Idealized RNA Folding

Ibrahim A. Apata

Mathematics and Physics Department, Morehouse College

Introduction

- RNA consists of the following nucleotides: adenine (A), guanine (G), cytosine (C), and uracil (U)
- In RNA folding C bonds with G and A bonds with U
- In an idealized RNA folding there are no unpaired nucleotides
- In this research I ask and answer questions about different RNA sequences

Definitions

- Plane Tree (p.tree)
  - A rooted tree with the vertices of each child having an order.
- Plane Tree Move (Figure 1)
  - Combines two edges on a tree to form a new pair of edges.
- Idealized RNA Folding (IRNAF)
  - A type of bonding in RNA where every nucleotide bonds with one other nucleotide. Additionally, the arcs that are used to represent these bonds do not cross.(example in figure 2)

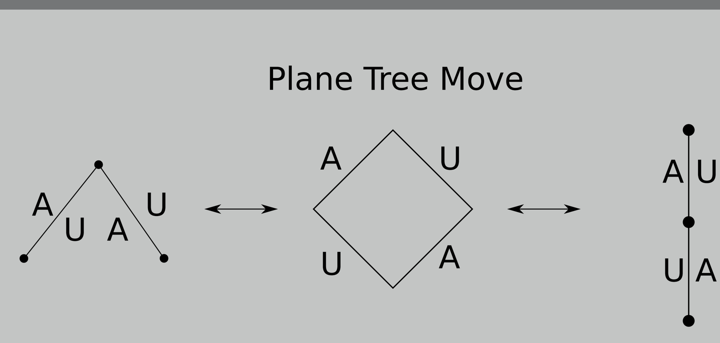

Figure 1: Making a PTM on a p.tree

Questions

1. How many RNA sequences fold into a given plane tree?
2. Given an RNA sequence and a tree, which tree moves are allowed?
3. How many ways can a given RNA sequence of length 2n fold?
4. Which RNA sequences have the most/fewest foldings?

Question 1

- Proposition
  - A given plane tree, with n edges, can be represented by $\mathbf{4^n}$ RNA sequences.
- Proof Sketch
  - A single edge can be represented as one of four bonds. As you add more edges you multiply **4** by the previous number of sequences which is the same as $\mathbf{4^n}$.
- Future Use
  - This could be used for the establishment of a relationship between multiple RNA sequences that fold into the same tree.

Question 2

- Proposition
  - A valid move on RNA p.trees involve either two siblings with the same type of bond or a parent and child, where the child is in an inversed order of the parent.
- Proof Sketch
  - See Figure 1.
- Future Use
  - By knowing which tree moves are allowed I am able to easily calculate how many moves a certain RNA plane tree can make.

Question 3

- Conjecture and Intuition
  - An RNA sequence of length 2n that is not in the forms: $\mathbf{(CG)^n}$, $\mathbf{(GC)^n}$, $\mathbf{(AU)^n}$, and $\mathbf{(UA)^n}$, have at most $\binom{n}{2}$ foldings.
  - The p.trees with the most foldings and n edges can fold into a maximum of $\binom{n}{2}$ trees. Since no other sequences can fold into every tree these sequences are limited certain trees.
- Future Use
  - By knowing the different ways that any given sequence can fold I limit the number of plane trees that sequence can fold into.

Question 4a

- Proposition
  - Some sequences with the fewest foldings have the following properties:
    1. No bonds can occur in the first half of the sequence
    2. The second half of this sequence is a reflection of the nucleotides that bond with the first half.
- Proof Sketch
  - See Figure 2.
- Future Use
  - By knowing some of the sequences with the fewest foldings I can establish a lower bound on the number of foldings of any sequence of length 2n.

Question 4b

- Proposition
  - The RNA sequences with the most foldings are: $\mathbf{(CG)^n}$, $\mathbf{(GC)^n}$, $\mathbf{(AU)^n}$, and $\mathbf{(UA)^n}$. (Ex $\mathbf{(CG)^2}$ means **CGCG**)
- Proof Sketch
  - One of the properties of these sequences is that they can fold into any plane tree with n edges. Additionally, the number of p.trees follow the Catalan numbers as the number of edges (n) increases.
  - Since there is a bijection between plane trees and non-crossing perfect matchings this allows me to conclude that these sequences have the most foldings.
- Future Use
  - By knowing the sequences with the most foldings I can establish and upper bound on the number of foldings of any sequence of length 2n.

P.Tree and IRNAF

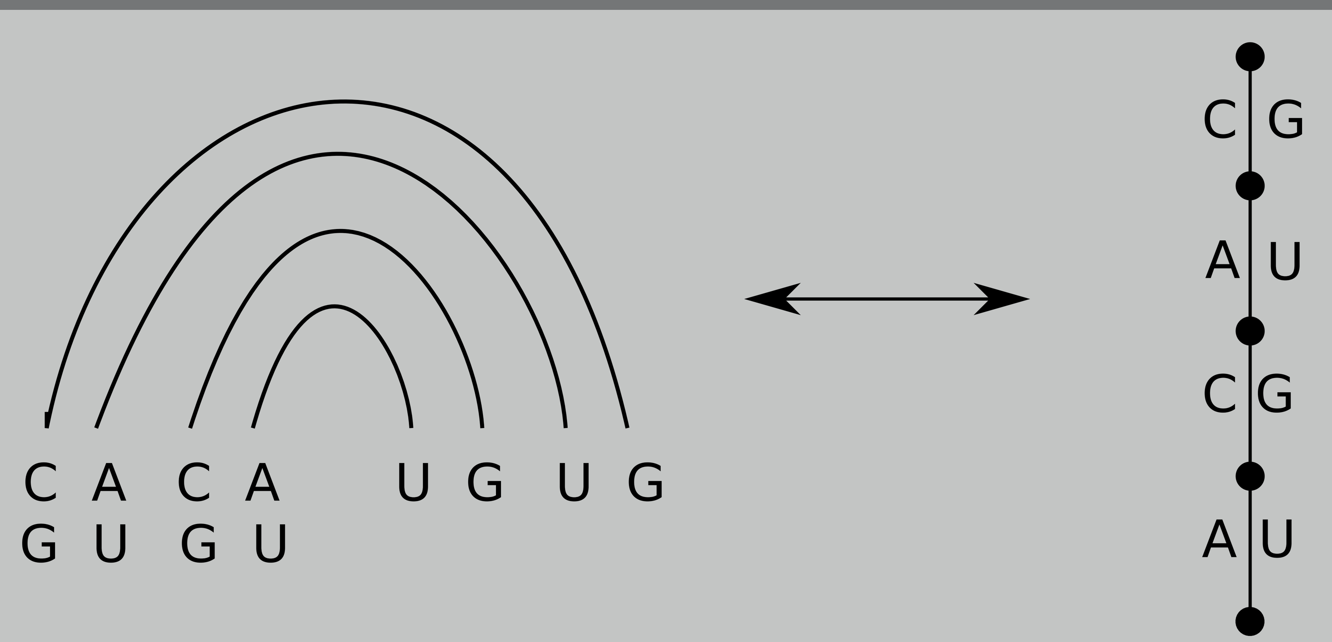

Figure 2: Showing how a IRNAF can be mapped as a p.tree.

Future Questions

- How many 2n letter RNA seqeuences can fold?
- What are the other sequences of RNA that have only one way of folding?

References

1. Keller, M. T., Trotter, W. T. (2014). Applied Combinatorics. Preliminary edn.

Acknowledgments

Thank You, Dr. Heitsch, the graduate students, and postdocs for presenting me with this opportunity and providing me with much support. Additional thanks to Dr. Cooper and Dr. Wilson for recommending me for this program. Finally a thank you to the NSF DMS grant #1344199 and Georgia Tech for making this program possible.