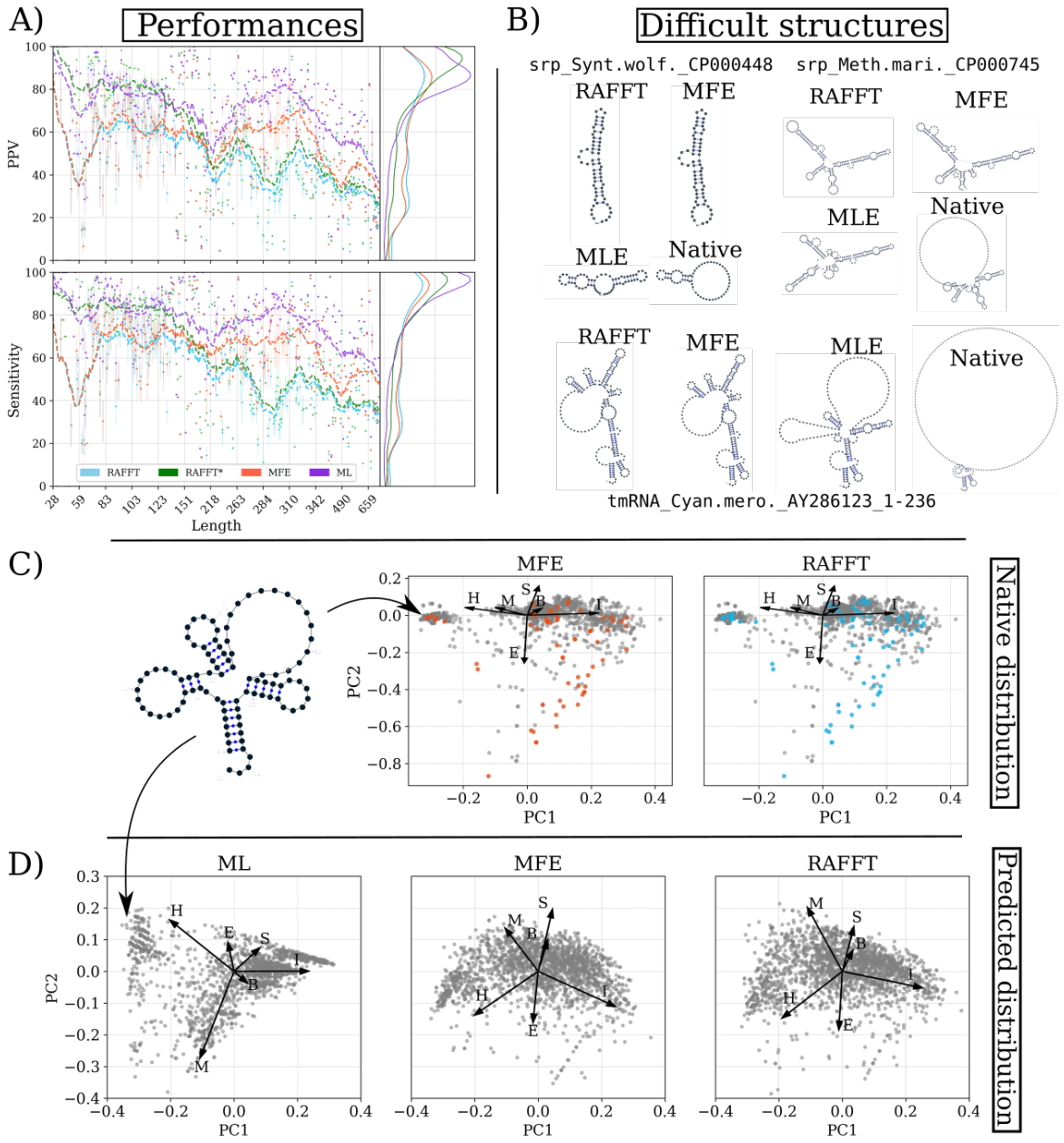

Figure 3: (A) Performance measured by PPV and sensitivity. RAFFT (blue) displayed the best energy found. RAFFT* (green) shows the best score found among the 50 saved structures for each prediction. The right pans of both figures show the distribution of PPV and sensitivity sequence-wise. (B) Structures found to be difficult to predict with the thermodynamic model. The sequence name where extracted directly from the dataset. (C) PCA analysis based on the native structures in the benchmark dataset. One example of structure found in the high density hairpins **H** is shown in the left. Both PCAs shows the same distribution, but on MFE (respectively RAFFT) shows in orange (respectively in blue) the structures with a PPV $\leq 10\%$. (D) The PCA for the predicted structures obtained with all three methods.

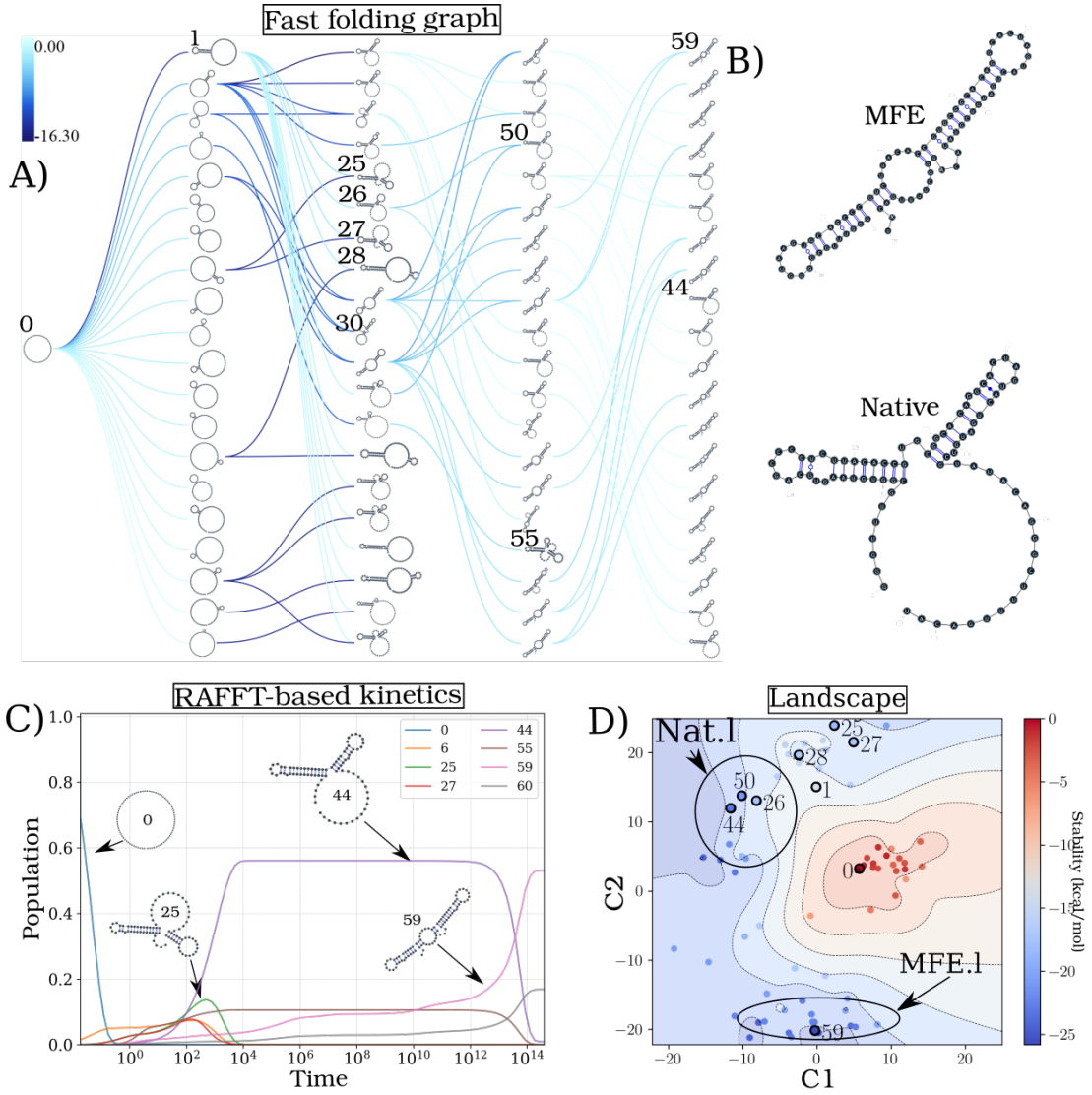

Figure 4: (A) Fast-folding path in four steps and 20 saved structures. The edges are colored according to the $\Delta\Delta G$ of stability. At each step, the structures are ordered by their stability from top to bottom. The minimum energy structure found is at the top left of the graph. Visited structures in the kinetic are annotated by a unique ID. (B) MFE (computed with RNAfold) and the native structures of the Coronavirus frameshifting element. (C) Kinetic of structures with an arbitrary time. (D) Folding landscape derived from the 68 unique structures found by RAFFT. The axes are the components optimized by the MDS algorithm so the base pair distances are mostly preserved. Observed structures are also annotated using the unique ID. The MFE like structures (**MFE.l**) are in the bottom of the figure while the native like (**Nat.l**) are in the top.

MFE, the simulation is directly dominated by the MFE after t=10^1.

To further investigate the fast-folding kinetics, we tested a classic bi-stable example, the sequence `GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC`. Because of its length, we were able to sample the whole space of sub-optimal structures from the unfolded state to the MFE structure. The ensemble is composed of 20×10^3 structures. The barrier tree derived from this ensemble displayed the bi-stable system. The two relevant structures are denoted SA and SB (figure 6). If initialized in the lower part of the tree, the kinetic produced is similar to the one obtained with RAFFT. First, the SB structure dominated the population from t=10^2 to t=10^{10}, then the SA structure took over the population. Otherwise, if the kinetic is initialized in the other basin, only the MFE dominated the population.

DISCUSSION

We have proposed a heuristic of the RNA structure and dynamic predictions called RAFFT. This heuristic uses simple folding rules based on the stem rate model. First, it searches for groups of consecutive base pairs, stems, and form them if they improve the stability. We implemented an FFT-based technique that uses a mirror encoding to quickly identify stems. Once a stem is formed, the sequence is split into two independent parts on which the procedure is recursively called. To mimic the parallel folding paths naturally observed, we implemented a stacking procedure where multiple parallel folding trajectories can be stored.

To assess the relevance of the folding trajectories produced, we compared the algorithm performance for the folding task. Two structure estimates were compared with: the MFE structure computed using `RNAfold`, the ML estimate using `MxFold2`. Other thermodynamic-based and ML-based tools were investigated but not shown here. We chose the MFE since it provides an intuitive interpretation in the structure landscape. The ML estimates gives a data view of the structure spaces. Although, it has better prediction accuracy, it also include signal from other folding mechanism such as the effects of chaperon proteins. Therefore, the ML estimate may include effects that are induced by the environment of those RNAs.

From our experiments, RAFFT had an overall performance below the MFE predictions by 8.1% of PPV and 10.3% of sensitivity. The ML-based approach dominated the predictions (70.4% of PPV and 77.1% of sensitivity). We observed some drastic loss of accuracies when the known structures contained large unpaired regions. However, those sequences were anecdotal in the dataset. Moreover, those regions are unlikely to be stable and assumed to be very flexible. Nevertheless, the effect of unpaired regions seemed less dramatic for the ML method since it can produce some of those atypical structures. We found no striking evidences of the length effect on prediction quality. In addition, no empirical effects of the base spanning was observed (see supp. mat.) as already pointed out in [2].

The PCA performed on the known structure compositions revealed a structure space prone to elongated structures where large unpaired hairpin loops and exterior loops can be observed. The PCA analysis performed on the structures predicted by the thermodynamic-based methods (RAFFT and MFE) shown similar structure spaces, where unpaired regions are of limited number. On the other hand, the ML method seemed to be closer to the natural structure space. According to the thermodynamic model, those unpaired regions have local stability equal to zero. Hence, those regions are not stable at regular experimental conditions in the sense that they may not have a unique stable structure. However, the ML-method was able to identify such structure more consistently than thermodynamic methods. This PCA revealed a

group of structures with high percents of hairpins. This may suggest some overfitting effects.

Although the overall performance of RAFFT was only fair in the folding task, we found one among the $k = 50$ predicted trajectories that had better accuracy than the low energy structure displayed. In fact, the gain of performance is substantial for the sequences of length below 200 nucleotides with 16% gain in PPV compared to the MFE predictions. The performance is similar to the ML-base method for this length range. Sequences of length < 200 nucleotides represent 86.4% (1983 sequences) of the total dataset. This shows that some of folding scenarios in all parallel folding path stored by RAFFT are relevant.

For the 140 sequences of length greater than 300 nucleotides, all 50 predictions per sequence performed worst than the other methods. This could be partially explained by the greediness of the algorithm, however, we also believe that the energy model assumption such as the additivity could give a complementary explanation [31]. However, the MFE did not show any notable discrepancy for large sequences (> 300 nucleotides) except for a few structures with large unpaired regions. This could be explained by the observation used in `kinwalker` where locally optimal substructures composed the native structures. Therefore, the MFE structures may be more often composed of locally optimal structures. We tried RAFFT with a larger number of saved structures in the stack, however, it only got closer to the MFE prediction quality and did not perform better on large sequences (see supp. mat.).

As an illustrative example, we applied our method on a natural RNA and a classic toy system. The folding trajectories produced by RAFFT started from the completely unfolded state and depicted simple "two-states" folding mechanisms. We showed that these fast-folding graphs obtained from RAFFT can be used to derived kinetic ansatz. Usual kinetic frameworks ([19]) involve the sampling of many thousands of structures then a coarse-grain procedure into basins. Then, an arbitrary choice of initialization is needed to simulate the structure competition. Here, since all folding graphs started from the completely unfolded structure, the proposed kinetic model can describe the complete folding process. When compared to the fast folding kinetics, we obtained similar results for some chosen initial conditions.

Given the results, we believe that RAFFT is a robust heuristic for the structure predictions since it can predict structure of high accuracy for 86.4% of this dataset. The folding paths as calculated by RAFFT are intuitive and coarse-grained since whole stems are formed sequentially. The folding landscape depicted by the proposed method also get along with the traditional two-states protein folding model. Furthermore, the proposed kinetic ansatz derived from the fast-folding graph was shown to approximate well the usual kinetic framework but with a gain in the number of structure needed. However, additional efforts are necessary to determine whether the folding paths followed were experimentally observed and when this folding model is likely to fail.

On the technical points, the mirror encoding as describe here is a versatile tool for RNA analysis. Since it contains the relative positions of base pairs in the whole sequence, we expect it to be extendable to other use cases such as sequence clustering, or the speed up of Nussinov-like algorithms. On the other hand, we are aware of the limits of choosing the largest stem at each step and will exploring different schemes. However, the greediness of the algorithm had a limited impact on the results. Another limit of the method is the choice of the relevant path among all parallel path predicted for one sequence. To alleviate this, we propose to compute ML-based scores to determined which structure is likely to be observed among the N saved structures by RAFFT.

DATA AVAILABILITY

One implementation in `python3.0` of RAFFT and the benchmarck data used in this manuscript are available from https://github.com/strevol-mpi-mis/RAFFT. We also provide the scripts used for the figures and kinetic analyses.

FUNDING

ACKNOWLEDGEMENTS

References

[1] Ambros, V. The functions of animal microrna. *Nature 431*, 7006 (2004), 350–355.

[2] Amman, F., Bernhart, S. H., Doose, G., Hofacker, I. L., Qin, J., Stadler, P. F., and Will, S. *The Trouble with Long-Range Base Pairs in RNA Folding*. Advances in Bioinformatics and Computational Biology. Springer International Publishing, 2013, pp. 1–11.

[3] Baranov, P. V., Henderson, C. M., Anderson, C. B., Gesteland, R. F., Atkins, J. F., and Howard, M. T. Programmed ribosomal frameshifting in decoding the sars-cov genome. *Virology 332*, 2 (2005), 498–510.

[4] Benson, D. C. Fourier methods for biosequence analysis. *Nucleic Acids Research 18*, 21 (1990), 6305–6310.

[5] Darty, K., Denise, A., and Ponty, Y. Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics 25*, 15 (2009), 1974–1975.

[6] Dill, K. A. Additivity principles in biochemistry. *Journal of Biological Chemistry 272*, 2 (1997), 701–704.

[7] Domin, G., Findeiss, S., Wachsmuth, M., Will, S., Stadler, P. F., and Mörl, M. Applicability of a computational design approach for synthetic riboswitches. *Nucleic Acids Research nil*, nil (2016), gkw1267.

[8] Felsenstein, J., Sawyer, S., and Kochin, R. An efficient method for matching nucleic acid sequences. *Nucleic Acids Research 10*, 1 (1982), 133–139.

[9] FLAMM, C., FONTANA, W., HOFACKER, I. L., and SCHUSTER, P. Rna folding at elementary step resolution. *RNA 6*, 3 (2000), 325–338.

[10] Flamm, C., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie 216*, 2 (2002), nil.

[11] Geis, M., Flamm, C., Wolfinger, M. T., Tanzer, A., Hofacker, I. L., Middendorf, M., Mandl, C., Stadler, P. F., and Thurner, C. Folding kinetics of large rnas. *Journal of molecular biology 379*, 1 (2008), 160–173.

[12] Guo, Z., and Thirumalai, D. Kinetics of protein folding: Nucleation mechanism, time scales, and pathways. *Biopolymers 36*, 1 (1995), 83–102.

[13] Han, K., Liang, Z., and Zhou, N. Design strategies for aptamer-based biosensors. *Sensors 10*, 5 (2010), 4541–4557.

[14] Hofacker, I. L. Vienna rna secondary structure server. *Nucleic Acids Research 31*, 13 (2003), 3429–3431.

[15] Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., and Mathews, D. H. Linearfold: Linear-time approximate rna folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics 35*, 14 (2019), i295–i304.

[16] Katoh, K. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research 30*, 14 (2002), 3059–3066.

[17] LEONTIS, N. B., and WESTHOF, E. Geometric nomenclature and classification of rna base pairs. *RNA 7*, 4 (2001), 499–512.

[18] Lorenz, R., Bernhart, S. H., zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. Viennarna package 2.0. *Algorithms for Molecular Biology 6*, 1 (2011), 26.

[19] Lorenz, R., Flamm, C., Hofacker, I., and Stadler, P. Efficient computation of base-pairing probabilities in multi-strand rna folding. In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies* (- 2020), p. nil.

[20] Martinez, H. M. An rna folding rule. *Nucleic Acids Research 12*, 1Part1 (1984), 323–334.

[21] Mathews, D. H. How to benchmark rna secondary structure prediction accuracy. *Methods 162-163*, nil (2019), 60–67.

[22] Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure. *Proceedings of the National Academy of Sciences 101*, 19 (2004), 7287–7292.

[23] Pan, J., Thirumalai, D., and Woodson, S. A. Folding of rna involves parallel pathways. *Journal of Molecular Biology 273*, 1 (1997), 7–13.

[24] Ramakrishnan, V. Ribosome structure and the mechanism of translation. *Cell 108*, 4 (2002), 557–572.

[25] Reuter, J. S., and Mathews, D. H. Rnastructure: Software for rna secondary structure prediction and analysis. *BMC Bioinformatics 11*, 1 (2010), 129.

[26] Rivas, E., Lang, R., and Eddy, S. R. A range of complex probabilistic models for rna secondary structure prediction that includes the nearest-neighbor model and more. *RNA 18*, 2 (2011), 193–212.

[27] Russell, R., Zhuang, X., Babcock, H. P., Millett, I. S., Doniach, S., Chu, S., and Herschlag, D. Exploring the folding landscape of a structured rna. *Proceedings of the National Academy of Sciences 99*, 1 (2001), 155–160.

[28] Sato, K., Akiyama, M., and Sakakibara, Y. Rna secondary structure prediction using deep learning with thermodynamic integration, 2020.

[29] Singh, J., Hanson, J., Paliwal, K., and Zhou, Y. Rna secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications 10*, 1 (2019), 5407.

[30] Solomatin, S. V., Greenfeld, M., Chu, S., and Herschlag, D. Multiple native states reveal persistent ruggedness of an rna folding landscape. *Nature 463*, 7281 (2010), 681–684.

[31] Tinoco, I., and Bustamante, C. How rna folds. *Journal of Molecular Biology 293*, 2 (1999), 271–281.

[32] Turner, D. H., and Mathews, D. H. Nndb: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research 38*, $\text{suppl}_1(2009), D280 - -D282$.

[33] Vitreschak, A. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends in Genetics 20*, 1 (2004), 44–50.

[34] Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., and Stadler, P. F. Efficient computation of rna folding dynamics. *Journal of Physics A: Mathematical and General 37*, 17 (2004), 4731–4741.

[35] Zhang, W., and Chen, S.-J. Rna hairpin-folding kinetics. *Proceedings of the National Academy of Sciences 99*, 4 (2002), 1931–1936.

[36] Zhang, W., and Chen, S.-J. Exploring the complex folding kinetics of rna hairpins: I. general folding kinetics analysis. *Biophysical Journal 90*, 3 (2006), 765–777.

[37] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research 31*, 13 (2003), 3406–3415.

[38] Zuker, M., and Stiegler, P. Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Research 9*, 1 (1981), 133–148.

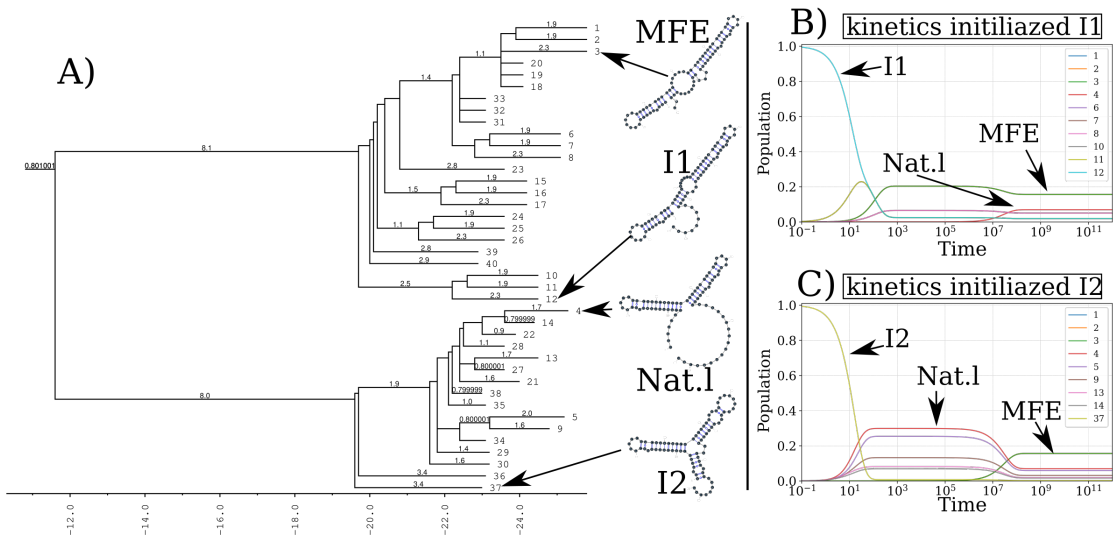

Figure 5: (A) Barrier tree obtained for the Coronavirus frameshifting element. (B) Kinetics where the initial population started with structure I1. (C) Kinetics where the initial population started with structure I2.

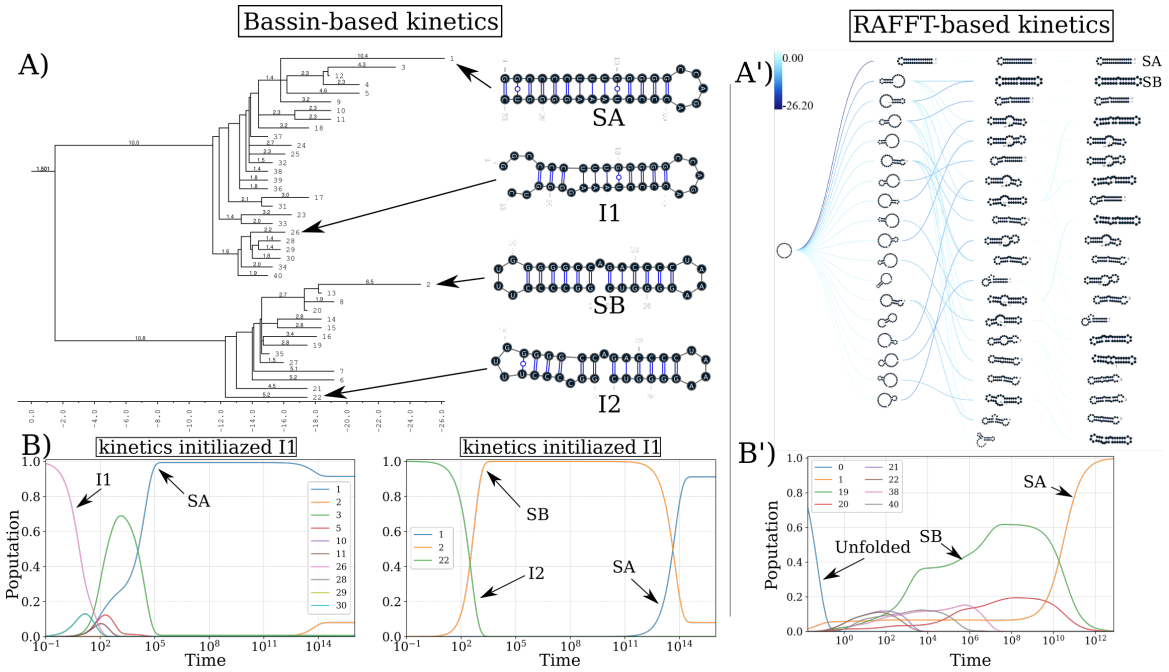

Figure 6: (A) Barrier tree for the sequence `GGCCCCUUUGGGGGCCAGACCCCUAAAGGGGUC`. (B) Left side is the kinetic trajectory when initialized with structure 22. (B) Right size is the kinetics trajectory when initialized with structure 26. A') The fast folding paths with 20 saved structures and 100 stem searched. B') Fast-folding kinetic trajectory obtained from the bi-stable system (indices are different from the barrier tree indices).