A Mirror encoding combined with the FFT for a fast heuristic of the RNA folding dynamics

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

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

March 28, 2021

1 Abstract

- Simple and fast heuristic for the folding path of RNAs.
- It is straightforward to model Pseudoknots
- It's performance is comparable to exact method on the RNA folding problem
- It follows a simple idea which naively corresponds to RNA folds mechanism (many BPs formed at once to compensate for the lost of entropy)
- Among the 50 predicted structures, in average, at least one has pvv ˜ 74% and sensitivity ˜ 76%.
- We propose a fast algorithm method based on the FFT to search for high density BP regions.
- There are smooth coarse-grain folding path which lead to near native structures.

2 Introduction

2.1 RNA folding introduction

bla bla dynamic of secondary structure relevant bla biological function.

- MFE and MEA not significantly different in term of performances (how to bench RNA)

2.2 RNA folding dynamics

1. Description of RNA structure
2. going up to the 2ndary structure only
3. Simple rules to compute a structure: multiple BPs compensate the lost of entropy during the folding process.

2.3 Energy model

1. issue with additivity principle in model. Might be worst when the sequence lengthens since more tertiary interactions interplay.

2.4 Existing methods

1. MC sampling: kinefold; atomic moves; MC-style simulation
2. Barrier trees from conformation landscape subopt tree: Sample from the boltzmann ensemble of structures
3. Vfold, simplified folding model

3 FFT based folding dynamic heuristic

We now describe the heuristic folding algorithm starting from one sequence S and its associated unfolded structure of lenght L. We first create a numerical representation of S where each type of nucleotide in 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:

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

where w_{AU}, w_{GC}, w_{GU} are tunable parameters for the next step. We call this new copy , the mirror of X.

For each of the 4 components, we compute the correlation between X and and simply sum up the four channels to 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. The correlation $cor(k)$ gives the average number of base pairs for a positional lag k. One channel correlation between the 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 $i+k$. $\mathrm{X^A(i)} \times {}^{\mathrm{A}}\mathrm{(i{+}k)}$ is non zero if sites i and $i+k$ can form a base pair, and will be the value of the chosen weight as described above. Although this operation requires $O(N^2)$ operation, it can take advantage of the FFT which reduces drastically its complexity to $O(Nlog(N))$.

The large correlation values between the two copies indicates the positional lag between 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 pair are formed. We perform the same search for the n highest correlation lags, which gives us n possible possibilities. Then, we added to the current structure the base pairs that gives the best change of energy.

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 "Breath First" fashion to form new consecutive base pairs, until no base pair formation can improve the energy. However, it is straightforward to consider pseudoknots by simply concatenating all the fragments left.

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 50 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.

4 Folding RNAs

To evaluate the relevance of the folding dynamic heuristic, we compared the algorithm performance for the folding task. In addition, to assess the effect of sequence lengthens on these predictions, we analyzed their performance length-wise.

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

Among the 50 configurations produced by RAFFT, we found in average at least one prediction with in average 59% of PPV and ¡SENS¿ of sensitivity (blue curve in **figure**). 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 ($\approx$ 16 % better than the MFE). The accuracy for those sequences is equivalent to ML performances. For sequence lengths greater than 300 nucleotides, we observed the same drastic lost of accuracy, although we took only the best prediction among the 50 saved configurations for each sequence.

Two regions of lack of performances were observed for all methods. A group of 28 sequences of length shorter than 80 nucleotides were evaluated with free of their known structures about 9.8 kcal/mol greater than the MFE structures. Some of them involve large exterior loop such as displayed in **figure**. The second region is around 200 nucleotides length. The known structure of these sequences also displayed large unpaired regions such as the one shown in **figure**.

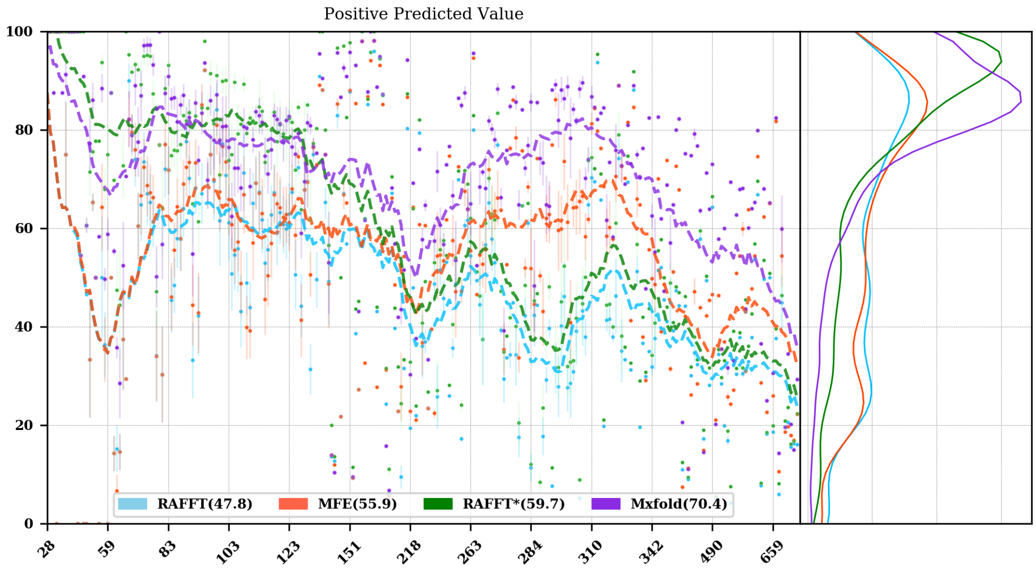

Figure 1: Folding comparison by taking the best energy among the 30 predicted trajectories

To investigate the region of the structure space where the thermodynamic model tends to fail, we computed the composition content of the known structures. **Figure** shows the prcent of base pairs or positions involved in the five loop types: interior, exterior, hairpin, stacking, and multi-branch loops. Those prcents were then represented in a principal component analysis.

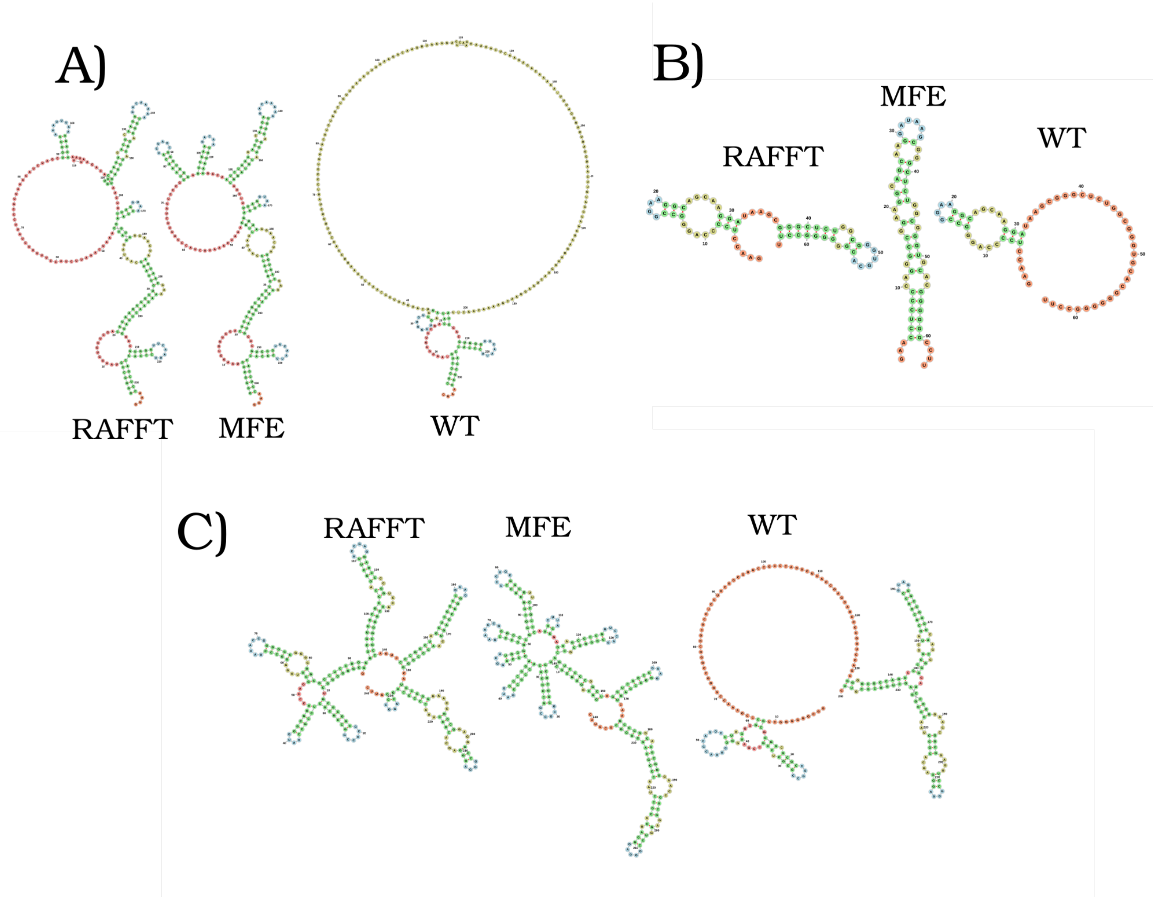

Figure 2: Difficult structures

From the PCA, we observed that the known structures are distributed in the structure space non-uniformly. 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.

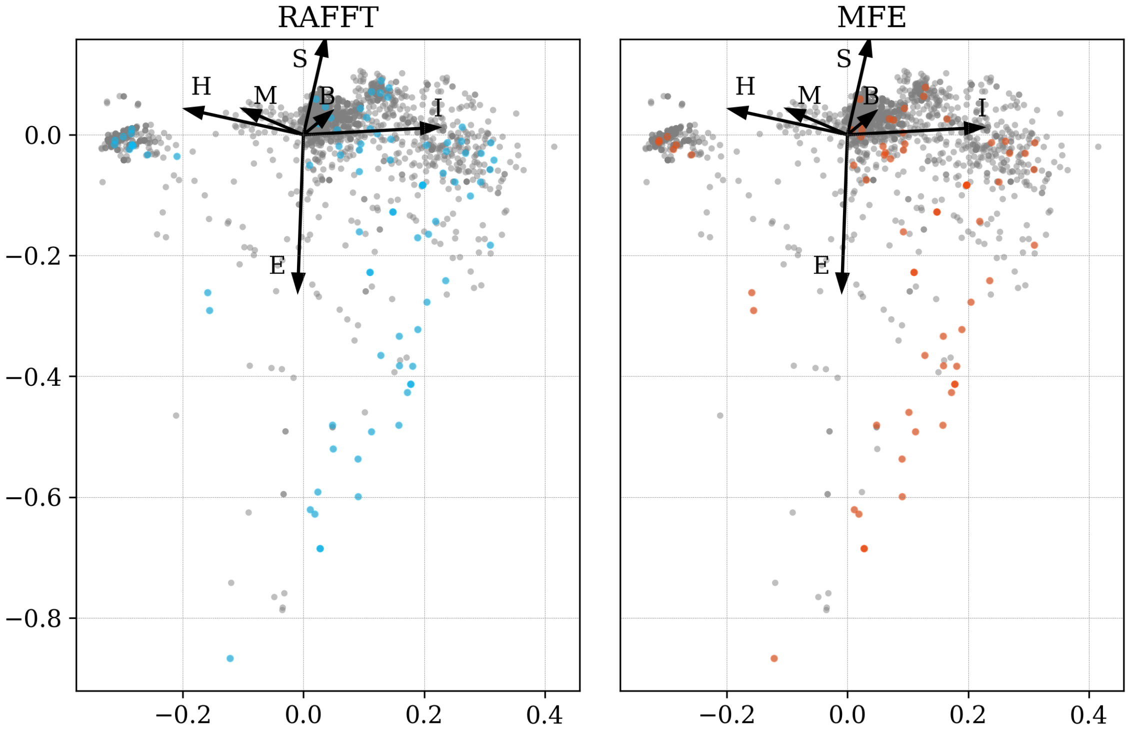

Figure 3: where does the methods failed? PCA RNAfold, Mxfold, FFT, and

The thermodynamic model tends to produce more diverse structures as shown in **figure**. Loops content were extracted from the predicted structures of each method and projected onto their respective two first principal components space. Both RAFFT and MFE predictions seems to produce a diverse structure space while the ML method does allow for long unpaired regions in long hairpins.

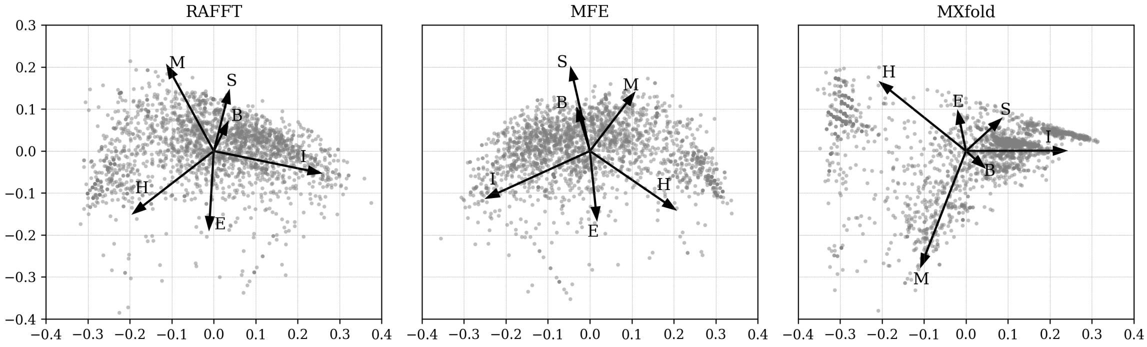

Figure 4: What kind of structure these methods naturally produced

5 Methods

We formed two sub-datasets based on the ArchiveII (**ref**) dataset. First, we removed from all the structure containing pseudoknot since all tool considered here don't handle pseudoknots. Next, we removed all the structures which were evaluated with a positive energy or null energy with the Turner 2004 energy parameters. Since positive energies means 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 kinfold, we performed for each sequence, 40 simulations of 10^4 (unit?). Then, we counted the occurrences of each structures and selected the 50 most populated structures. The best structure in terms of PPV was displayed and used for the statistics.

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

To measure the predictions accuracy, we used two metrics from epimiology. The positive predictive value (PPV) which 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 shift are allowed.

The loop composition were extracted in terms of proportion to have an overall measure of the structure distribution. We first convert all natural structures into Shapiro notation using Vienna Package utilies. From the notation, we extracted the proportion of base pairs involved into the interior, exterior, bulge, stacking, and multibranch loops. For each true structure, we obtained a prcent of type of loops from which we extracted the principal components. Next, the structure compositions where projected on the first two principal components for visual conveniences. The composition arrows represents the eigen vectors obtained from the diagonalization of the covariance matrix.

6 Concluding discussion

We have proposed a simple heuristic of the RNA folding dynamic called RAFFT. This heuristic uses a greedy rule to fold RNAs. Groups of consecutive base pairs found to improve the energy are formed along the procedure in such a way a smooth and coarse grained fashion. To search for consecutive base pairs, we implemented a FFT-based technique which takes advantage of the mirror encoding. Once a group of base pairs are formed, the sequence is split into two unrelated segments on which one can recursively search for new group of consecutive base pairs. For one sequence, the algorithm can follow k folding paths. Finally, the path which leads to the structure with the lowest energy is chosen.

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 estimate using MxFold tool and the kinetic approach using kinefold. Other thermodynamic-based and ML-based tools where investigated but not shown here. We chose the MFE since it provide a 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 $\approx$ 10% of PVV and ¡SENS¿ of sensitivity. The ML-based approach dominated the predictions (70.4% of PPV and 77.7% of sensitivity). We observed some drastic lost of accuracies when the known structures contained large of unpaired regions. These regions are unlikely to be stable and assumed to be very flexible regions which could explain their presence. However, the effect of unpaired regions seemed less dramatic for the ML method.

The principal component analysis performed on the known structure compositions revealed a structure spaces prone to elongated structures were 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 a local stability equal to zero. Hence, we suppose that those regions are actually not stable in the sens that they don't have a unique stable structure. However, the ML-method was able to identified such structure more consistently than thermodynamic methods. This may suggest some overfitting effects. We argue that not being able to recover such structures would be a proof of robustness.

Although the overall performance of RAFFT was weak compared to the state of the art in the folding task, we observed that among the $k = 50$ predicted trajectories, one was found to have a better accuracy than the low energy trajectory. 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 sequences were similar and performed worst than the other methods.

Given the experiment results, we believe that RAFFT is a robust heuristic for the folding dynamic 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 many base pairs, if it improves the energy, can be formed at once and can lead to near-native structures. This near native coarse grained folding path is an intuitive idea which get 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 are 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 to the speed up of nussinov-like algorithms. On the other hand, we are aware of the limits of chosing the maximal number of base pairs each at each step. The greedyness of the algorithm as shown in **figure**, however, it had a limited impact on the results. We are not planning to provide yet another folding tool, in this already crowded area of excellent softwares, but one could combined this tool with a ML-base scoring for such a purpose.