Plane Trees as a Model for RNA Secondary Structures

MCMC Sampling from A Probability Distribution

Chelsea Huston

Spelman College, Department of Mathematics, Atlanta, GA

chuston@scmail.spelman.edu

Spelman College

A Choice to Change the World

Abstract

This project uses samples from a Markov Chain to investigate what a random plane tree looks like under a certain probability distribution. The statistics calculated from the randomly generated plane trees are compared with those from known RNA foldings.

Importance

The folding of RNA secondary structures is the motivation for this project. RNA, formally known as **ribonucleic acid**, is a molecule in cells of all living things. RNA folds to more complicated shapes than DNA, because the single strand bonds with itself and its helices fold onto itself. The **secondary structure** of an RNA molecule is the two-dimensional representation of the base pairings A-U and G-C. The folding of an RNA molecule gives insight to how it functions in the cell. The ability to accurately predict how RNA folds can be useful in understanding different diseases.

Objectives

- Design and implement a Markov Chain Monte Carlo to sample from a probability distribution on plane trees.
- Predict potentially biologically interesting parameters of the energy function graph features to study the resulting plane tree structures.
- Use samples from the MCMC to analyze the features as you vary the parameters.
- Compare with features of known RNA secondary structures.

Plane Trees

A **plane tree** is a rooted tree for which the children of the vertices are linearly ordered. For this project, we use plane trees to model RNA secondary structures. The root represents the start and end of an RNA sequence. A vertex corresponds to a loop and the edge between vertices represent a helix. The degrees of vertices signify the types of loops:

- Hairpins in RNA secondary structures correspond to leaves
- Internal loops are degree 2
- Branches are degree ≥ 3

The number of plane trees with n edges is enumerated by the Catalan numbers, C$_n$. **Ladder distance** is longest path of consecutive edges one leaf to another leaf.

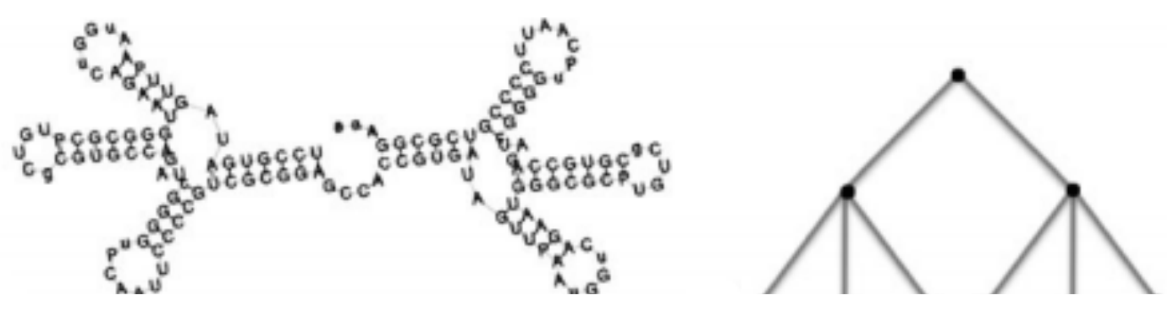

Figure 1: RNA secondary structure as a plane tree.

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC): *Given an irreducible, reversible Markov chain on Ω with transition matrix P and stationary distribution π, a Markov chain on Ω will have stationary distribution μ if the transition matrix $\tilde{P}$ has entries*:

$$\text{when } x \neq y, \tilde{P}(x,y) = P(x,y) \cdot \min\left(1, \frac{\mu(y)\,P(y,x)}{\mu(x)\,P(x,y)}\right)$$

$$\tilde{P}(x,x) = 1 - \sum_{y \in \Omega} \tilde{P}(x,y)$$

.

Methods

In this project we use SAGE, a mathematics software system, to build a Markov chain to sample from a probability distribution μ on plane trees. In order to sample from μ, we use Markov chain Monte Carlo (MCMC) to construct a Markov chain with stationary distribution μ. μ is motivated by the energy functions in the NNTM (nearest neighbor thermodynamic model) for the folding of RNA secondary structures. The energy function used is

$$E(T) = -.4r + 2.3d_0 + 1.3d_1 - .1n,$$

where r is the degree of the root vertex, d_0 is the number of non-root vertices with 0 children, d_1 the number of non-root vertices with 1 child, and n is the number of edges. The stationary distribution μ is defined by

$$\mu(T) = e^{-E(T)}/z,$$

where

$$z = \sum_T e^{-E(T)}.$$

Results

Plasmodium falciparum

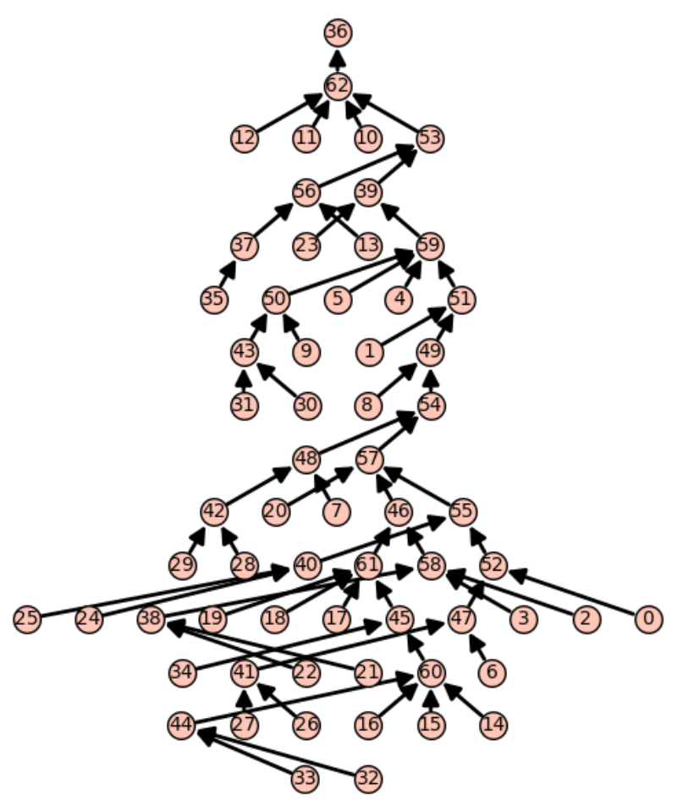

Figure 2: Plane Tree generated from the RNA sequence for Plasmodium Falciparum, a protozoan parasite that causes malaria in humans.

Plasmodium Falciparum Plane Tree Characteristics	
Children of Root	1
Non-Root Vertices with One Child	1
Longest Path	14
Leaves	36
Ladder Distance	15

Figure 3: Statistics computed from the above tree.

Samples from MCMC on 62 Edge Plane Trees

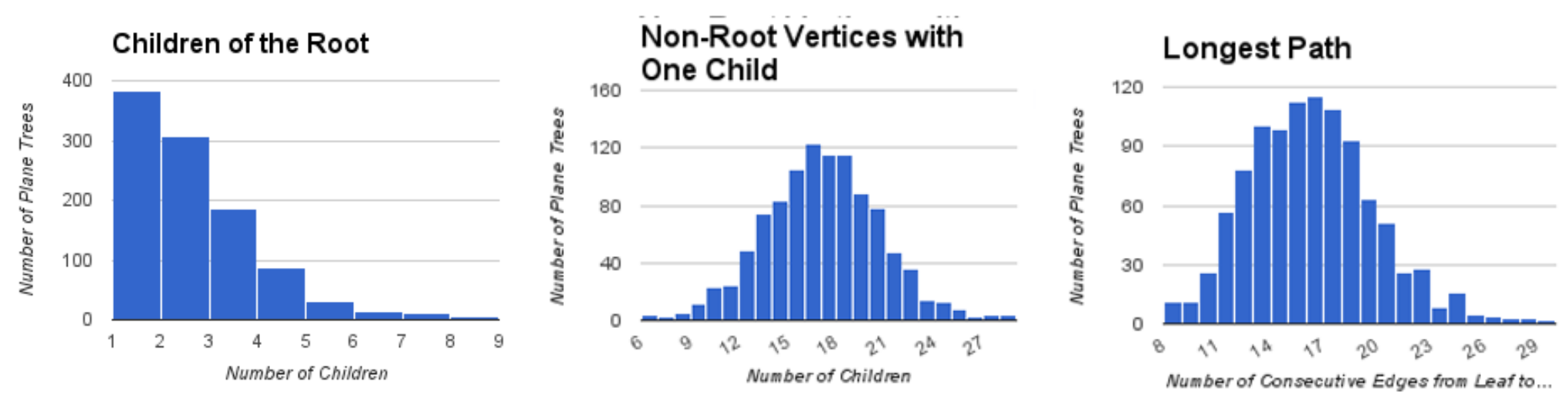

Figure 4: Three charts of the different statistics collected from the randomly sampled plane trees with 62 edges. Children of the root had a mean of 2.148 and variance of 1.572. Non-root vertices with one child had a mean of 16.604 and variance of 11.578. Longest path had a mean of 15.763 and variance of 12.079.

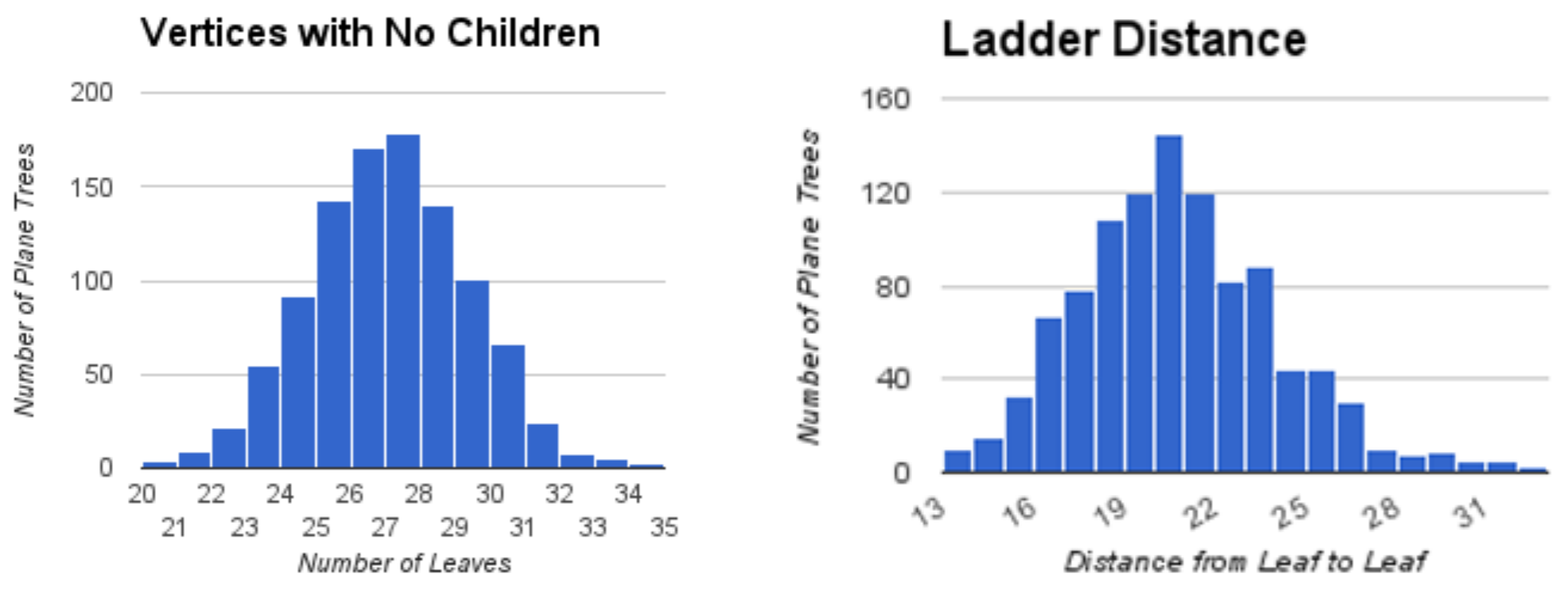

Figure 5: Two charts of the different statistics collected from the randomly sampled plane trees with 62 edges. Vertices with o children had a mean of 26.58 and variance of 5.015. Ladder distance had a mean of 20.246 and variance of 10.524.

Albinaria turrita

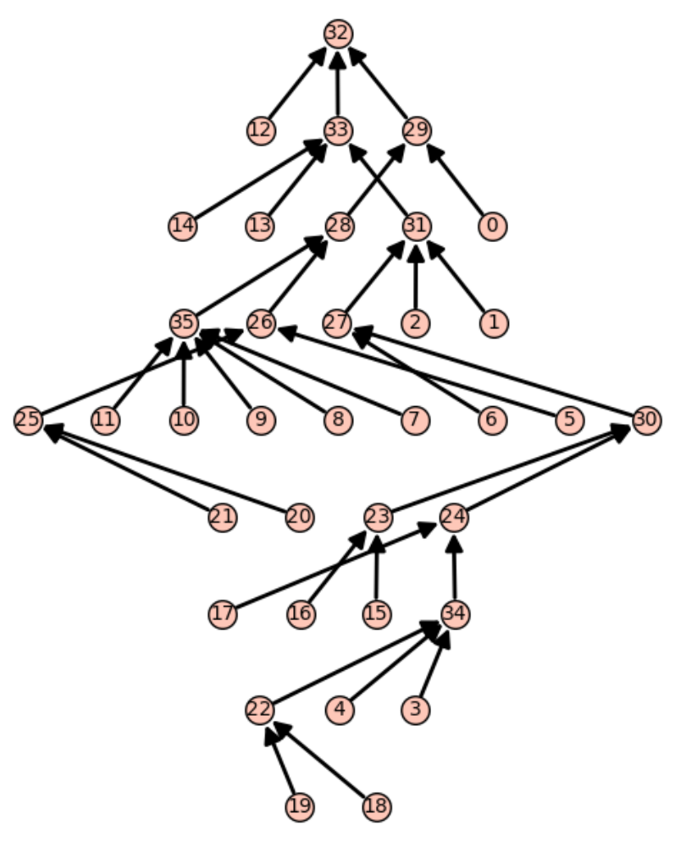

Figure 6: Plane Tree generated from the RNA sequence for Albinaria turrita, Albinaria is a genus of air-breathing land snails (terrestrial pulmonate gastropod mollusks).

Albinaria Turrita Plane Tree Characteristics	
Children of Root	3
Non-Root Vertices with One Child	0
Longest Path	8
Leaves	22
Ladder Distance	13

Figure 7: Statistics computed from the previous tree.

Samples from MCMC on 35 Edge Plane Trees

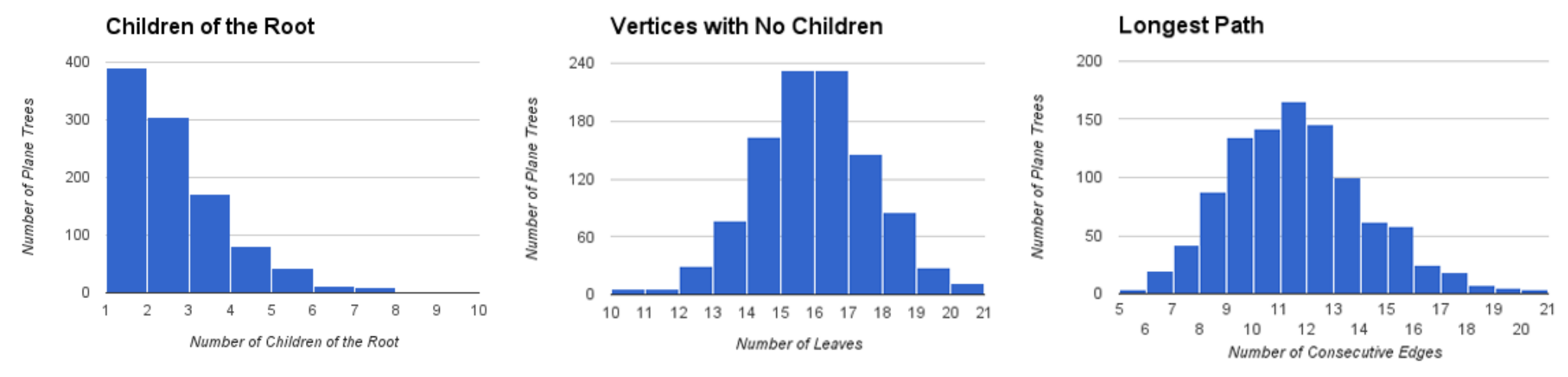

Figure 8: Three charts of the different statistics collected from the randomly sampled plane trees with 35 edges. Children of the root had a mean of 2.154 and variance of 1.666. Non-root vertices with one child had a mean of 9.162 and variance of 6.07. Longest path had a mean of 11.131 and variance of 6.54.

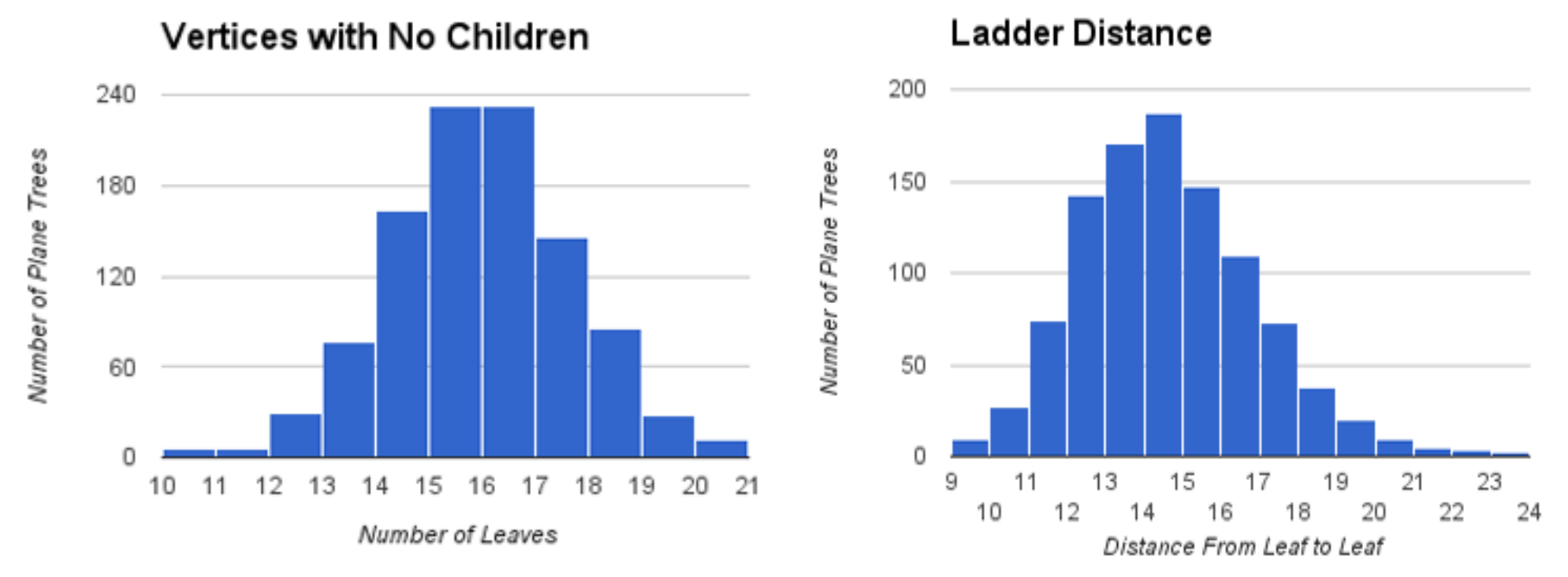

Figure 9: Two charts of the different statistics collected from the randomly sampled plane trees with 35 edges. vertices with no children had a mean of 15.495 and variance of 2.869. Ladder distance had a mean of 14.103 and variance of 4.971.

Conclusions

The statistics computed on the random samples yielded close to normally distributed results. For the most part, the statistics on the plane trees coincides with the actual features. The major exception is that the number of non-root vertices with one child were significantly lower than the average value of the random samples.

Forthcoming Research

In the future we will compare random samples to actual RNA foldings for molecules and look at additional statistics. Different energy functions will be tested in the program that we developed to run MCMC and eventually create an original weighted function to obtain samples that closely resemble biological structures.We will compare the results to the projected outcomes in the papers referenced.

Acknowledgments

I would like to thank Peter Ralli, Anna Kirkpatrick, Emily Rogers, Dr. Megan Bernstein, and Dr. Christine Heitsch for their contributions to my research. Also, I would like to thank Spelman College for nominating me for this REU. This research was supported by the NSF DMS grant #1344199.

References

[1] Yuri Bakhtin and Christine E. Heitsch. Large deviations for random trees and the branching of RNA secondary structures. *Bull. Math. Biol.*, 71(1):84–106, 2009.

[2] Valerie Hower and Christine E. Heitsch. Parametric analysis of RNA branching configurations. *Bull. Math. Biol.*, 73(4):754–776, 2011.