RNA Secondary Structure Prediction

Background: Though the primary structure of RNA, single stranded and linear such as in mRNA, is highly important for the functions of a cell, it accounts for only between 1%-5% of cellular RNA. The rest of the cellular RNA is composed of RNA in its secondary structure, the majority in the form of tRNA or rRNA. The secondary structure of RNA is arrived at through RNA molecules folding in upon themselves and forming bonds between their own complementary nucleotide bases. The structure of the resulting molecule determines its function within the cell.

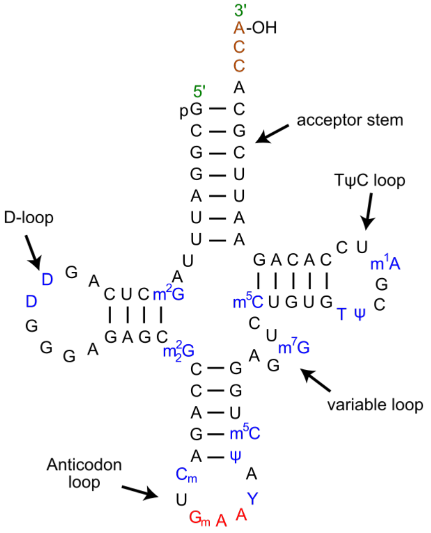

Figure 1. Secondary *cloverleaf structure* of tRNA^{Phe} from yeast.

The secondary structure of RNA is difficult to predict because the number of possible structures increases exponentially as the sequence length increases. It is generally accepted that the most stable possible structure is the most likely to occur, and that the most stable possible structure will be the one with the most bonded complementary base pairs.

Data Collection: Data was collected from the RNA STRAND database from the BETA Lab of the University of British Columbia. This includes the sequence of the RNA molecule and the predicted secondary structure of that sequence.

High Level Program Steps: Given a string of text the program will find all the possible ways of 'splitting' the string into two new ones to simulate the folding of RNA. It essentially cleaves the string at a given point and arranges the two resulting strings into the directions they would be had it really folded. For each split possible it calculates the number of aligning complementary bases and returns the split with the highest score. It then finds sections within the alignment that are unpaired and performs the function again on the sequences of both sides of the unpaired section. Finally it performs the function again on any part of the sequence that extended past the alignment based on the length of the two strings.

Discussion of Results: The prediction of RNA secondary structures using this method was highly accurate for predicting both the number of paired bases structure for shorter sequences, such as 17 base pair tRNAs like that shown in Figure 2. It was generally accurate in predicting the number of paired bases but not the structure for mid-length sequences such the 65 base pair cis-regulatory element pictured in Figure 3. It begins to become more and more inaccurate as the sequence lengths become longer, and it appears to be unable to handle pseudoknots, as displayed in Figure 4.

The three main assumptions were made for the sake of the program. First, that unbonded loops can be as small as a single nucleotide. Second, that when a sequence folds the two nucleotides at the point at which the fold takes place have the capacity to bond to each other. Third, that a single unmatched pair of nucleotides embedded within a string of matched pairs is just as bad as any other mismatch.

In expanding this project further I would first allow it to compensate for minor errors, ignoring mismatches flanked on either side by multiple matches. Secondly, I would change it to call the entire function, (the main split, the internal split, and the extension split) for every possible iteration, not just the best fold of the initial string. This may cause it to run much slower than it does now, but it would give more accurate result. Finally, I would create a better way to visualize the secondary structure, possibly by denoting on each larger string where and which smaller folds begin and end, or by creating the actual structure itself within a table.

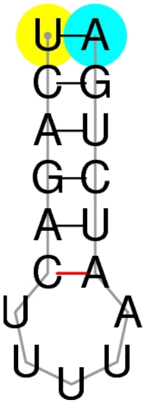

Figure 2. The secondary structure of a tRNA (Molecule ID: PDB_00024). My program recreates this structure with one exception, it does not bond the ‘C’ with the ‘A’ towards the end of the stem, instead bonding the ‘A’ and ‘U’ that begin the loop.

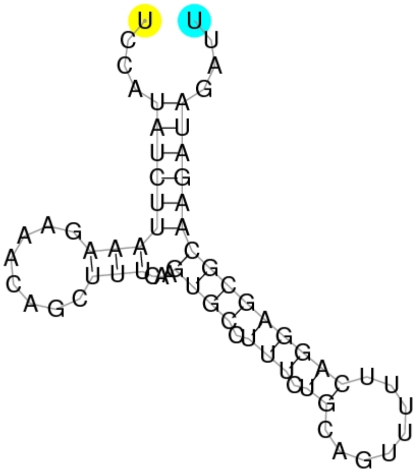

Figure 3. The secondary structure of a cis-regulatory element (Molecule ID: RFA_00640). It contains 36 paired bases, my program predicts 38. The structures are notably different, however.

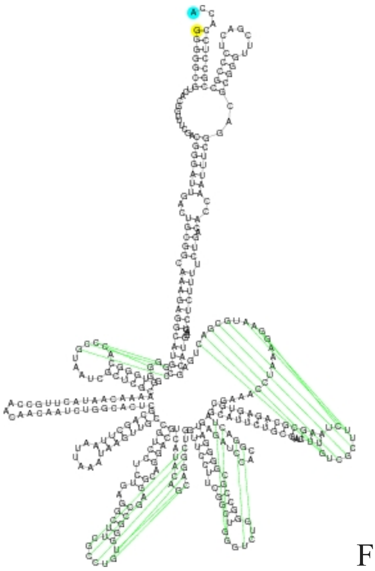

Figure 3.The secondary structure for a Transfer Messenger RNA Molecule ID: TMR_00007). It contains 216 paired bases, my program predicts 180, and the structures are very different.

Bibliography

https://www.qiagen.com/kr/resources/faq?id=06a192c2-e72d-42e8-9b40-3171e1eb4cb8&lang=en

Figure 1. https://en.wikipedia.org/wiki/Transfer_RNA

Figure 2. http://www.rnasoft.ca/strand/show_results.php?molecule_ID=PDB_00024

Figure 3. http://www.rnasoft.ca/strand/show_results.php?molecule_ID=RFA_00640

Figure 4 http://www.rnasoft.ca/strand/show_results.php?molecule_ID=TMR_00007