A Mirror encoding and FFT for an RNA fast-folding path heuristic

Vaitea Opuu, Nono S. C. Merleau, and Matteo Smerlak

Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany

March 30, 2021

We propose a heuristic to the folding dynamic making use of a mirror encoding and the fast Fourier transform (FFT) called RAFFT. Based on simple folding rules, it can create many parallel folding paths. The performance in the folding task when compared to the state-of-the-art folding tools was reasonable on a well-curated dataset. However, when all parallel folding paths were analyzed, it revealed near-native predictions (PVV $\approx$ 79% and sensitivity $\approx$ 81%) for sequences of length below 200 nucleotides. On average, those predictions were found to be of similar quality to recent deep-learning-based methods. The dynamics were built with the stem-based rate model which displays coarse-grained folding paths. Moreover, stems are sequentially added to the predicted structures if it improves their overall stability. Those simple rules create smooth coarse-grained folding paths which are intuitive to analyze and get along the traditional two states view of the protein folding landscape and the fast-folding path concept. Hence, those paths could well approximate fast-folding paths. Since the algorithm was designed toward speed, it can be readily applied to large RNAs.

1 Introduction

Natural RNA molecules as proteins have biologically relevant functions in many cellular contexts such as protein translation (mRNA, tRNA...), but also in protein-like functions where RNA can perform enzyme functions. Generally, those functions can be better understood through the light of their static tri-dimensional structure. However, with the discovery of riboswitch RNA molecules (**ref**), the dynamic behavior of RNA is getting more interest. Several other RNA systems were found to allow multiple metastable structures allowing for highly tunable functions (**ref from kinwalker**).

RNA molecules are bio-polymers composed of nucleotides. These nucleotides are simple molecules composed of a phosphate-based backbone, a ribose, and a variable nucleobase. Four different canonical bases/nucleobases compose the RNA, namely adenine (A), guanine (G), uracil (U), and cytosine (C). As amino-acid sequences, these nucleotide sequences can form complex tri-dimensional structures critical for their biological functions. Four main structure types are usually considered: the primary structure which is the nucleotide sequence itself.

The secondary structure is usually defined by interacting pairs of nucleobases called canonical base pairs interactions, such as G-C, A-U, and G-U. Next, the tertiary interactions involve other non-trivial interactions within the same sequence. Finally, the quarternary structure is defined by the interaction of multiple structured RNA molecules. However, unlike proteins, these structures are usually hierarchically formed. The secondary structure is generally formed first, then the tertiary structure (**ref tinoco how RNA fold**). Moreover, unlike proteins, the secondary structure provides an accurate enough description of the thermodynamics and kinetics of RNA molecules. Although base pairing can be formed with various configurations (**ref, westhof**). It is possible to consider a more coarse-grained interaction model where the geometries are not explicitly considered. Many subtleties can be used to define the secondary structure, but we used here a well-accepted formal definition called pseudoknot-free. This forces the RNA secondary structure to be drawable onto a plan where base pairs cannot cross (**ref**). This has computational and theoretical benefits as will be described later.

The structure space of an RNA molecule is described by the stability of individual possible structures. The stability ΔG_s of a structure s is measured by the free energy changes with the completely unfolded state. To predict secondary structures, thermodynamic-based methods use empirical data to estimate RNA stability. By assuming the additivity of the loop contributions to the overall stability, the NNM energy model was developed (**ref**). It associates free energy values, determined by optical melting experiments, to individual loop types and compositions (**turner**). The functional form allows for generalizable energy parameters and the use of an efficient dynamic programming algorithm. It can determine the minimum free energy (MFE) structure of a sequence. The MFE is considered a gold standard for free-energy-based predictions. Other estimates exist such as the MEA, however, it was not found to be significantly better than the MFE. Also, the MFE has an intuitive interpretation. Several tools implement this algorithm, namely Zucker algorithm (**ref**), such as RNAfold, Mfold, seqfold (**ref**). Although those methods were found to be consistently good at predicting RNA secondary structures (**ref bench**), the additivity foundation is expected to be doomed when sequences get larger and structures complexify. The non-additivity of tertiary interactions and pseudoknots pairings can partially explain this discrepancy. Therefore, pseudoknots loop contributions are not defined in the main parameters sets like the Turner2004 model. In addition, the thermodynamic energy models are not perfect. Another limit of such prediction estimates is the static picture that it gives to the RNA folding landscape. From a dynamic standpoint, the sequence navigates the structure space by following the landscape drawn by the stability.

Kinetic information on RNA molecules was found to give valuable complementary information (**riboswitch, regulatory**). To follow the dynamic of individual RNA molecules, three rate models describing elementary steps in the structure space are currently used. The base stack model uses base stacks formation and breaking as elementary moves (**ref vfold**). The base pair rate model uses base pairs as elementary steps as implemented in kinfold (**ref kinfold**). kinfold implemented a continuous-time Monte Carlo simulation to follow the RNA folding. It gives the finest resolution in the secondary structure folding landscape but the cost of computation time. The stem-based model uses the creation or deletion of stems to construct the folding dynamics. It is among the first strategy explored (**ref**) and provides a coarse-grained description of the kinetic. 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. An alternative approach, implemented in kin-

walker (**ref**), used the observation that folding paths are generally composed of locally optimal conformations. Therefore, locally optimal structures are formed using the standard dynamic programming algorithm and aggregated together along with the folding dynamic.

Here we propose a complementary approach by approximating fast-folding paths based on simple folding rules. The basic idea is to use the stem rate model to create multiple parallel folding paths. Here, stems are not allowed to be removed and can be formed only if it improves the stability. It uses a mirror encoding and relies on the fast Fourier transform to speed up the search of stems. This method is inspired by MAFFT, a well-known multiple-sequence-alignment tool. Another FFT-based algorithm was already developed, FFTbor2D, but uses the FFT differently (**ref**). The mirror encoding is a simple numerical orthogonal representation of nucleotide sequences. Several other similar encodings were developed for the analysis of DNA (**ref tally vectors, first FFT**) and combined with the FFT. To assess the reliability of the paths predicted, we compared its performance on the folding task for a well-curated dataset, archive II (**ref**). The algorithm is compared to two estimates: the MFE computed by RNAfold and an ML-based prediction computed by MxFold2. Next, we applied the algorithm to a simple test case, the Coronavirus frameshifting stimulation element (**ref**), where it performed better than the MFE.

2 FFT based folding dynamic heuristic

We now describe the heuristic starting from one sequence S and its associated unfolded structure of length L. We first create a numerical representation of S where each type of nucleotide is replaced by a unit vector of 4 components:

$$A \to \big(1000\big)\, U \to \big(0001\big)\, C \to \big(0100\big)\, G \to \big(0010\big) \tag{1}$$

which gives us a $4 \times L$ matrix we call X where each row is a nucleotide type channel. Here, the first row would be the A channel which we refer to as X^A. Then, we create a second copy for which we revert the order of the sequence and use the following complementary encoding:

$$\bar{A} \to \big(000 w_{\text{AU}}\big)\, \bar{U} \to \big(w_{\text{AU}} w_{\text{GU}} 00\big)\, \bar{C} \to \big(00 w_{\text{GC}} 0\big)\, \bar{G} \to \big(0 w_{\text{GC}} 0 w_{\text{GU}}\big) \tag{2}$$

Where $\bar{A}$ (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 for the next step. We call this new complementary copy $\bar{X}$, the mirror of X.

Next, for each of the 4 channels, we compute the correlation between X and $\bar{X}$ and by simply summing up the channel correlations, we obtain the correlation between the two copies:

$$cor(k) = (c_{X^A,\bar{X}^A}(k) + c_{X^U,\bar{X}^U}(k) + c_{X^G,\bar{X}^G}(k) + c_{X^C,\bar{X}^C}(k))/min(k, 2 \times L - k) \tag{3}$$

where $c_{X^A,\bar{X}^A(k)}$ is the correlation in the A channel between the two copies. $cor(k)$ gives the average number of base pairs for a positional lag k. One channel correlation between copies is given by:

$$c_{X^A,\bar{X}^A}(k) = \sum_{1 \le i \le L, 1 \le i+k \le M} X^A(i) \times \bar{X}^A(i+k) \tag{4}$$

where $X^A(i)$ and $\bar{X}^A(i+k)$ are the A channel of site i and $i+k$. $X^A(i) \times \bar{X}^A(i+k)$ 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. Although this requires $O(N^2)$ operations, it can take advantage of the FFT which reduces drastically its complexity to $O(Nlog(N))$.

The large correlation values between the two copies indicate the positional lag at which the base pair density is high. Therefore, we use a sliding window strategy to search for the longest consecutive base pairs within the positional lag. Since the copies are symmetrical, we only need to slide over one-half of the positional lag. Once the longest base pairs are identified, we simply compute the free energy change when those base pairs are formed. We perform the same search for the n highest correlation lags, which gives us n potential stems. Then, we add to the current structure the base pairs that give the best change of free energy. Free energies were computed using Turner 2004 energy parameters through Vienna RNA package API (**ref**).

We are now left with two segments, the interior, and exterior of the group of consecutive base pairs formed. The two exterior fragments are concatenated together. Then, we simply apply recursively the same procedure on the two segments separately in a "Breadth First" fashion to form new consecutive base pairs, until no base pair formation can improve the energy. Hence, it is straightforward to consider pseudoknots by simply concatenating all the fragments left. When multiple stems can be formed in these independent fragments, we combine those possible independent stems and pick the composition that has the best overall stability.

The algorithm described so far tends to be stuck in the first local minima found along the folding trajectory. To alleviate this, we propose a stacking procedure where the best trajectories are stored in a stack and evolved in parallel. Hence, it offers the flexibility of overcoming some energy barriers. **Figure** shows the whole procedure. Once no stem can be formed, the algorithm stops and output the structure with the best energy found among the structures saved in the stack.

3 Application to the folding task

To evaluate the relevance of the folding dynamic heuristic, we compared its performance for the folding task. Also, to assess the effect of sequence lengthens on these predictions, we analyzed their performance length-wise. To localize its performance, we compared with two estimates: the MFE computed by RNAfold (**ref**) and the ML-based prediction computed by MxFold2 (**ref**). RAFFT predictions were performed using non-optimized weights. 50 structures are evolved in parallel for each sequence and 100 positional lags were explored at each step for each store structure.

Figure 1 shows the performance in predicted positive values (PPV) and sensitivity for the four methods. It shows that the ML method is consistently better than thermodynamic-based methods. 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 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 produced by RAFFT, we found on average at least one prediction with 59% of PPV and 63% of sensitivity as shown figure 1. The overall gain of performances is not significantly different from the MFE predictions. However, for the sequences of length lesser than 200 nucleotides, this gain was found to be substantial and significant (≈ 16 % better than the MFE) with PVV $\approx 79\%$ and sensitivity $\approx 81\%$. The accuracy for those sequences is equivalent to ML performances. For sequence lengths greater than 300 nucleotides, we observed

the same drastic loss of accuracy, although we took only the best prediction among the 50 saved configurations for each sequence.

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 exterior loops such as displayed in figure 2. The second region is around 200 nucleotides in length. The known structure of these sequences also displayed large unpaired regions such as the one shown in 2.

Figure 1: **Predicted positive values and sensitivity results.** RAFFT (blue) displayed the best energy found. RAFFT* displayed the best score found among 50 saved structures.

To investigate the region of the structure space where the thermodynamic model tends to fail, we computed the composition content of the known structures. Loop type length was computed in percents. Figure 3 shows principal component analysis (PCA) of the structural composition. From the PCA, we observed that the known structures are distributed in the

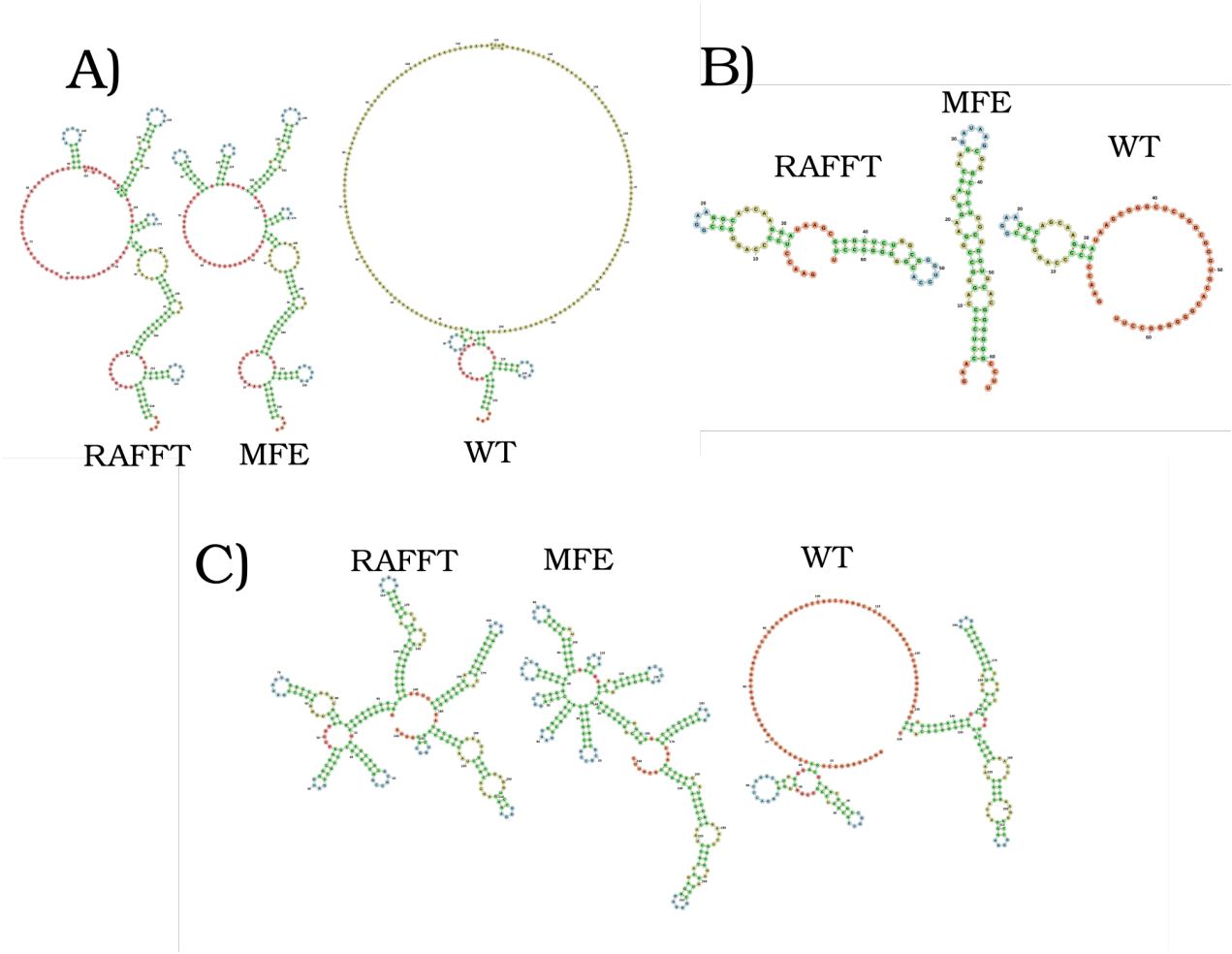

Figure 2: **Structures found to be difficult to predict with thermodynamic models.** WT is the known structure.

structure space toward interior loops. Also, some natural structures, as observed above, have large exterior loops. The center of mass in the principal component space is located in between the high-density stacking and interior loops. This shows that the dataset contains many elongated structures.

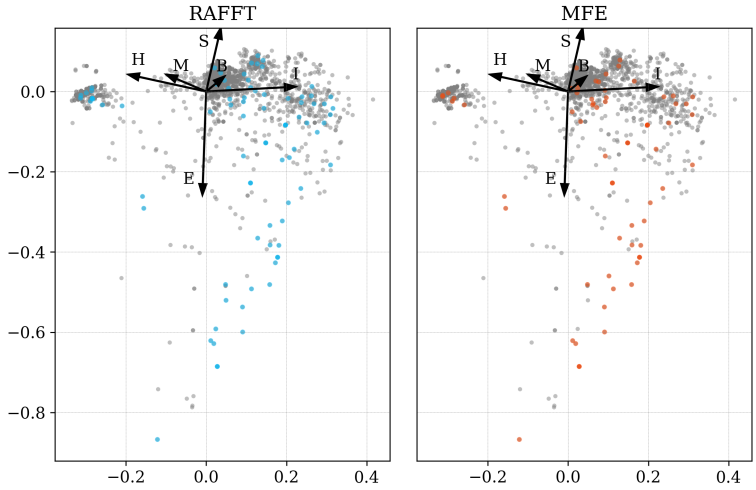

Figure 3: **PCA analysis of the known structure space where prediction with PPV = 0 were colored for RAFFT and the MFE prediction.**

Next, we investigated the structure space produced by the three methods. The thermodynamic model tends to produce more diverse structure spaces as shown in figure 4. Loops content was extracted from the predicted structures of each method and projected onto their respective two first principal components space. Both RAFFT and MFE predictions seem to produce similar structure spaces while the ML method does allow for long unpaired regions in long hairpins which tend to be closer to the dataset structure space.

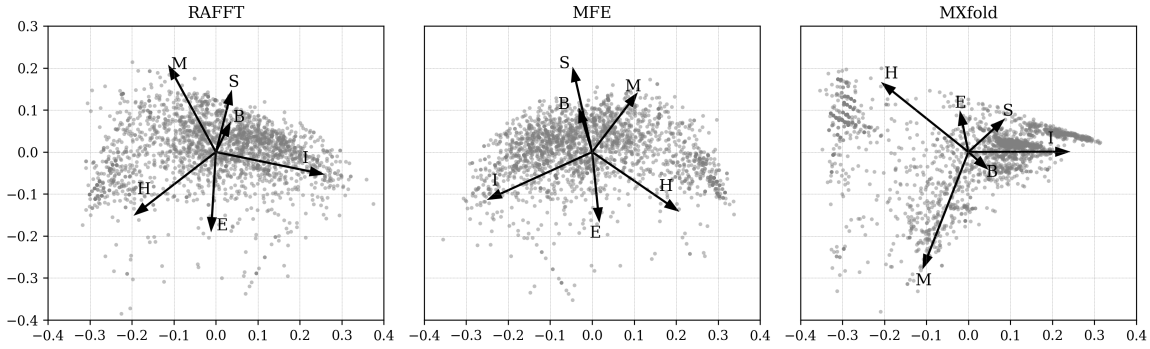

Figure 4: **PCA analysis of the predicted structures for the three methods.**

4 Test case to predict fast-folding paths

Finally, to illustrate RAFFT folding heuristic, we applied it to the Coronavirus frameshifting stimulation element. It is an RNA sequence of about 82 nucleotides with a secondary structure determined by sequence analysis and obtained from RFAM database. The assumed known structure has a pseudoknot but was not displayed here. Figure 5 shows the folding path predicted, the MFE prediction, and the assumed known structure. The approximated fast-folding path is predicted in three steps where 5 structures were store in the stack and 10 positional lags were searched for stems. As shown, some structures explored were not saved or evolved since no further improvement (relative to all possibilities) was found. RAFFT was able to recover near-native structures, found to be closer than the MFE, and depicted simple folding paths.

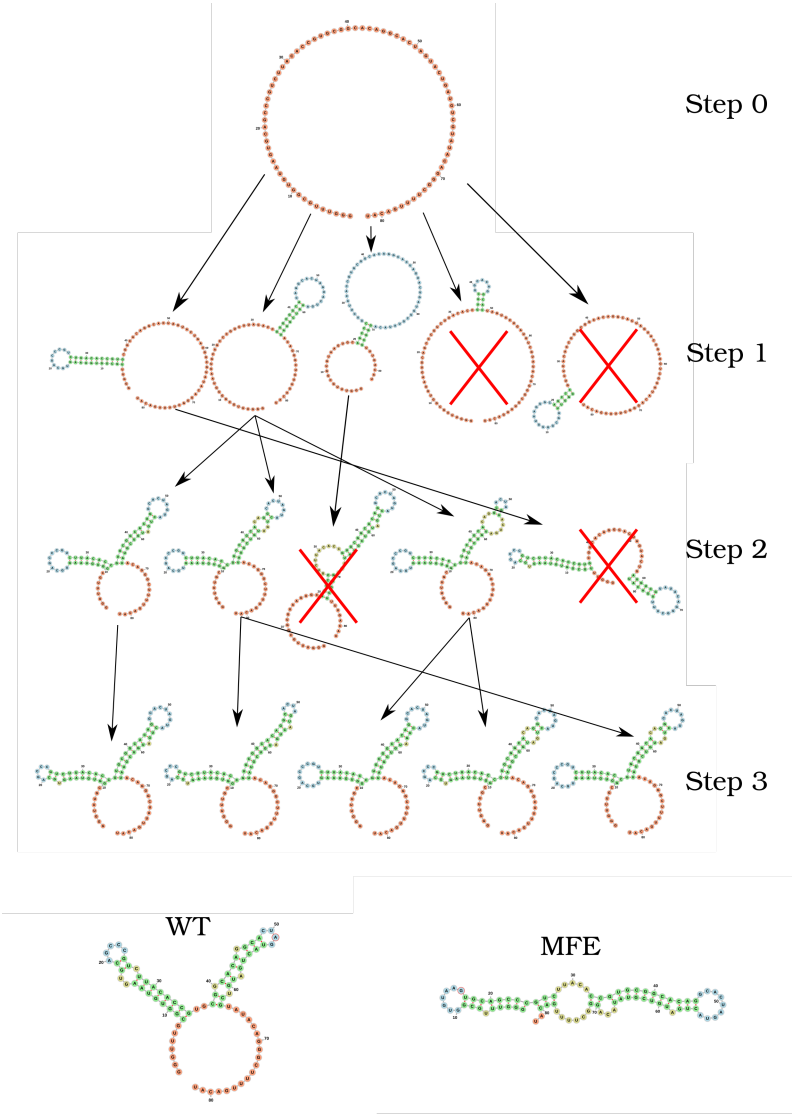

Figure 5: **Fast folding path prediction for the Coronavirus frameshifting stimulation element.**

5 Concluding discussion

We have proposed a simple heuristic of the RNA folding dynamic called RAFFT for fast-folding paths. This heuristic uses a greedy rule. Groups of consecutive base pairs, stems, found to improve the energy are formed along with the procedure. Hence, it produces smooth and coarse-grained trajectories. To search for consecutive base pairs, we implemented an FFT-based technique that uses a mirror encoding. Once a stem is formed, the sequence is split into two independent segments on which one can recursively search for new stems. For one sequence, the algorithm can follow k folding paths.

To assess the relevance of the folding trajectories produced, we compared the algorithm performance for the folding task. We considered three methods to compare with: the MFE structure computed using RNAfold, the ML-based 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, and the MEA prediction was not found to be significantly more accurate (**ref how bench**).

From our experiments, RAFFT had an overall performance below the MFE predictions by 8.1% of PVV 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, these regions are unlikely to be stable and assumed to be very flexible. However, the effect of unpaired regions seemed less dramatic for the ML method since it can produce some of those atypical structures. Also, we found no striking evidence of the length effect on prediction quality. Moreover, no empirical effect of the base spanning was observed (see supp mat) as already pointed out in (**ref book bp spanning**).

The PCA performed on the known structure compositions revealed a structure space prone to elongated structures where large unpaired hairpins 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 space, where flexible loops such as long hairpins or exterior loops 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 (**ref turner**) in the sense that they don't have a unique stable structure. However, the ML-method was able to identify such structure more consistently than thermodynamic methods. The PCA revealed a group structure with high percents of hairpins. This may suggest some overfitting effects. We argue that not being able to recover such structures is proof of robustness.

Although the overall performance of RAFFT was weak compared to the state of the art in the folding task, we found one among the $k = 50$ predicted trajectories that had better accuracy than the low energy trajectory displayed. In fact, the gain of performance is substantial for the sequences of length lesser than 200 nucleotides with about 16% better in PPV than the MFE predictions. The performance is significantly similar to the ML-base method for that length range. Sequences of length ¡ 200 nucleotides represent 86.4% of the total dataset. However, for the 140 sequences of length greater than 300 nucleotides, all k predictions per sequence were similar and performed worst than the other methods. This could be partially explained by the greediness of the algorithm, however, we also believe that the thermodynamic could be a complementary explanation. Indeed, the additivity of the loop contributions to the stability

assumed is likely to be doomed for large sequences (**ref tinoco**). However, the MFE did not show any notable discrepancy for large sequences (> 600 nucleotides). This could be explained by the observation used in kinwalker, where locally optimal substructures composed the native 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 (see supp mat) on large sequences.

Given the experiment results, we believe that RAFFT is a robust heuristic for the fast-folding path since it can produce predictions of high accuracy for 86.4% of this dataset. The folding paths as calculated by RAFFT are smooth and coarse-grained since whole stems are formed, if it improves the energy, and leads to near-native structures. This near-native coarse-grained folding path is an intuitive idea that gets along with the funnel protein folding landscape. We expect this heuristic to give valuable and complementary information to the MFE-like predictions. However, some additional work is necessary to determine whether the folding paths followed were experimentally observed.

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 maximal number of base pairs each at each step. However, the greediness of the algorithm had a limited impact on the results. We are not planning to provide yet another folding tool, in this already crowded area of excellent software, but one could combine this tool with an ML-base scoring to discriminate the folding path that is likely to be observed.

6 Methods

We formed two sub-datasets based on the ArchiveII (**ref**) dataset. First, we removed from all the structures containing pseudoknot since all tools considered here don't handle pseudoknots. Next, we removed all the structures which were evaluated with positive energy or null energy with the Turner 2004 energy parameters. Since positive energies mean that the completely unfolded structure is more stable than the native one, we assume that those structures are not well modeled by the energy function used here. 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.

To compute the MFE structure, we used RNAfold (version) with the default parameters and the Turner 2004 set of energy parameters. For the machine learning tool, we computed the prediction using Mxfold2 with the default parameters. The structures for both were used for the statistics.

For the FFT-based algorithm, we used two sets of parameters. First, we used a search for consecutive base pairs in the 50 best positional lags and stored 50 conformations for which we displayed the best energy found. The correlation was computed using the weights $w_{GC}=3$, $w_{AU}=2$, and $w_{GU}=1$.

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:

$$PPV = \frac{TP}{TP + FN} \quad \text{Sensitivity} = \frac{TP}{TP + FP} \tag{5}$$

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 in **ref Mathews**, which provide also a more flexible estimate where shifts are allowed.

The loop compositions were extracted in terms of proportion to have an overall measure of the structure distribution. 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. Hence, we convert those sizes into percents of types of loops from which we extracted the principal components. Next, the structure compositions were projected on the first two principal components for visual conveniences. The composition arrows represent the eigenvectors obtained from the diagonalization of the covariance matrix.

References