k-Local Folding: A Local Alignment Approach to RNA Folding

Ben Chugg   Coulter Beeson

1 Introduction

Ribonucleic acids (RNA) play a crucial role in all living organisms, serving both as information storage as well as providing catalytic activity. Given their diverse functions, RNA come in many different varieties, such as mRNA encoding genetic information for translation into proteins, tRNA for the mapping of codons to amino acids, and ribozymes with catalytic activity such as ribosomes and spliceosomes ([2],[9],[1]), as well as numerous other less understood forms. As opposed to DNA, which is double stranded, RNA is often single stranded and forms — usually complex — three dimensional structures by pairing with itself. As with proteins, the three dimensional structure of RNA is critical to its function, and structural prediction is a natural first step when aiming to ascribe function to a given RNA, as well as in the construction of synthetic sequences with novel properties ([10],[4]).

To predict the three dimensional structure of a given RNA sequence it is often necessary to first determine the secondary structure. RNA, as with proteins, will adopt structure(s) that minimize their total energy. The major stabilizing interaction for RNA comes from their intramolecular base pairing. That is, sequences of similar length base pair internally with other complimentary regions [3]. Accordingly, most algorithmic approaches seek to maximize the number of these base pairings. Alternate approaches aim to measure other energetic interactions between bases, such as base stacking, and search for a structure of minimal global energy ([5],[8],[2]). Regardless of the approach used, most modern RNA folding algorithms use a similar recurrence that is amenable to dynamic programming.

We present a modified heuristic approach to RNA folding which seeks to maximize the interactions among regions which are highly complimentary. Our approach may be viewed as a preprocessing step to the typical RNA folding approach. We pair specific regions of the strand using a variant of local alignment, extract these regions from the original strand and run the usual dynamic programming algorithm on the remaining parts of the strand.

2 Background

RNA consists of the four base pairs Adenine ($\texttt{A}$), Guanine ($\texttt{G}$) , Cytosine ($\texttt{C}$) and Uracil ($\texttt{U}$). As opposed to DNA, the base pairs of RNA pair in a complementary fashion: Adenine to Uracil ($\texttt{A}-\texttt{U}$) and Cytosine to Guanine ($\texttt{C}-\texttt{G}$).

There are several models for determining the secondary structure of RNA. A common approach is energy minimization ([3]) which is formulated as a dynamic program as follows. Let $r = r_1, \ldots, r_n$ be a strand of RNA, where $r_i \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{U}\}$ and let $S(i, j)$ denote the optimal score of folding the subsequence $r_i, r_{i+1} \dots r_j \subset r$. Then,

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1, & \text{if } i,j \text{ are a base pair} \\ S(i+1, j) & \\ S(i, j-1) & \\ \max_{i<k<j}\{S(i,k) + S(k+1,j)\}, & \text{bifurcation} \end{cases}$$

This is typically known as the Nussinov algorithm, named after Ruth Nussinov. We will refer to this algorithm as Nussinov. The time and space complexity of this approach are $O(n^3)$ and $O(n^2)$ respectively.

3 A Heuristic Approach: k-Local Folding

Assuming no pseudoknots, Nussinov returns the mathematically optimal alignment under the energy minimization model. Therefore, any new approach cannot hope to beat the scores, but only improve the running time.

We propose a heuristic, k-Local Folding, which aims to take advantage of the quadratic running time afforded by the Smith-Waterman local alignment algorithm. The following definition will help in underlining the intuition behind this approach.

Definition 1. *Let $r = r_1, \dots, r_n$ be a strand of RNA. We say the two regions $r_i, \dots, r_j$ and $r_k \dots, r_\ell$, $i < j < k < \ell$ are complimentary palindromes if, when one region undergoes the mapping $\mathtt{A} \mapsto \mathtt{U}, \mathtt{U} \mapsto \mathtt{A}, \mathtt{C} \mapsto \mathtt{G}, \mathtt{G} \mapsto \mathtt{C}$, they become palindromes of one another.*

For example, AGUUAC and GUAACU are complimentary palindromes.

Regions of an RNA strand r which pair with one another are likely to be almost complimentary palindromes. The idea behind k-Local Folding is to pair these regions together using Smith Waterman local alignment, then run the remaining unpaired regions through Nussinov to determine their most likely secondary structure. We can then concatenate all regions together to obtain a global secondary structure. This approach makes sense biologically as these regions of high complementarity represent many stabilizing hydrogen bonds, and are likely present in biologically active structures.

Formally, we define an algorithm k-Local Folding which takes as input an RNA strand r and a parameter k, runs a local alignment algorithm on the strand to find k high scoring and disjoint complimentary palindromic regions of r. It then passes the remaining unpaired regions to Nussinov's algorithm independently to be folded as usual.

To find complimentary palindromes of r we can run local alignment on r and $\overline{r}$, as shown in algorithm 1, using the following scoring matrix which encourages matchings between complimentary base pairs.

	A	U	G	C
A	−1	1	−1	−1
U	1	−1	−1	−1
G	−1	−1	−1	1
C	−1	−1	1	−1

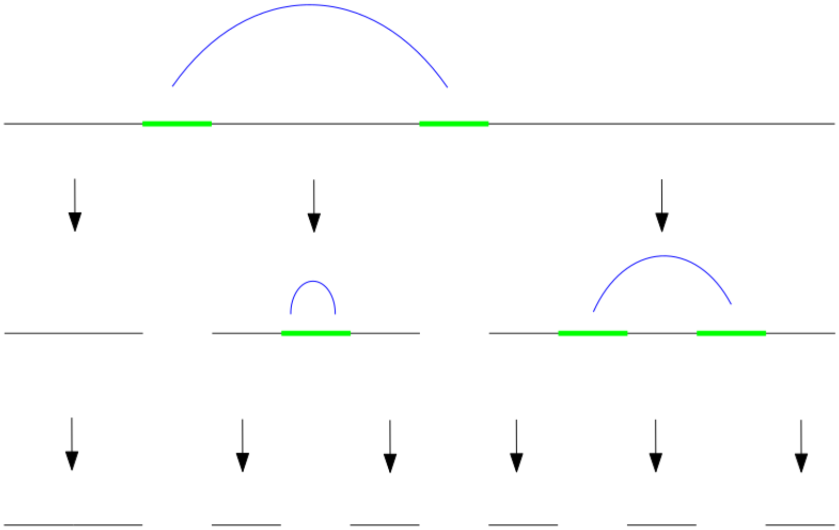


Figure 1: A sequence of RNA undergoing k-Local Folding. Here, we have found 3 local alignments.

Algorithm 1 k-Local Folding

```
procedure k-Local Folding($(r, k)$)
    Initialize stack $S \leftarrow r$
    while The number of local alignments found is $\leq k$ and $S$ not empty do
        $s \leftarrow \text{pop}(S)$
        Let $\overline{s}$ be the reverse of $s$
        Call Local Alignment on $s$ and $\overline{s}$.
        if local alignment found then
            Remove the aligned regions from $s$
            Push all unmatched regions of $s$ back onto stack
        end if
    end while
    Call Nussinov on all unmatched regions of $r$.
end procedure
```

3.1 Runtime Analysis

Fix an RNA strand r with length n.

Lemma 1. *Let $\mathscr{A}$ be the set of local alignments found. Finding k disjoint local alignments of r takes time $O(n^2k)$.*

Proof. We can view the progress of the algorithm as a ternary tree: amortized across each level we perform local alignment on a sequence of size $n - \sum_{A \in \mathscr{A}} \ell_A$. Local alignment takes time squared in the size of the input. Since the tree in the worst case has depth k, the result follows. ◀

Lemma 2. k-Local Folding *runs in time*

$$O\left(\left(n - \sum_{A \in \mathscr{A}} \ell_A\right)^3 + n^2k\right),$$

where ℓ_A is the length of an alignment $A \in \mathscr{A}$.

Proof. Running Nussinov takes cubic time in the size of the input. k-Local Folding first finds k disjoint alignments, then runs Nussinov on the remaining unmatched regions, which have total length $n - \sum_{A\in\mathscr{A}} \ell_A$. The result follows by applying lemma 1. ◀

It is worth remarking that it is very unlikely that the depth of the tree is k: it will more likely be $\log(k)$. Therefore, an average case analysis will yield that k-Local Folding runs in time

$$O\left(\left(n - \sum_{A\in\mathscr{A}} \ell_A\right)^3 + n^2\log(k)\right).$$

Now, it is clear that since $O((n - \sum_{A\in\mathscr{A}} \ell_A)^3) = O(n^3)$, asymptotically there is no difference between k-Local Folding and Nussinov. However, we will demonstrate that there is indeed a difference in practice. To illustrate this it is useful to look at two extreme cases.

1) Consider a perfect hairpin in which case the entire sequence of r is a complementary palindrome eg $r = GGG\ldots CCC$ Smith-Waterman will match the entire strand to itself and Nussinov will not do any work.
2) Consider $r = AAA\ldots AAA$ where there is no complementarity and can be no internal binding. There will be no pruning done by the preprocessing, and Nussinov will do its maximum amount of work.

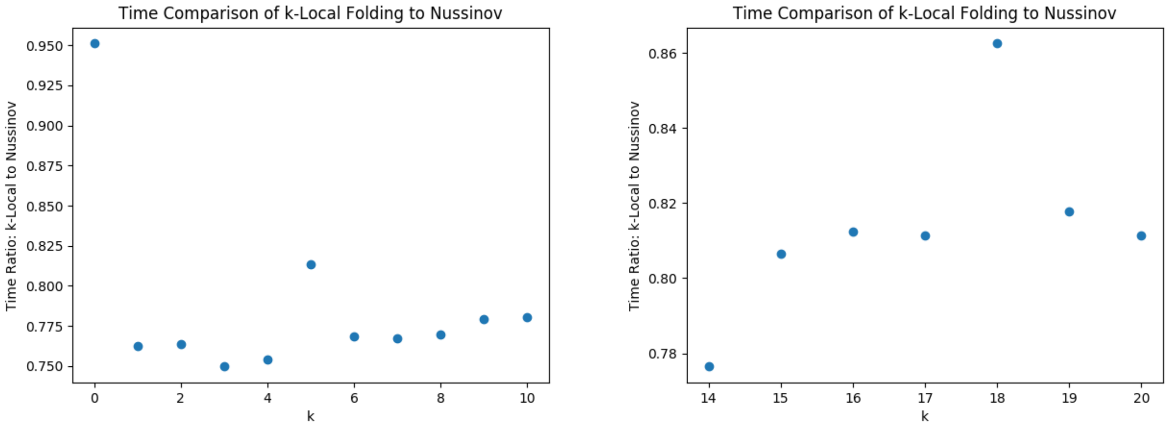
At this point, it is worth reiterating that k-Local Folding is not a method intended to replace Nussinov. Instead, it might be used when a lower score accuracy is acceptable when an increase in speed is desired.

4 Results

Experiments were run on several RNA datasets. A random set of twenty RNA strands were chosen from each set and, for each k, their scores and runtimes were averaged. Only the first 250 base pairs of each strand were kept, for the purposes of time.

The code was written in python, and can be accessed at `https://github.com/bchugg/bwt`.

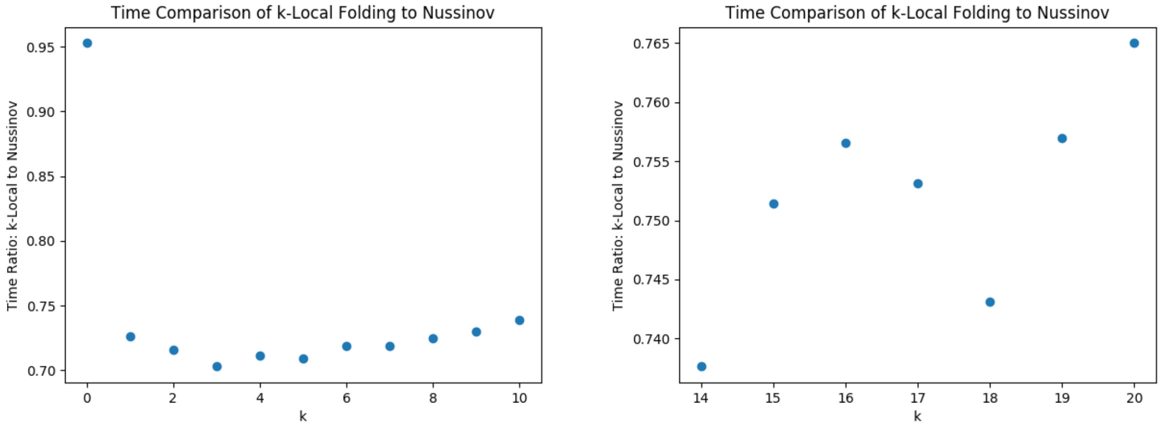
4.1 Runtime Comparison



(a) $k \in \{0, 1, \dots, 10\}$

(b) $k \in \{14, \dots, 20\}$

Figure 2: 16S Ribosomal Subunit RNA



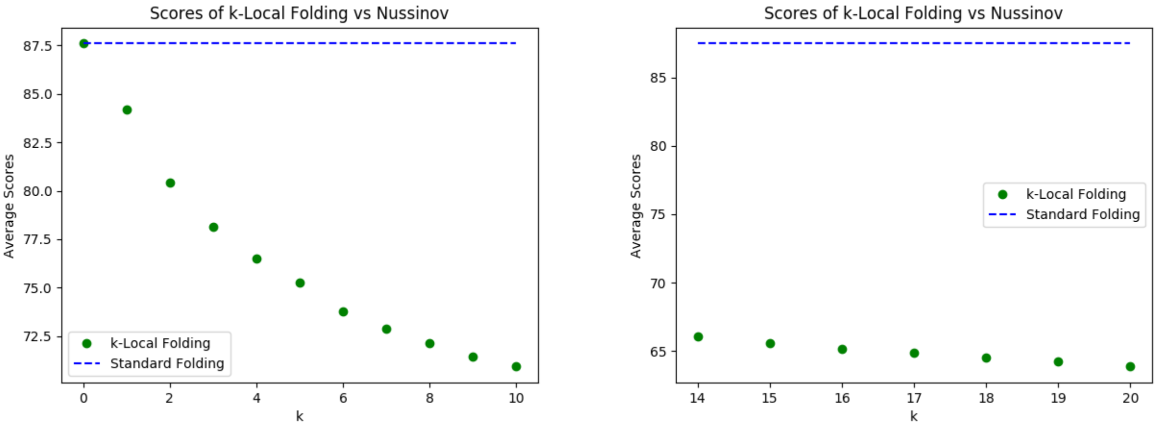
(a) $k \in \{0, \dots, 10\}$

(b) $k \in \{14, \dots, 20\}$

Figure 3: Ciliate Telomerase RNA

For $k = 0$, as expected the ratio is approximately 1:1, since k-Local Folding is not actually finding any local alignments. However, for $k > 0$, we see an immediate drop in the runtime. Interestingly, for all $k > 0$, there does not seem to be much difference across runtimes. Indeed, if we notice the scale of figure (b), we see that relative to figure (a) it remains almost linear. The results for both data sets look relatively similar, potentially indicating some sort of universality for the runtime of k-Local Folding.

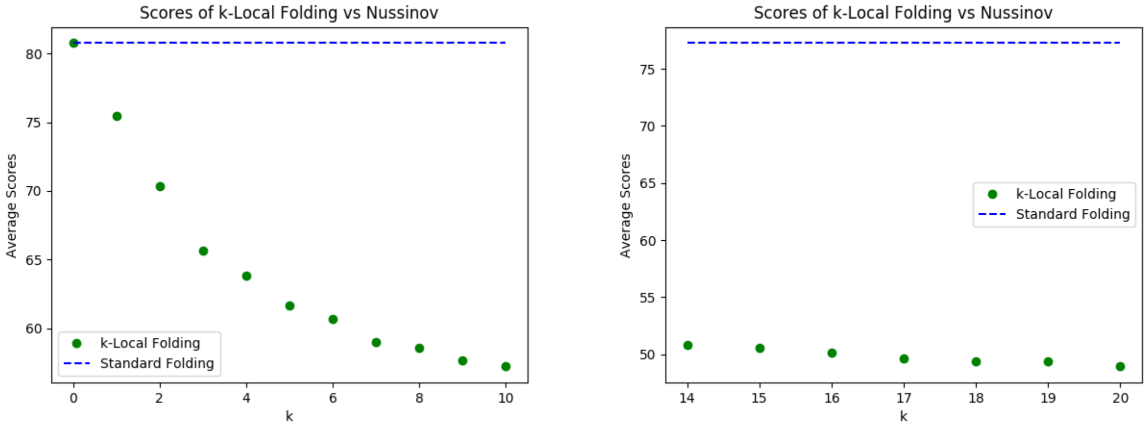
4.2 Scoring Comparison



(a) $k \in \{0, 1, \ldots, 10\}$

(b) $k \in \{14, \ldots, 20\}$

Figure 4: 16S Ribosomal Subunit RNA



(a) $k \in \{0, \ldots, 10\}$

(b) $k \in \{14, \ldots, 20\}$

Figure 5: Ciliate Telomerase RNA

For $k = 0$, the scores are identical as is expected. As k increases, we see a significant decrease in the scores. This can be explained by recalling that Nussinov is mathemtically optimal. As we greedily align more subregions of the RNA sequence, we veer further away from the global alignment returned by Nussinov, hence our score decreases. This makes sense, because by choosing local alignments to pair before running Nussinov we constrain the possible structures reducing the search space, but likely removing the optimal structure.

5 Conclusion

From the two data sets tested, it looks as though k-Local Folding does indeed beat Nussinov in terms of runtime. However, as k increases, the score returned by k-Local Folding decreases. As such, we propose k-Local Folding as only a potential preprocessing step when an increase in speed is more important than a precise score. It is worth noting that even state of the art energy minimization models only get within about 5% of the minimum energy, but this gap still represents potentially tens of thousands of structures, as such the divergence of scores in our approach likely corresponds to drastically different structures. However when dealing with more sophisticated energy minimization models it could also be the case that our heuristic is more likely to approach the optimal structure as these models often score contiguous runs of base pairs higher due to base stacking interactions.

6 Extensions and Further Research

There are several questions which arise from this research.

1. Can we decrease the runtime of k-Local Folding if we choose better data structures?

   For instance using a heap to store the location of where alignments start means we could run the local alignment once resulting in a run time of $O(n^2 \log n)$ but this is worse if $k \in o(\log n)$

2. k-Local Folding can be modified to accommodate pseudoknots.

   When looking for additional alignments, or when handing the remaining problems over to Nussinov if the segments are simply concatenated back together then this approach will find pseudoknotted structures. As k increases these structures could become extremely knotted and likely are not biologically relevant. Additionally we did not explore this option so that our results were still directly comparable to Nussinov.

3. Implement k-Local Folding using the Four Russians speedup for Nussinov and using the Burrow-Wheelers Transform to speedup Smith-Waterman.

   Both of these optimizations are applicable and would likely result in this approach being even faster. Additionally any other new optimizations to Smith-Waterman or Nussinov would also improve our technique.

4. Can the sub-optimal alignments for k-Local Folding be found using probabilistic sampling using a Hidden Markov Model?

   This approach could be successful, and deserves further research. It was not addressed in this project due to lack of time and a concern that a large number of samples would be needed to find extremely short alignments.

5. Lower bounding the expected size of alignments found in a sequence of RNA to provide stronger guarantees about how much the problem size is reduced prior to being solved by Nussinov

6. Convert our heuristic approach to an approximation by proving bounds on how far our scores deviate from the optimal scores.

References

[1] Jamie H Cate, Anne R Gooding, Elaine Podell, Kaihong Zhou, and et al. Crystal structure of a group i ribozyme domain: Principles of rna packing. *Science*, 273:5282:1678–1685, 1996.

[2] Jennifer A Doudna and Jon R Lorsch. Ribozyme catalysis: not different, just worse. *Nature Structural and Molecular Biology*, 12:395–402, 2005.

[3] Sean R Eddy. How do rna folding algorithms work? *Nature Biotechnology*, 22:1457–1458, 2004.

[4] Zemora Georgeta and Christina Waldsich. Rna folding in living cells. *RNA Biology*, 7.6:634641, 2010.

[5] T.W. Lam, W.K. Sung, S.L. Tam, C.K. Wong, and S.M. Yiu. Compressed indexing and local alignment of dna. *Bioinformatics*, 24:6:791–797, 2008.

[6] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 36:5:589–595, 2010.

[7] Giovanni Manzini. An analysis of the burrows-wheeler transform. *Journal of the Association for Computing Machinery*, 48:3:407–430, 2001.

[8] David H Mathews and Douglas H Turner. Prediction of rna secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16:3:270–278, 2006.

[9] Markus C Wahl, Cindy L Will, and Reinhard Lurhmann. The spliceosome: Design principles of a dynamic rnp machine. *Cell*, 136:4:701–718, 2009.

[10] Christian Hner zu Siederdissena, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. A folding algorithm for extended rna secondary structures. *Bioinformatics*, 27, 2011.