Improving RNA Secondary Structure Prediction Accuracy by Implementing a Hidden Markov Model on SHAPE Data

Jason Kolbush & Taylor Strickland

IMPACT REU

jkolbush3@gatech.edu

tstrickland@agnesscott.edu

Introduction

RNA has risen to the forefront in molecular biology research [1].

The four nucleotides in RNA follow the Watson-Crick pairings of (A) paired to (U) and (C) paired to (G). RNA itself is single stranded, but it has the ability to fold on itself to give many possible structures. Predicting the structure of an RNA sequence is important because its structure determines its function.

- As of now the predominant method for predicting RNA secondary structures uses free energy minimization (MFE) [2].
- Thermodynamic optimization like other methods does not produce high accuracies for secondary structure foldings.

Recent research has shown that the introduction of SHAPE data into prediction programs improves secondary structure prediction considerably [4].

- SHAPE data is a thermodynamic representation of the reactivity of a nucleotide.
- The overlap in distribution of SHAPE data for different nucleotide states creates ambiguity in folding prediction algorithms.
- Alternative methods for incorporating SHAPE data into predicting RNA secondary structures may improve prediction accuracies [4].

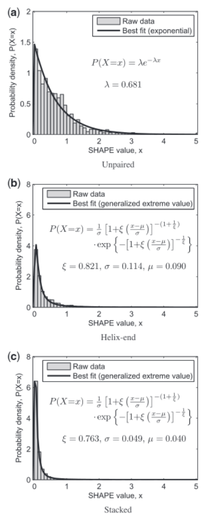

Figure 1: SHAPE Value Probability Density Function[4].

Main Objective

Design a Hidden Markov Model that can be used with experimental SHAPE data to create separation between overlapping SHAPE data distributions. This new SHAPE data can then be used with a MFE folding algorithm to produce more accurate secondary structure predictions.

Structure of HMM

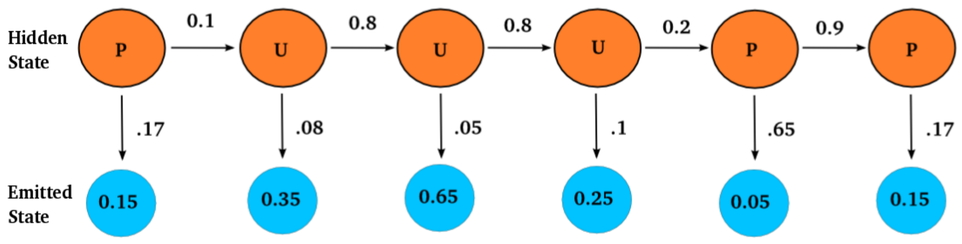

Figure 2: Hidden Markov Model Structure Example

- A Hidden Markov Model has promise to predict nucleotide states correctly because it takes into account not just the value of a given SHAPE value, but the values of its neighbors.
- The chosen Markov Model contains 16 paired states and 16 unpaired states. This allows us to directly take into account the exact probabilities of specific substring lengths occurring.

Methodology

- Analyzed an equally weighted set of tRNA and 5S rRNA in order to gather length distribution for paired and unpaired subsequences and used these distributions to create our transition matrix between our 32 state matrix.
- Discretized SHAPE probability distribution functions found in [4] by subdividing the functions in 0.1 intervals and used this discretization as emission probabilities for our Hidden Markov Model.
- Simulated 1000 sets of SHAPE data for a test tRNA sequence, ran data sets through the Viterbi algorithm of our HMM, observed it's accuracy in correctly predicting nucleotide states.

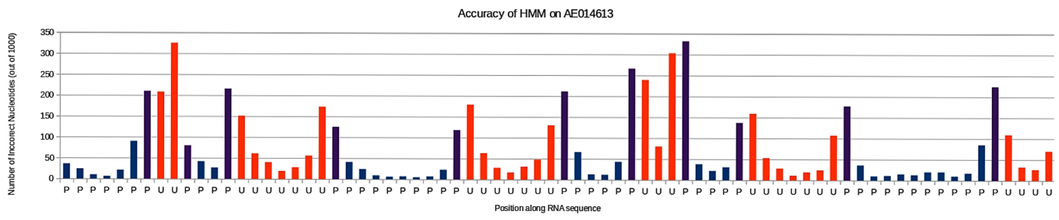

Figure 3: tRNA nucleotide position vs. inaccuracy of predicted state.

- Chose to emit new SHAPE values based on parabolic function. Ran optimization to determine best parameters.
- Ran the Viterbi algorithm of our HMM on simulated SHAPE data from a set of 16s rRNA sequences and assigned new SHAPE values based on optimized parabolic function. Calculated accuracy of folding prediction.

Emitted SHAPE Values: Parabola Function

$$S = a\left(x - \frac{L-1}{2}\right)^2 + M \qquad (1)$$

$$a = \frac{4(E-M)}{(1-L)^2} \qquad (2)$$

Note: When L=1 S=E
M: Min/Max SHAPE Value
S: Emitted SHAPE Value
L: Length of Sequence
E: Edge SHAPE Value
x: Position Along Sequence
a: Horizontal Stretch of Parabola

Results

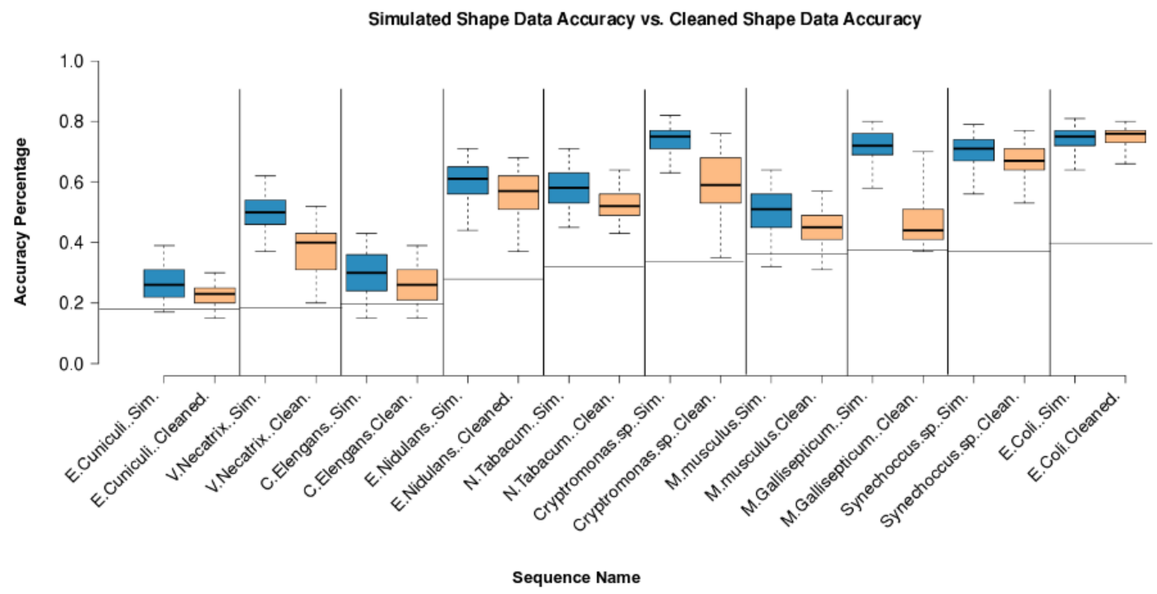

Figure 4: Accuracy ranges for 1000 simulated SHAPE data (blue), accuracy ranges for 1000 simulated "cleaned" SHAPE data (orange), and non-directed sequence accuracy (gray lines).

- For the ten 16s rRNA sequences our HMM was approximately 76% correct.
- Of the ten 16s rRNA sequences tested only one (*E.Coli*) showed improvement.
- Median accuracies for all "cleaned" sequences maintained accuracies above that of non-directed sequences.
- On average the change in median accuracy was -8.139%.

Conclusions & Future Research

- Despite decrease in median accuracy the method of using a HMM and "cleaned" SHAPE values shows promise for improving RNA secondary structure prediction accuracy as shown in (Results).
- Potential methods to improve our model include: changing hidden state structure, including different RNA training sets, and altering the function that assigns "cleaned" SHAPE values.

References

1. Couzin J. Small. (2002) RNAs make big splash. Science, 298, 22962297.
2. Mathews,D.H. (2006) Revolutions in RNA secondary structure prediction. J. Mol. Biol., 359, 52632.
3. Turner, D. H., Sugimoto, N., & Freier, S. M. (1988). RNA structure prediction. Annual review of biophysics and biophysical chemistry, 17(1), 167-192.
4. Zsuzsanna Sksd, M. Shel Swenson, Jrgen Kjems, and Christine E. Heitsch. (2013) Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions Nucleic Acids Res., 41, 2807-2816.

Acknowledgements

We would like to thank our mentor Torin Greenwood for supporting us through our research this Summer, Christine Heitsch for running the IMPACT REU program and allowing us to reference and extend her research, and the NSF grant DMS-1344199 for funding this research.