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

March 25, 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.

2 Introduction

2.1 RNA folding dynamics

bla bla dynamic of secondary structure relevant bla biological function.

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 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 FFT based folding dynamic heuristic

1. Encoding into two complementary strands
2. Search for high BPs regions
3. Use a sliding window to form large consecutive BPs
4. split the strands into interior and exterior
5. start again from 2) for the two sub-sequences

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:

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

where 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 $i+k$. $\text{X}^{\text{A}}(\text{i}) \times {}^{\text{A}}(\text{i+k})$ 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

1. comparisons to DP folding algorithm -¿ RNAfold and MFE prediction or MEA
2. Comparisons to ML folding algorithm -¿ Mxfold or Contextfold
3. The discrepancy between FFT and RNAfold for the folding task can be explained by the greedyness of the algorithm.
4. Show the best trajectory among the 50 predicted and its PVV performance =¿ means that one trajectory is relevant most of the case. Could be combine with ML method to determine which one.
5. How natural loop compositions are distributed -¿ bias toward some specific composition while.
6. Show two folding trajectories, one where it works, and one where the greedyness is a problem.

To evaluate the relevance of the folding trajectories produced, we benchmarked the algorithm performance for the folding task. In addition, we want to assess the effect of sequence length on these predictions.

We compared the performance of the MFE structure computed by RNAfold and the performance of a machine learning approach implemented in Contrafold/Mxfold.

The performance analysis were performed sequence-length-wise to asses the impact of length increase.

To investigate the region of the structure space where the thermodynamic model tends to fail, we computed the composition content of the natural structures and mapped the ones that had a low PVV score. It shows that the MFE tends to fail when the structure contains a high proportion of interior loops as shown in **figure**.

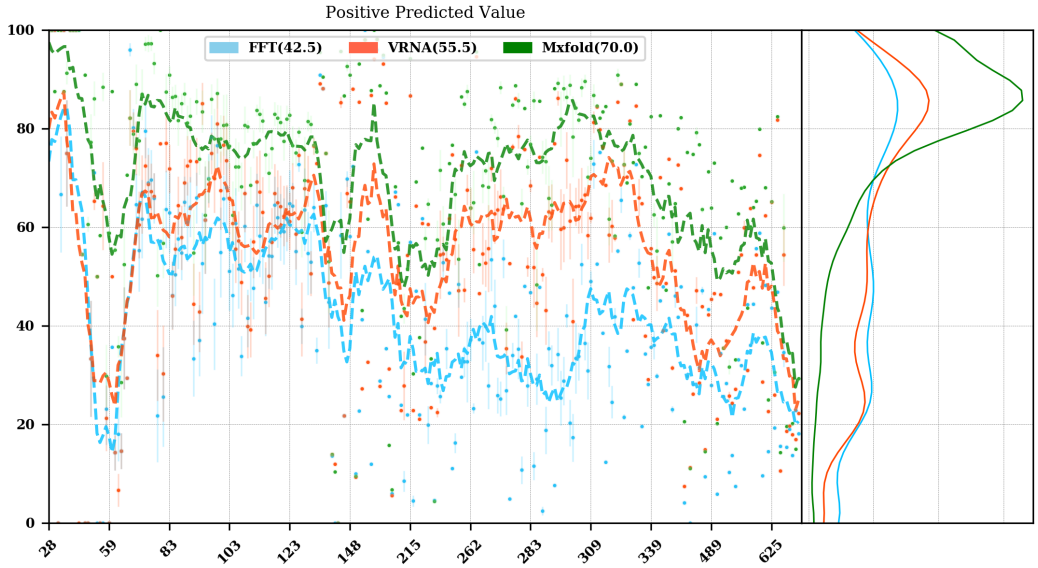

Figure 1: Folding comparison by taking the best energy among the 30 predicted trajectories

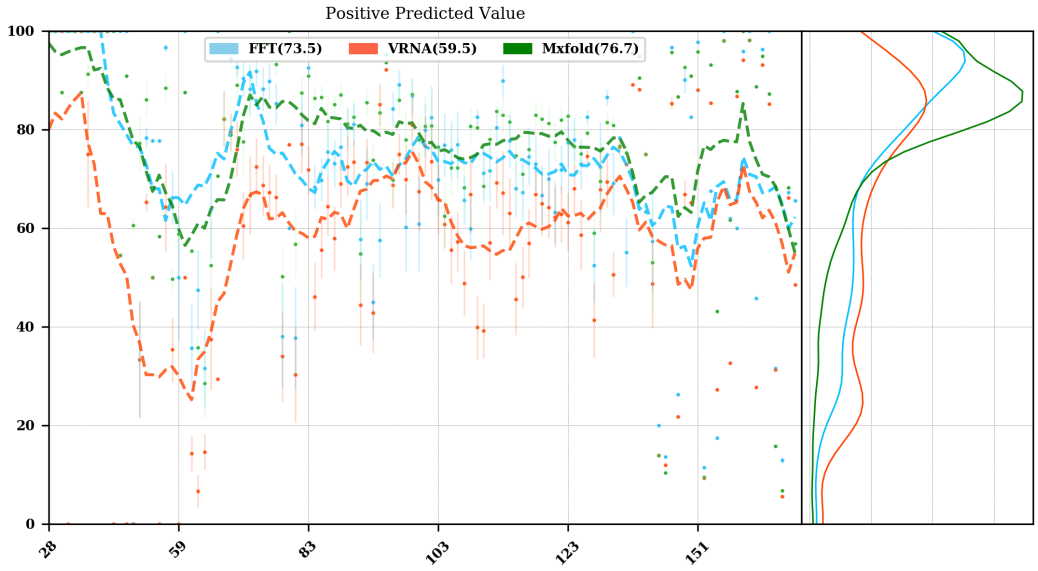

Figure 2: Is there a good trajectory among 50 saved trajectories

5 Methods

1. Dataset used
   (a) We considered all structures with nrj ¡ 0 and no pseudoknot (since the energy parameters doesn't take them into account).
   (b) We studied a smaller subsets of shorter sequences length ¡= 200 nuc in which we expect the thermodynamic model to be the most accurate. (maybe put that above)
2. Folding parameter applied for all methods considered
3. Analysis: PVV and sensitivity + PCA and composition extraction

5.1 Datasets

We formed two sub-datasets based on the ArchiveII (**ref**) dataset. We first 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 structure are not well modeled by the parameter set.

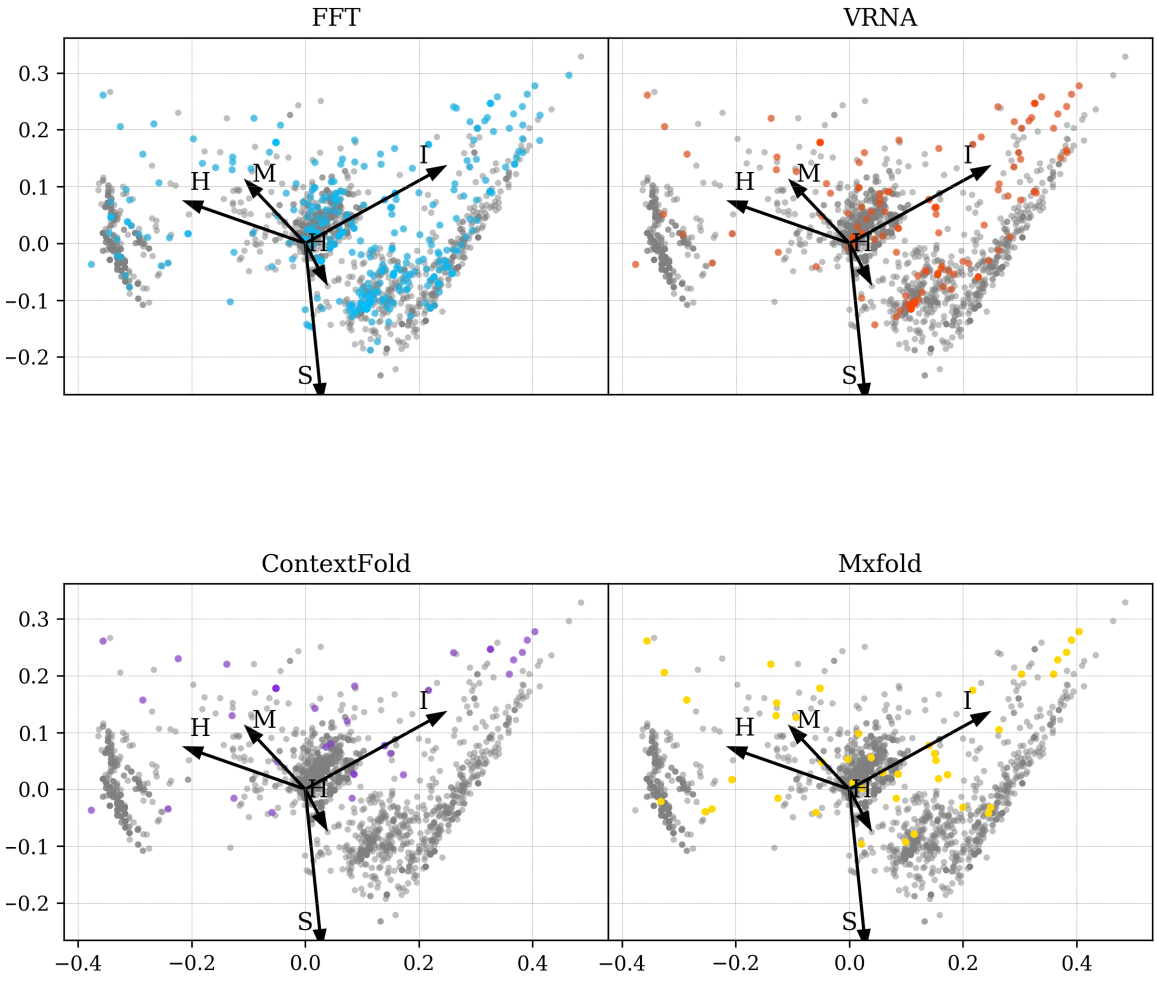

Figure 3: where does the methods failed? PCA RNAfold, Mxfold, FFT, and

6 Concluding discussion

6.1 Good stuff

1. Simple heuristic to compute folding path
2. Versatile method: allow simple modeling of pseudoknot and more information can be encoded in the mirror representation.
3. Performance is comparable although not as good as state of the art in the folding task.
4. One trajectory among the selected produce good structures (close with better accuracy than ML methods).

6.2 limits

1. Choosing the maximum number each time is not an optimal choice
2. In average, the scores are not good. Only a few out of the predicted structures have good scores.
3. The quality of the prediction degrade drastically when the size ¿ 250 from 74% -¿ 50%.
   (a) The stacking method might one cause however, since MFE is degraded as well, we believe that it might partly explain by the thermodynamic model accuracy.
4. The distribution of loop types composition seems to differ between the Boltzmann ensemble and the natural structures.