RNA Secondary Structure Prediction

Gary Ho
BTRY 4840 Final Project

Introduction

Ribonucleic acids, also known as RNA, are a kind of biological macromolecule that is essential to all forms of known life. RNA, like DNA, is a nucleic acid that is composed of nucleotide monomers. RNA has the canonical monomers cytosine (C), guanine (G), adenine (A), and uracil (U) in which C pairs with G and A pairs with U. In contradiction with Watson-Crick base pair rules, there exists multiple "wobble pairs" in RNA, one of them allowing G and U to pair. Unlike DNA, RNA is single-stranded and thus has the ability to base pair with itself. The self base pairing interactions within RNA is known as its secondary structure. We would like to explore how the given sequence of RNA nucleotides, the primary structure, can be used to determine the secondary structure.

Since RNA folding is considered a hierarchical process, the secondary structure will affect the three dimensional conformation, the tertiary structure. Biologists would like to have a way to predict RNA conformation given primary structure because due to technological limitations, it is still a challenge to experimentally determine tertiary structures[1]. As a result, many research groups developed computational pipelines to predict secondary structure from primary structure and tertiary structure from the resulting secondary structure. Therefore, determining the secondary structure is a crucial step intermediate step.

In particular, biologists would like to know the secondary/tertiary structure of non coding RNAs because it would help explain their biochemical mechanism of action. One such example of non coding RNAs are ribozymes, which have catalytic activity. Molecules that have catalytic activity are thought to work by having an active site, in which substrates can bind to the active site and achieve a lower activation energy, allowing more substrates enough energy to undergo a reaction. Active sites are optimized to bind to a specific substrate due the substrates' intermolecular forces with the enzyme's unique tertiary structure. In addition to ribozymes, many new non coding RNAs such as signal recognition particle RNAs, riboswitches, and ribonucleases have been discovered to have important roles in biological processes[1].

Nussinov Algorithm

A simple approach to determine secondary structure is to maximize the number of base pairings within a fold. This can be done by dynamic programming (modified from Durbin from recursive

to iterative)[2].

Let us formally define the problem. Given a sequence of nucleotides $S = x_i \dots x_j$, in which each element $\in \{A, C, G, U\}$, give the fold P that maximizes $|P|$, the total number of base pairings in P. P is an ij-substructure of S iff $P \subseteq \{i, \dots, j\}^2$, the power set of bases in the ij sequence. Create a 2D matrix to memoize the resuts of max $|P|$ within an ij substructure of S.

Initialize a nxn matrix where $n = |S|$, the length of the nucleotide sequence; call this matrix opt. Then fill the diagonal with zeros. More specifically, $opt(i, i-1) = 0$ for i=2 to n and $opt(i, i) = 0$ for i=1 to n. The recurrence is given by:

$$opt(i,j) = max \begin{cases} opt(i+1, j) \\ opt(i, j-1) \\ opt(i+1, j-1) + \delta(i,j) \\ \max_{i<k<j} opt(i,k) + opt(k+1, j) \end{cases}$$

where

$$\delta(i,j) = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ can base pair} \\ 0 & \text{if } x_i \text{ and } x_j \text{ cannot base pair} \end{cases}$$

There are four ways to get the maximal pairing of $x_i, ..., x_j$ given that we know the maximal pairings of the ij-substructures. The first case is add unpaired base x_i onto best structure for subsequence $x_{i+1}, \dots, x_j$. Second case adds unpaired base x_j onto best structure for subsequence $x_i, \dots, x_{j-1}$. The third case considers a match between x_i and x_j, adding the pair onto best structure found for subsequence $x_{i+1}, \dots, x_{j-1}$. The fourth case considers when x_i and x_j do not form a base pair with each other, but each instead forms a base pair with an base within $x_i, \dots, x_j$ (no crossing), combining two optimal ij-substructures. Once the opt matrix is filled, we can traceback from $opt(1, n)$ to obtain one of the maximal matching folds. Therefore, the overall time complexity is $O(n^3)$ and space complexity is $O(n^2)$.

Zuker Algorithm

Since the Nussinov algorithm produces the fold with the maximal pairings, it does not yield biologically relevant structures. One way is to achieve biologically relevant structures is to minimize the free energy of RNA given experimentally measured energies. Recall that the Gibbs free energy of a system is given by $G = H - TS$ where H is the enthalpy, T is the temperature, and S is the entropy. We want to compute a structure P that minimizes the change in free energy ΔG due to folding. This is done by summing the free energy change that is generated by structural motifs (stacks, loops, bulges, etc.)[3,4].

Let us define these motifs more formally. As above, let S be a sequence of RNA nucleotides where $|S| = n$ and P be a ij-substructure (fold) of S. (For sake of simplicity, x_i notation is now simplified to i.) If i is unpaired in P_{folded}, then there is no other nucleotide base j such that $(i, j) \in P$ or $(j, i) \in P$.

- A stack is closed by $(i, j) \in P$ iff $(i+1, j-1) \in P$. $eS(i, j)$ is the free energy of stacking pair closed by (i, j); depends on all four bases[3].

- A hairpin loop is closed by $(i, j) \in P$ iff all bases k such that $i < k < j$ are unpaired in P. $eH(i, j)$ is the free energy of hairpin closed by (i, j); depends on unpaired bases adjacent to (i,j), terminal mismatches (helix to opening), and loop size[3].
- An internal loop (i, j, i', j') is closed by $(i, j) \in P$ and $(i', j') \in P$. In the $5' \rightarrow 3'$ direction $i < i' < j' < j$. In an internal loop $i' - i > 1$ and $j - j' > 1$, where all bases $i + 1, \ldots, i' - 1$ and $j' + 1, \ldots, j - 1$ are unpaired.
- A bulge (i, j, i', j') is similar to an internal loop except $i' - i = 1$ (exclusive) or $j - j' = 1$. It is a right bulge iff $i' - i = 1$ and a left bulge iff $j - j' = 1$. $eS(i, j, i', j')$ is the free energy of internal loop/bulge closed by (i, i') and (j', j); depends on the four paired bases and loop size[3].

Note that all thermodynamic parameters are given by Turner[3].
Now initialize a nxn matrix W and another nxn matrix V, where $W(i, j)$ contains the minimum $E(P)$ (the minimum energy ij substructure (fold) of S). $V(i, j)$ contains the minimum $E(P)$ where $(i, j) \in P$ meaning that i and j must be paired. The $W(i, j)$ matrix is initialized as 0 for all positions. The $V(i, j)$ matrix is initialized as ∞ for all positions. The recurrence for each is given as[4]:

$$W(i,j) = min \begin{cases} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ \min_{i<k<j} W(i,k) + W(k+1,j) \end{cases}$$

$$V(i,j) = min \begin{cases} eH(i,j) \\ eS(i,j) + V(i+1,j-1) \\ \min_{i<i'<j'<j} eL(i,j,i',j) + V(i',j') \\ \min_{i+1<k<j-1} W(i+1,k) + W(k+1,j-1) \end{cases}$$

Note that if i cannot base pair with j, $V(i, j) = \infty$. We can see than W matrix is analogous to scoring matrix in Nussinov. The traceback starts from $W(1, n)$ like in Nussinov with a similar traceback, except we look at $V(i, j)$ in a matching instead of $W(i + 1, j - 1)$ as in Nussinov.
The overall time complexity is $O(n^4)$ and space complexity is $O(n^2)$. However, we can limit the double loop within the eL case by restricting the size of internal/bulge loops, making runtime closer to $O(n^3)$ as restriction becomes more strict. In practice, as the sequence gets very large, the runtime is much longer than Nussinov.

Pseudoknot Prediction using Iterative Loop Matching

A pseudoknot is formed if a fold P of a nucleotide sequence S, contains at least two base pairs that are crossing. That is if (i, j) and (i', j') are two base pairings and either $i < i' < j < j'$ or $i' < i < j' < j'$, then these two base pairs are crossing.

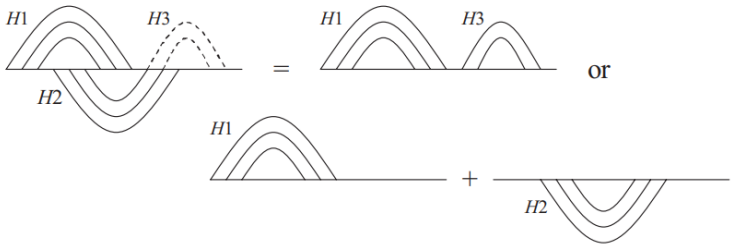
So far, both Nussinov and Zuker's algorithms fail to consider the existence of pseudoknot structures in the secondary structure. This is because the recurrence breaks down because we have assumed that base pairs outside of (i, j) do not contribute to computed substructures. In fact, it was shown by Lyngso and Pedersen [5] that RNA folding with pseudoknots is NP-hard by reduction from 3SAT.

Therefore, in order to model RNA folding with pseudoknots, we need to give an approximation algorithm that uses some heuristic.

One such heuristic outlined by Ruan et. al[6], is to keep the most reliable helix in the final fold after running Zuker (or another folding DP algorithm), and perform hierarchical folding by continually repeating (finding most stable helix and then deleting it from sequence in next iteration and running again). The program terminates when there are no helices left in the current iteration, but does not add any base pairings (the leftovers) that are not in a helix. Notice that when we "delete" base pairs from the sequence, we are "joining" two segments together in the next iteration so we must add a constraint, VLOOPLENGTH, which describes the minimum virtual distance required between two paired bases after the first iteration. That is, after the first iteration, two bases with virtual distance less than VLOOPLENGTH are not allowed to pair. VLOOPLENGTH defaults to 3.[6]

The most stable helix is the helix with lowest free energy, computed using the stack free energies from Turner[3]. Biochemically speaking, this assumes that the helix with the most minimal free energy is folded first, followed by the next most stable helix, and so forth. This approach is titled iterative loop matching (ILM) by Ruan.

This works much better than running Zuker (or another DP algorithm) twice consecutively and then adding the results together. This is because during the first run, it is very likely that these pairs in the true pseudoknot would have already formed some false positive base pairs, that precludes the pseudoknot from forming during the second run[6].



If H1, H2, and H3 are potential helices found by a DP algorithm, the approach of running the DP algorithm twice will result in the top (H1+H3) because during the second iteration, H3 prevents H2 from forming since the bases are already paired. In the ILM approach, we can form the correct fold by taking H1 in the first iteration and then picking up H2 in a subsequent iteration.

The time complexity of the algorithm is pseudopolynomial $O(n^3H)$ where H is the number of helices found, and space complexity remains $O(n^2)$.

Results

Database used for non-pseudoknot data is RNAStrAlign[7].

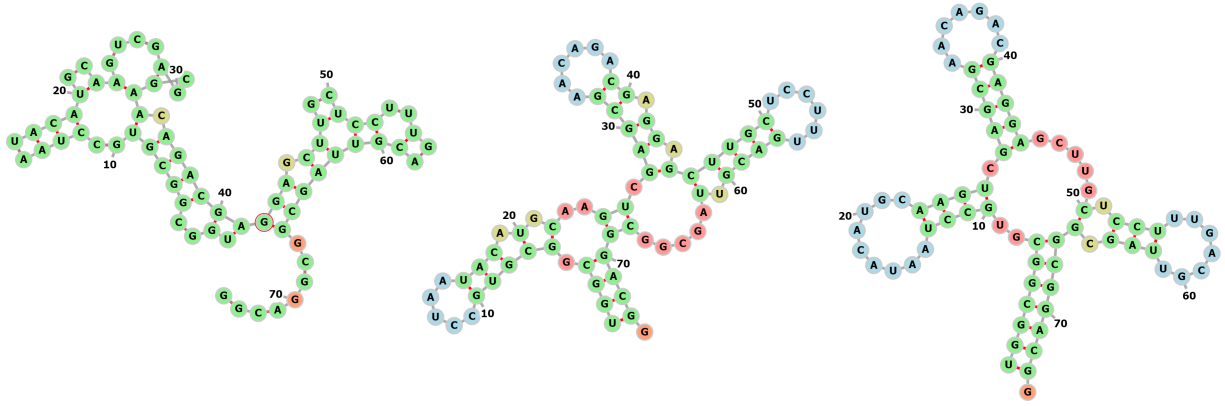


Figure 1: On the left is a tRNA (tdb0007697) folded with Nussinov algorithm. Middle is folded with Zuker. Right is ground truth. (Diagrams generated by ViennaRNA's forgi[8] on their online server)

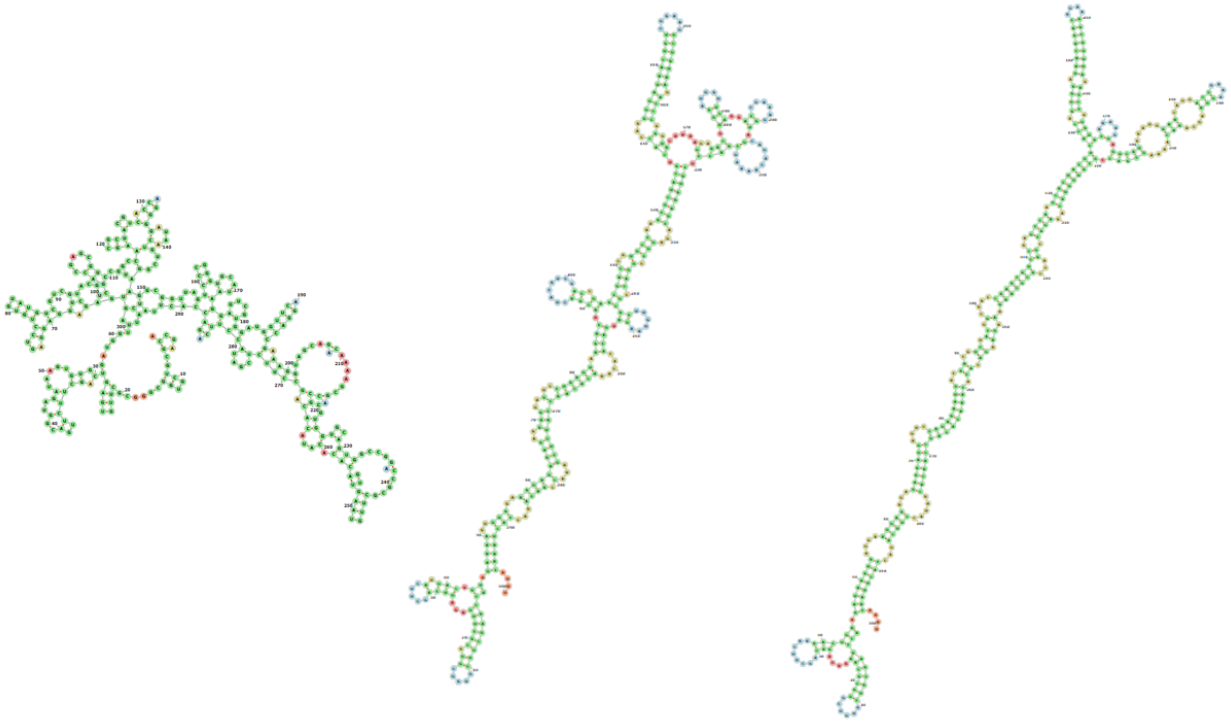


Figure 2: Folds of a SRP (signal recognition particle, Caen.eleg.AY948607). Outputs in same order as Figure 1.

Here are typical (average in closeness of structural motifs) output from Nussinov and Zuker implementations that shows how close the foldings are from ground truth. Note that in both figures, the sequence does not have a pseudoknot in the ground truth. I chose these examples because they are instructive examples that display the shortcomings of each implementation.

When looking at Nussinov implementations, one can see that the resulting structures are extremely

different from ground truth. Base pairing is maximized resulting in many 0 length hairpins and include many helices (stacking of consecutive bases).

In comparison, Zuker implementation gets much closer to the ground truth (in terms of structural motifs). In Figure 1, the three hairpins from the ground truth are approximately in the correct location. However, there are also differences such as having additional multiloops and more internal loops (including bulging). The additional multiloops/internal loops in Zuker implementation can best be seen in Figure 2. A multiloop is a junction where three or more stacks are joined together. There are two multiloops in the Zuker diagram from Figure 1 above.

Table 1. This table contains the ratio of correct base pairs found / all base pairs in ground truth. All sequences ran contained no pseudoknots in ground truth. (See results folder to see which sequences were chosen).

RNA Type	Zucker Avg	Zuker Std Dev	Nussinov Avg	Nussinov Std Dev
5S	0.345	0.124	0.080	0.098
16S	0.279	0.120	0.059	0.041
RNAse	0.401	0.091	0.069	0.043
SRP	0.294	0.166	0.014	0.03
tRNA	0.519	0.149	0.081	0.133
All	0.366	0.153	0.059	0.078

The ILM algorithm for computing pseudoknots can miss completely, give a fold that is close, or give a fold that is "lukewarm" in its closeness. Since ViennaFold is unable to display the directed graph diagrams of RNA folds with pseudoknots, I will using a text and arc diagram notation for pseudoknot folds. The text display (ground truth) is called a dot bracket notation where open parentheses indicate that the base is paired to another base ahead of it. The arc diagram (ILM output) shows which bases are paired together.

Here is one example where the ILM algorithm is is mostly correct (sequence taken from PseudoBase[9]).

```
            10        20        30        40        50        60
#   123456789|123456789|123456789|123456789|123456789|123456789|12345678
$ 1 GGGGUUCAUGUUGUCGACCUUCACGUGGUGAGCCCUGUCAACUGACUGCUGUCAGGCUAACAGACAAC=68
% 1 (((((((::[[[[[[[[(((::[::::)))))))))]::::(((((::::)))))::::]:]]]]]]
```

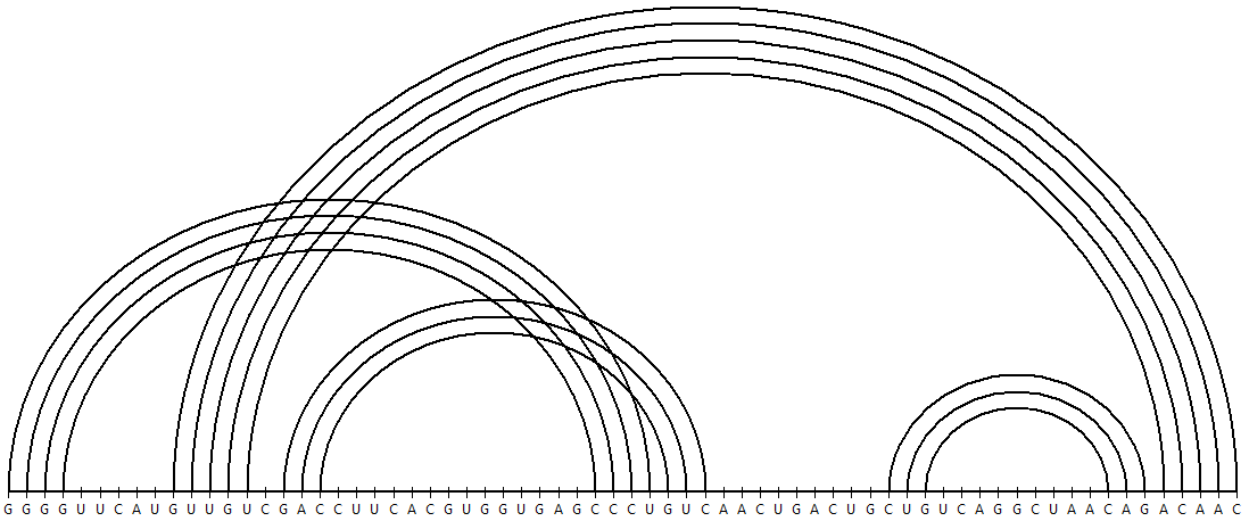


Figure 3: A folding created by ILM for the sequence BK006888 (from a HDV-like ribozyme)

We can see that the main base pair helices are correct, namely the left most and right most helix stems.

Here is an example where ILM is not very close to ground truth.

```
     1050       1060       1070       1080       1090
#       56789|123456789|123456789|123456789|123456789|(59)
$ 1045 UUUAAACUGUUGAGAGGUGCCUGGAGCGCCUGCAGGCAUCUCUGUU=1090
% 1045 ::::::::::::((((((((((::::[[[[[))))))))))):::
     1150       1160
# (59) |123456789|12
$ 1150 UAUAAUGCAGGCA=1162
% 1150 :::::::]]]]]:
```

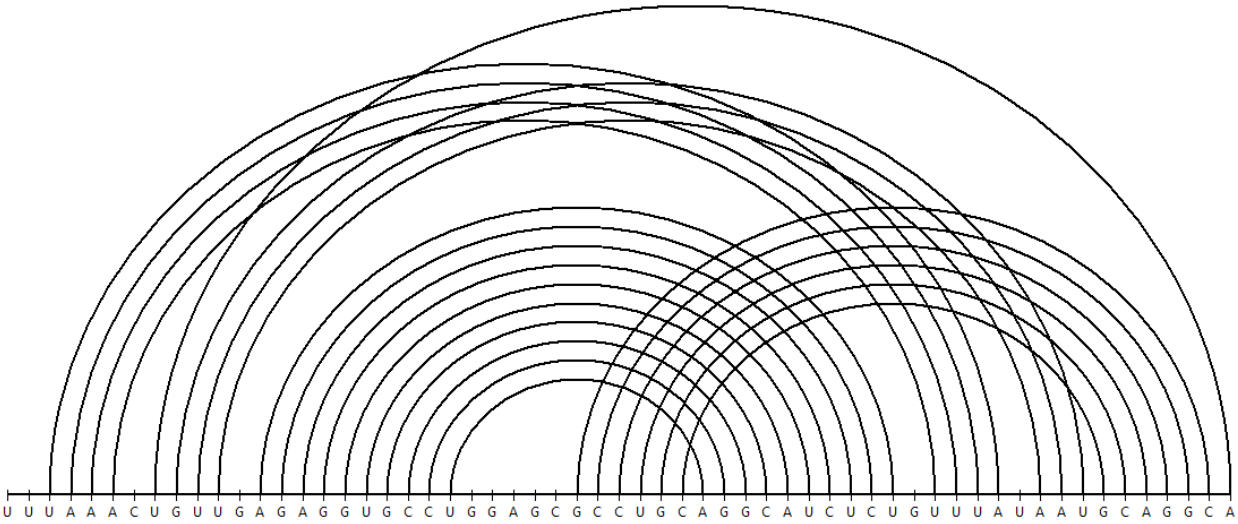


Figure 4: A folding created by ILM for the sequence X52374 (from ORF1a/ORF1b (polymerase) ribosomal frameshift site of Berne virus)

Now here is an example where ILM completely misses:

```
        1590       1600       1610       1620
#       123456789|123456789|123456789|123456789|1
$ 1581 GGGAAAUGGACUGAGCGGCGCCGACCGCCAAACAACCGGCA=1621
% 1581 ::::::::::::::((((:[[[[:))))::::::::]]]]:
```

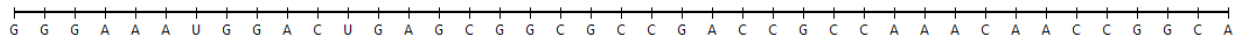


Figure 5: A folding created by ILM for the sequence AF352025 (ribosomal frameshift signal ORF-2/3 of beet chlorosis virus). Note that there is no fold!

Sometimes, ILM gets one helix of a pseudoknot correct, and misses the other one.

```
     1590       1600       1610       1620       1630
#       |123456789|123456789|123456789|123456789|123456
$ 1590 AAAAAACUAAUAGAGGGGGGACUUAGCGCCCCCCAAACCGUAACCCC=1636
% 1590 ::::::::::::::[[[[[[:::::(((]]]]]]::::)))::::::
```

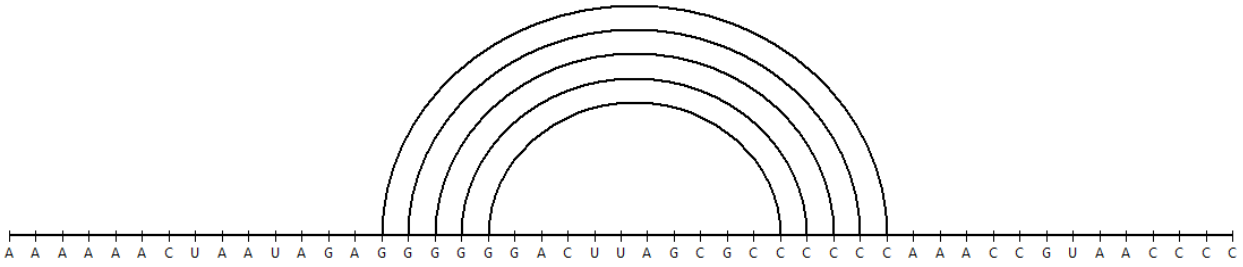


Figure 6: A folding created by ILM for the sequence AF352025 (ribosomal frameshift signal ORF-2/3 of beet chlorosis virus). Only one of the two helices detected.

So far all of these sequences are short, with one pseudoknot. Here is an example with a telomerase sequence (AF221916.94-481, from RNAStrAlign) that is 388 base pairs long.

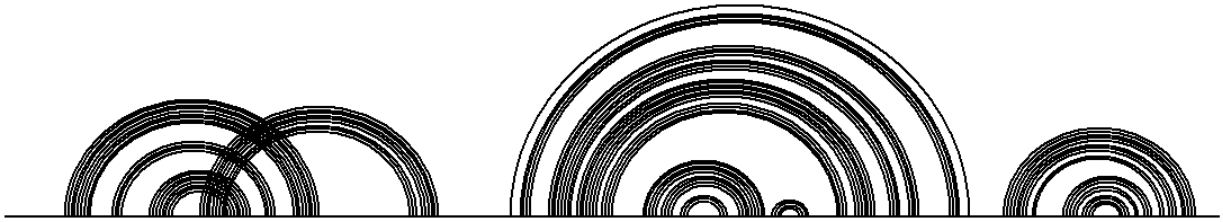


Figure 7: The ground truth fold of the telomerase sequence.

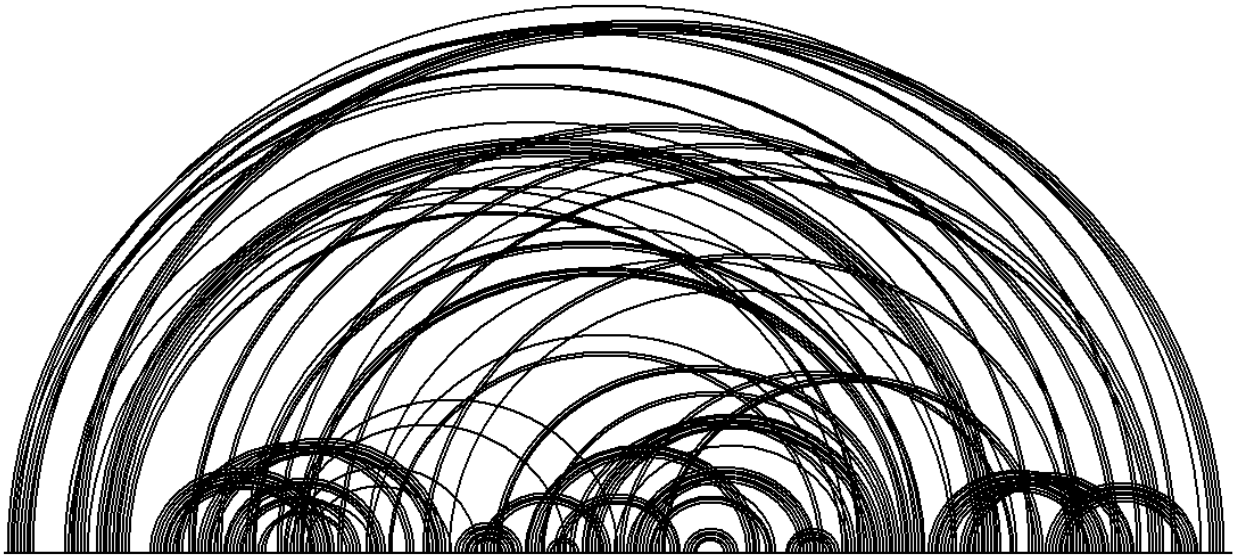


Figure 8: The calculated fold using ILM.

I only ran a few long sequences using ILM due to the blowup of pseudopolynomial runtime. And in all of them, the folds are not really close to the ground truth. The ILM outputs many more overlapping pseudoknots than ground truth, and does not exhibit the "coherency" (i.e. helices

with larger runs of consecutive base pairs) of the helix matches that the ground truth has.

Now lets try ILM on a sequence where ground truth has no pseudoknots.

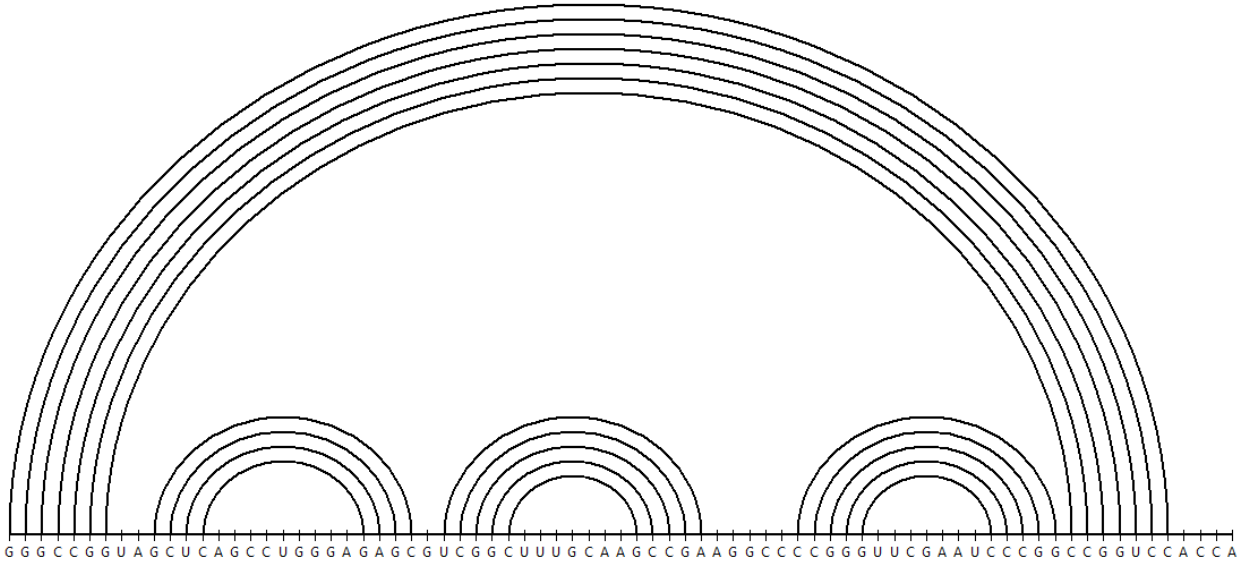


Figure 9: The ground truth fold of tRNA sequence (tdbD00000022).

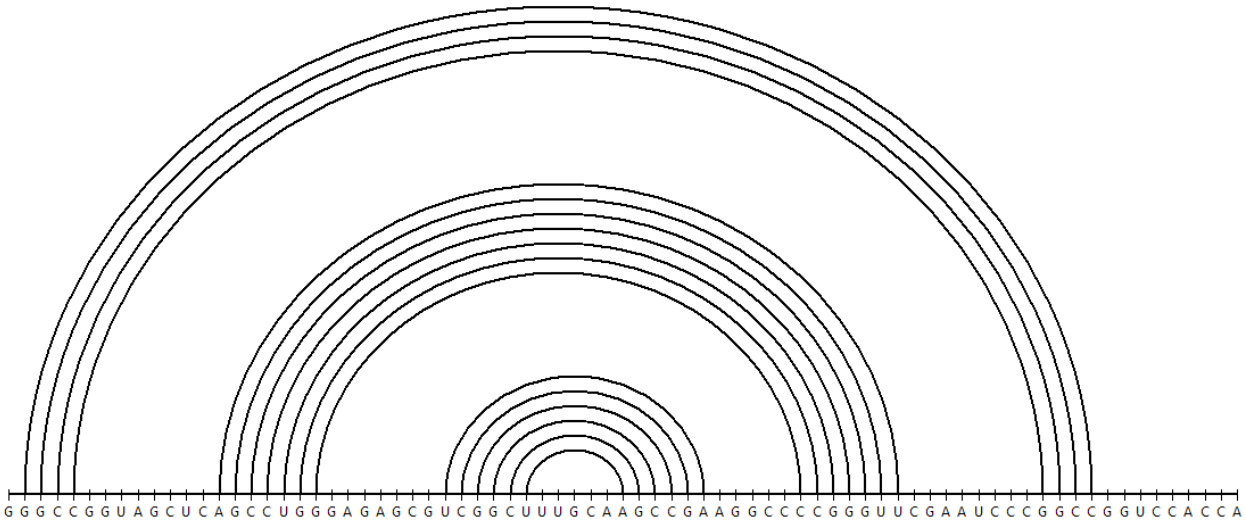


Figure 10: The calculated fold using ILM (same sequence run in Figure 9).

ILM computes that there is no pseudoknot (the output is very similar to Zuker output). Let's try a larger sequence.

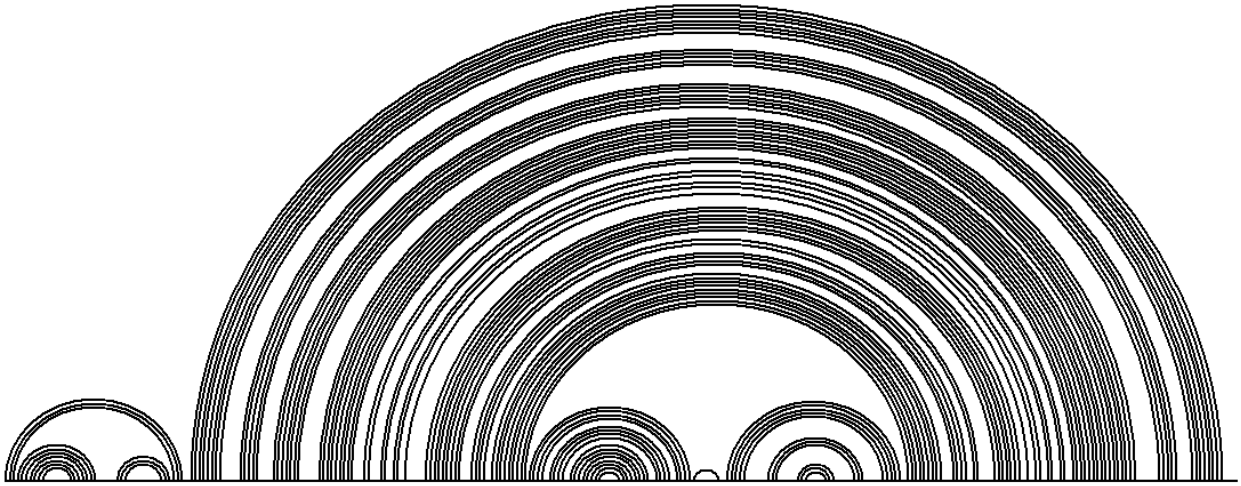


Figure 11: The ground truth fold of tRNA sequence (SRP, Caen.eleg.AY948607).

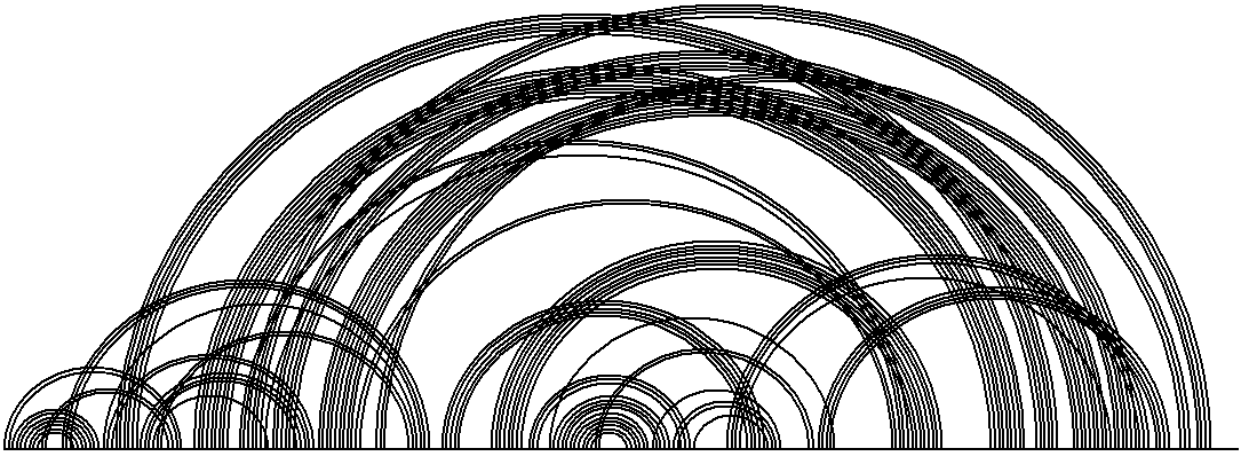


Figure 12: The calculated fold using ILM (same sequence run in Figure 11).

We can see that ILM computes that there is a pseudoknot when there is not, which is usually the case upon when running sequences that are longer, regardless if they contain a pseudoknot.

Discussion

We can see that the fold generated by the Nussinov algorithm differs wildly from ground truth in terms of structural motifs. In particular, since there is no minimum size hairpin, there are multiple hairpins with size 0, which can be seen by the sharp turns at many helix ends. There are also lots of multiloops that are chained together, which best can be seen in Figure 2. The best and only explanation of this is that Nussinov results in the fold with the maximal base pair matchings, which is unlikely to be the most thermodynamically stable. Such a matching will almost always have hairpins of size 0 or 1.

In contrast, the Zuker algorithm creates folds that are quite similar to the ground truth in terms of structural motifs. However, there are still some differences in the fold to the ground truth. Most notably, there are lots of multiloops, bulges, and internal loops that are not found in the ground truth. In the original (this) implementation of Zuker's algorithm[4], there is no recurrence that penalizes a multiloop. We can see this by how multiloops are "chosen" in the algorithm; multiloops are given by the "bifurcation loop" recurrence in $V(i,j)$, $\min_{i+1<k<j-1} W(i+1,k) + W(k+1,j-1)$. Thus if there are multiple helices within (i+1,k) and (k+1,j-1) and the bifurcation loop is chosen to be the lowest free energy, there will be a multiloop. From experimental data from Turner[3], we can this is not actually the case in reality and that multiloops will incur a penalty (increased free energy) around 3 to 6 kcal/mol for a multiloop with 4 helices. This prevents many multiloops from forming, which is for instance why it is likely there is not a 2 multiloops in the ground truth of Figure 1.

I think that the appearance of multiple internal loops and bulges is partly due to the multiloop implementation since the fold would be slightly incorrect to begin with, but also due to how free energy of internal loops and bulges are calculated. For example, Turner's parameters show that an internal loop of size 1 to 3 incurs no penalty in free energy (for initialization). Also, bulges of size 2 somehow have a lower free energy initialization penalty than bulges of size 1 (bulge of 2 has ΔG= 2.8 kcal/mol and bulge of 1 has ΔG=3.8 kcal/mol). We can see many bulges of size 2 in the Zuker diagram for Figure 2. I don't think that the data is somehow wrong, but rather that the results of the algorithm are dependent on the experimental data that is fed to it.

From table 1, we can see that although Zuker algorithm's accuracy in predicting base pairs is not great (average of 36.6% overall), it does much better than Nussinov (average 5.9% overall). This makes sense because Zuker produces folds with more similar motifs to ground truth, so it is also more likely to have a correct base pair. Of course, when an incorrect motif is chosen in Zuker, the rest of the RNA will likely be affected in its base pairing, so accuracy suffers even if overall structural motifs are similar.

To analyze the effectiveness of ILM algorithm, let us first look at Figure 3, the arc diagram generated by running on the sequence above the figure. We can see that there is a general match between ground truth and the ILM output as the dot bracket notation coincides with the arc diagram for the majority of the left (5′) end of the RNA. We can see that ILM fails to pick up the small psuedoknot pair between base 22 and base 27 since it is not a helix of considerable size (no stacking stabilization).

Now look at Figure 4. We can see that ILM did indeed pick up the two helix stems in the pseudoknot, but has also added additional helices not found in the ground truth. This is just a side effect of running ILM; it believes there are still more helices since it has found a helix in the next iteration after finding the true pseudoknot. Also note the "stray pairing" between U-A found in Figure 4, which adding the constraint VLOOPLENGTH attempts to solve, but in the iteration in which that pair was added it was not large enough to stop this pairing. Increasing VLOOPLENGTH too drastically will lead to less helix matches so it cannot be set too high. As the RNA sequence gets longer, this becomes more of a problem (see Figure 8 and Figure 12).

Ultimately, the effectiveness of ILM is heavily impacted by which DP algorithm it chooses to iterate with. Figure 5 shows this point because no helix found in ILM means that no helix was

found in the Zuker algorithm to begin win. Figure 6 also shows this because ILM could not find the second helix, but it could also be due to the VLOOPLENGTH constraint.

If the sequence length is larger than about 100 nucleotides, there tends to be a pseudoknot in the output, regardless if there is one in the ground truth, just due to increased chance of finding a pseudoknot forming helix (see Figure 12). As RNA sequence length decreases, ILM tends to be more correct for pseudoknot free sequences (see Figures 9 and 10).

Overall, Nussinov algorithm is quite frankly terrible at estimating the true fold of a RNA sequence, not creating the correct structural motifs or the correct base pairing. The Zuker algorithm in its current implementation falls a bit short of expectation, generating folds that have similar structural motifs to the ground truth, but ultimately has a subpar accuracy in determining which base pairs are created. Perhaps accuracy would have increased if multiloops were taken into consideration. The ILR method of finding pseudoknots with a Zuker iteration is one that makes intuitive sense, but breaks down with larger datasets.

References

1. Zhao, Yunjie, Jun Wang, Chen Zeng, and Yi Xiao. 2018. "Evaluation of RNA Secondary Structure Prediction for Both Base-Pairing and Topology." Biophysics Reports 4 (3): 123–32. https://doi.org/10.1007/s41048-018-0058-y.

2. Durbin, Richard. 2010. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Cambridge Univ. Press.

3. Turner, Douglas. 2009. "Turner 2004 RNA Folding Parameters." Rna.Urmc.Rochester.Edu. June 30, 2009. https://rna.urmc.rochester.edu/NNDB/turner04/index.html.

4. Zuker, Michael, and Patrick Stiegler. 1981. "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information." Nucleic Acids Research 9 (1): 133–48. https://doi.org/10.1093/nar/9.1.133.

5. Lyngsø, Rune B., and Christian N. S. Pedersen. 2000. "RNA Pseudoknot Prediction in Energy-Based Models." Journal of Computational Biology 7 (3–4): 409–27. https://doi.org/10.1089/106652700750050862.

6. Ruan, J., G. D. Stormo, and W. Zhang. 2003. "An Iterated Loop Matching Approach to the Prediction of RNA Secondary Structures with Pseudoknots." Bioinformatics 20 (1): 58–66. https://doi.org/10.1093/bioinformatics/btg373.

7. "Mathews Lab Publications." n.d. Rna.Urmc.Rochester.Edu. https://rna.urmc.rochester.edu/publications.html. (Citation 112)

8. "TBI - Forna: RNA Secondary Structure Visualization Using a Force Directed Graph Layout." n.d. Rna.Tbi.Univie.Ac.At. http://rna.tbi.univie.ac.at/forna/.

9. "FHD(Eke)van Batenburg: Pseudoknot Retrieval by Class." n.d. Www.Ekevanbatenburg.Nl. http://www.ekevanbatenburg.nl/PKBASE/PKBGETCLS.HTML.