KnotAli: Informed Energy Minimization Through the Use of Evolutionary Information

by

Mateo Gray
B.Sci., University of Vermont, Vermont, USA, 2019

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science

Supervisory Committee

Dr. Hosna Jabbari, Supervisor
(Department of Computer Science)

Dr. Sean Chester, Co-Supervisor
(Department of Computer Science)

ABSTRACT

Motivation:

Improving the prediction of structures, especially those containing pseudoknots (structures with crossing base pairs) is an ongoing challenge. Current alignment-based prediction algorithms only find the consensus structure, and their alignments can come from structure-based alignment algorithms, which is more reliable, but come with an increased cost compared to sequence-based alignment algorithms. This step can be removed; however, non-alignment based algorithms neglect structural information that can be found within similar sequences.

Results:

We present a new method for prediction of RNA pseudoknotted secondary structures that combines the strengths of MFE prediction and alignment-based methods. KnotAli takes an RNA sequence alignment and uses covariation and thermodynamic energy minimization to predict secondary structures for each individual sequence in the alignment. We compared KnotAli's performance to that of three other alignment-based algorithms, on a large data set of 10 families with pseudoknotted and pseudoknot-free reference structures. We produced sequence alignments for each family using two well-known sequence aligners (MUSCLE and MAFFT). We found KnotAli to be superior in 6 of the 10 families for MUSCLE and 7 of the 10 for MAFFT. We find KnotAli's predictions to be less dependent on alignment quality. In particular, KnotAli is shown to have more accurate predictions compared to other leading methods as alignment quality deteriorates.

Availability:

The algorithm can be found online on Github at `https://github.com/mateog4712/KnotAli`

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

My advisors Hosna Jabbari and Sean Chester for their support and encouragements. Their help allowed me to constantly better myself.

My mother who remained an unyielding emotional support during a hard time.

Chapter 1

Introduction

Understanding RNA structure is essential to understanding its function. RNA plays an active role in many processes that occur within the cell, such as in transcription [7], translation [7, 40], splicing [45, 60], catalysis [7, 63] and regulating gene expression [7, 35, 43, 45]. RNA's function is mainly determined by its structure. As experimental methods are largely expensive for finding these structures, computational methods, and their improvements, have consistently remained relevant.

The majority of computational methods focus on secondary structures — the two dimensional folding of an RNA molecule. Due to similar functions, homologous RNA molecules conserve their common structure. Conservation takes the form of compensatory mutations in response to point mutations that would cause a change in the structure [38, 61]. Compensatory mutations leave a detectable correlation between positions on a multiple sequence alignment — referred to as covariation. Given enough sequences from a related family and an alignment of high structural consistency, comparative sequence analysis (CSA) has been shown to accurately predict secondary structures [32]. Despite the usefulness, circumstances for CSA are limited — homologous sequences and an accurate alignment are not always available especially in cases of novel sequences. A prevalent approach, when such information is not available, is to predict for a single sequence a structure with the minimum free energy (MFE), as structures with minimum free energy are assumed to be the most stable [44]. These algorithms use a set of empirical parameters to define the folding energies of a structure, where every structural feature has been assigned a specific energy value. These parameters are not always accurate or known. In addition, these methods assume that an RNA molecule forms a structure in isolation or with minimal interaction with other molecules. These simplifications may result in discrimination

between predicted structures and structures found in nature.

Current alignment-based methods couple their covariation with another metric for determining structure. RNAalifold [4] uses thermodynamic energy minimization, Hxmatch [64] uses maximum weighted matching (MWM), and Cacofold [50] uses an RNA-based grammar. Despite their coupling, these algorithms still heavily rely on the quality of the alignment to make accurate predictions. These algorithms, in addition, only predict the consensus structure rather than the structures for all sequences. Within alignment-based algorithms, there is an opportunity to address these present shortcomings.

In this work, we present KnotAli, a novel RNA pseudoknotted secondary structure prediction algorithm, which enhances its minimum-free-energy prediction using conserved structural information. Given a sequence alignment of functionally similar RNA molecules, KnotAli finds their individual structures. KnotAli combines two types of information into the prediction. It first uses covariation to find a conserved consensus structure and then uses this consensus structure to guide the minimum free energy prediction for each sequence that makes up the alignment. We consider probable base pairs found in the comparative step separately from the MFE structures found at the end and call them *intermediary pairs*. This grants two qualities: a) the intermediary pairs are not static; and b) the final pairings are thermodynamically informed. We also introduce *restricted bases* and define them as improbable such that their likelihood of pairing with any other base is sufficiently low. We find these bases to be unfavorable toward the final structure.

KnotAli's prediction accuracy was benchmarked against other existing alignment-based prediction algorithms, both pseudoknot-free (RNAalifold [4]) and pseudoknotted (Hxmatch [64], and Cacofold [50]). We find KnotAli to produce predictions which are more robust to alignment quality deterioration and to perform better to a significant degree on the majority of families compared to current algorithms.

1.1 Thesis Objectives

Our motivation in this thesis was to address the lack of evolutionary information in traditional single sequence RNA secondary structure prediction. Alignment-based algorithms utilize this information but solve for the alignment and not for the individual sequences. For these reasons, and for improving the predictive accuracy of these methods, we proposed KnotAli and compared it to some of the existing alignment-

based methods. We focused on algorithms which do not simultaneously solve the alignment and the structure.

1.2 Thesis Contributions

We describe the contributions of our thesis in the current section.

1. We designed and developed KnotAli, a possibly pseudoknotted alignment-based structure prediction algorithm.

2. We discussed an important topic in alignment-based methods — their ability to predict accurately as the alignment quality decreases. We presented multiple examples which explore this topic. From this, we asserted that alignment quality plays a significant role in the predictive ability of alignment-based algorithms.

3. We performed an analysis on our algorithm, KnotAli, against three well-known algorithms (RNAalifold [4],Hxmatch [64], and Cacofold [50]) to measure their accuracies on a large dataset of both pseudoknot-free and pseudoknotted structures. We showed in our analysis that KnotAli has superior accuracy on the majority of our dataset.

4. We compared two different accuracy measures for structure prediction (MCC and F-measure), and we found F-measure to be a superior metric for evaluating structure prediction.

Chapter 2

RNA secondary structure

2.1 Definitions

Definitions We represent an RNA molecule with its sequence, S, and its length n. An RNA sequence is made up of four bases: Adenine (A), Cytosine (C), Guanine (G), and Uracil (U). When referring to an alignment of multiple RNA sequences, in addition to the four bases we sometimes observe a "-" (gap) which holds the position of an insertion/deletion (indel) in the alignment. Note that due to indels an alignment might be longer than the RNA sequences — we denote this length as n_a.

When an RNA sequence forms a structure, its complementary bases pair together and form hydrogen bonds. 'A' pairs with 'U' and 'G' pairs with either 'C' or 'U' — termed *canonical base pairs*. We refer to bases by their position in S. A *base pair* is then defined as the pairing of two bases i and j where $1 \leq i < j \leq n$. We identify base pairing by a "." (dot). We note that each base can pair with maximum one other base (i.e. no base triplets are allowed). In Figure 2.2, we note that the sequence is comprised of 43 bases and each loop signifies a base pairing.

An RNA structure is considered *pseudoknotted* when at least two of its base pairs, $i.j$ and $i'.j'$ cross: $1 \leq i < i' < j < j' \leq n$. Both $i.j$ and $i'.j'$ are considered pseudoknotted base pairs. The example of a pseudoknot shown in Figure 2.1 consists of three base pairs at 17.32, 18.31, and 19.30 crossing the larger stem. All base pairs are pseudoknotted within this example. In contrast, structures without crossing base pairs, are called *pseudoknot-free structures* — see Figure 2.2.

The formation of base pairs within an RNA structure partition the unpaired bases into loops. We define these below. Furthermore, these loops are visualized in Fig-

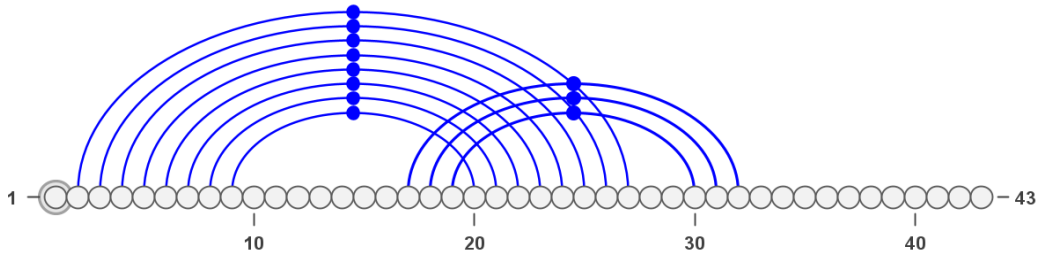


Figure 2.1: An example of a H-type pseudoknotted structure. Base pairs at 17.32, 18.31, and 19.30 cross the larger stem. This figure was made using the VARNA software [8].

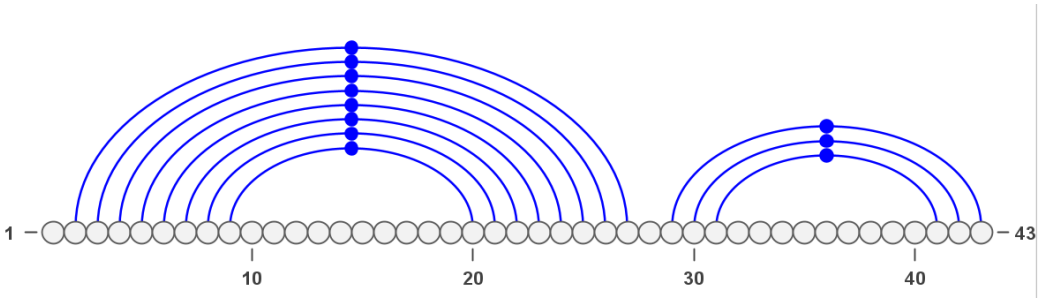


Figure 2.2: An example of a pseudoknot-free structure. We notice that the stems are non-overlapping. This figure was made using the VARNA software [8].

ure 2.3.

- A *hairpin loop* is comprised of one closing base pair $i.j$ where all bases between i and j are unpaired.
- A *stack* is where two consecutive base pairs form. A pair $i.j$ is considered a stack when the following bases $i+1$ and $j-1$ pair. A consecutive number of stacks are then termed a *stem*. We can visualize a stem in Figure 2.3.
- An *internal loop* is comprised of two base pairs $i.j$ and $k.l$ where $i+1 < k < l < j-1$ and bases between $[i+1, k-1]$ and $[l+1, j-1]$ are unpaired. Variations of the internal loop include a bulge loop and a stack (defined above). A *bulge loop* is where there are no unpaired bases on one side of the loop. A stack can similarly be thought of as the case where there are no unpaired bases on either side.
- A *multi-loop* is a region comprised of at least three distinct branches.

Continuing, a *band* is a set of pseudoknotted base pairs, which form a stem or stems, such that the removal of these base pairs leave the structure pseudoknot-free. Within Figure 2.1, base pairs 17.32, 18.31, and 19.30 form a band. Similarly, we termed *spanning a band* as when the exterior base pair of an internal loop or multiloop cross a band.

2.2 Algorithms

We start with a high level definition of how the different types of algorithms work and their complexities. We then move to examples of well-known algorithms in the given area.

2.2.1 Alignment-based algorithms

Alignment based algorithms [4, 64, 50] measure the interdependence of two columns of an alignment in cubic time. The best-known time and space complexity of alignment-based algorithms are $\mathcal{O}(Nn_a^3)$ and $\mathcal{O}(n_a^2)$ where N is the number of sequences in the alignment. They then use this measured interdependence in one of two ways: 1) They merge the interdependence score with the score of the metric they couple with, or 2) They use the measured interdependence to select pairs which they work off of with their coupled metric.

Within alignment-based algorithms, there are two methods of solving the problem: Solving the alignment and the structure at the same time, and solving the structure from a given alignment. Our focus is on the latter; examples, excluding those used in this paper (discussed later) are given for both, however. Our motivation in working on the latter stems from a benefit of separating the alignment and the structure prediction — namely, the union of the two requires confidence solely in the algorithm while separation allows for curation by experts.

LocARNA LocARNA [24, 23, 19] is a sequence-structure aligner which finds the consensus sequence and alignment for a set of unaligned sequences. Consensus structure prediction is accomplished through the use of RNAFold [66] of the ViennaRNA package [22]. Locarna solves the problem in $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ space for pseudoknot-free structures.

Foldalign Foldalign [13] is a sequence-structure aligner which, likewise, finds the alignment and consensus structure from a set of unaligned sequences. It is based

off the Sankoff algorithm [55] and runs, through heuristics, in $\mathcal{O}(n^2 n_a^2 \delta^2)$ time and $\mathcal{O}(n^2\delta^2)$ space where δ is the maximum distance between two subsequences being aligned. Foldalign predicts the pseudoknot-free structure

KnetFold KnetFold [5] is a consensus structure prediction algorithm which returns a possibly pseudoknotted structure for a given alignment. This is accomplished through a hierarchical network of k-nearest neighbor classifiers as well as an implementation of a RNAFold-based [34] consensus probability matrix. It's time and space complexity being $\mathcal{O}(Nn^3)$ and $\mathcal{O}(n_a^2)$.

2.2.2 Thermodynamics-based algorithms

Thermodynamics-based algorithms [15, 36, 49] find the structure with the minimum free energy for an individual sequence using dynamic programming. Every substructure is assigned an empirically tested energy, and the energy of a structure is the sum of the energies for each substructure. Consequently, one selects, from the set of all possible structures, the structure whose sum minimizes the free energy. For pseudoknot-free structure prediction, the standard time and space complexity is $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$.

MFE pseudoknotted structure prediction is found to be NP-hard [1, 42] and inapproximable [57]. Polynomial-time algorithms require limiting the class of pseudoknotted structures as time complexity is traded off with generality [44]. The most general thermodynamics-based algorithm is PKnots [49] but it comes with a prohibitively expensive time and space complexity of $\mathcal{O}(n^6)$ and $\mathcal{O}(n^4)$. While pseudoknot-free MFE-based prediction is sufficient for a subset of RNA, especially smaller molecules, the biological importance of pseudoknots [26, 59] within the remainder of RNA gives cause for creating algorithms which can predict them.

We focus on thermodynamics-based algorithms which find the MFE structure for a given sequence. Examples are given for this type of algorithm.

Pknots Pknots [49] is a possibly psuedoknotted MFE-based structure prediction algorithm. Pknots covers a broad set of pseudoknots and find the MFE structure in $\mathcal{O}(n^6)$ and $\mathcal{O}(n^4)$ time.

Mfold Mfold [66] is an MFE-based structure prediction algorithm which returns a pseudoknot-free structure for a given sequence. It accomplishes this problem in $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space.

Iterative HFold Iterative HFold [36] is an MFE-based possibly pseudoknotted prediction algorithm that returns a structure in $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space for a

given structure. It differs from previous examples in that it requires a pseudoknot-free approximation of the structure for the given sequence.

2.2.3 KnotAli

KnotAli, using the coupling of covariation and thermodynamics, is capable of finding possibly pseudoknotted structures in $\mathcal{O}(Nn^3)$ time and $\mathcal{O}(n^2)$ space. KnotAli does not cover all topologies of structures but does cover many of the important types of pseudoknots, i.e. kissing hairpins [27] and H-type pseudoknots [2] with arbitrarily nested substructures. Specifically, KnotAli covers all density-2 structures (no base is enclosed by more than two pseudoknotted stems and the structure can be partitioned into two pseudoknot-free structures). This trade-off allows KnotAli, however, to scale to much larger sequences.

2.3 Energy Model

Computational methods for the prediction of secondary structures, namely those which use dynamic programming use a set of parameters to model the free energy of the substructures which make them up. These sets of free energy parameters are called an *energy model*. Energy values are determined experimentally and incorporated into the the thermodynamic portion of the algorithm. The free energy of a loop is dependent on the temperature of the environment when folding. For the energy parameters listed, the temperature is assumed to be 37°C.

2.3.1 Notation

We give the notation for the parameters which make up the energy model. Values are first given for the parameters used in a pseudoknot-free structure and then move to those within a pseudoknotted.

Name	Description	Energy ($KCal/mol$)
$e_H(i,j)$	Energy of a hairpin closed by $i.j$	
$e_S(i,j)$	Energy of a stack closed by $i.j$	
$e_{stP}(i,j)$	Energy of a stack that spans a band	$.89 \cdot e_S(i,j)$
$e_{int}(i,k,l,j)$	Energy of a pseudoknot-free internal loop	
$e_{intP}(i,k,l,j)$	Energy of a internal loop that spans a band	$.74 \cdot e_{int}(i,k,l,j)$
P_s	Exterior pseudoloop initiation penalty	-1.38
P_{sm}	Penalty for initiation of pseudoloop in a multiloop	10.07
P_{sp}	Penalty for initiation of pseudoloop in a pseudoloop	15.0
P_b	Penalty for initiating a band	2.46
P_{up}	Penalty for unpaired base of a pseudoloop	.06
P_{ps}	Penalty for closed subregion	.96
a	Penalty for initiation of a multiloop	3.36
b	multiloop base pair penalty	.03
c	Penalty for unpaired base of a multiloop	.02
a'	Penalty for initiation of a multiloop that spans a band	3.41
b'	Branch penalty in a multiloop that spans a band	.56
c'	Penalty for unpaired base in a multiloop that spans a band	.12

Table 2.1: Free energy parameters used within the thermodynamic portion of our algorithm. The table shows the name of the parameter, its description, and its corresponding energy value. The values are taken from the parameters from HotKnots V2 [3]. All values were found at a temperature of 37°C and with 1 M salt concentration

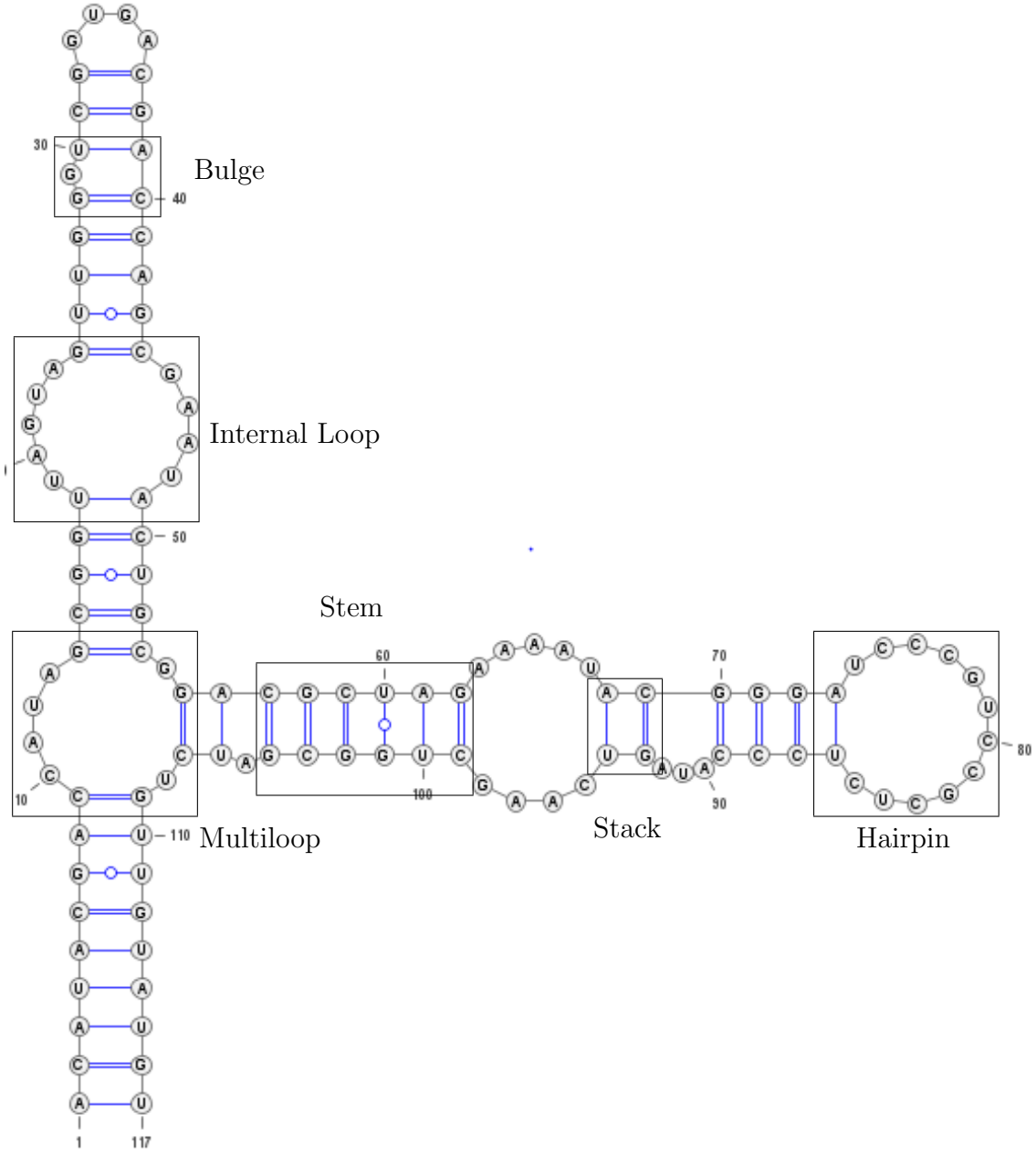
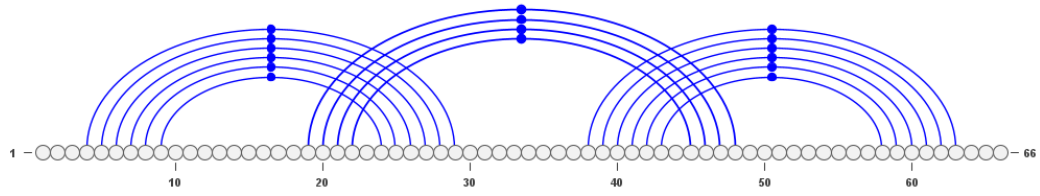
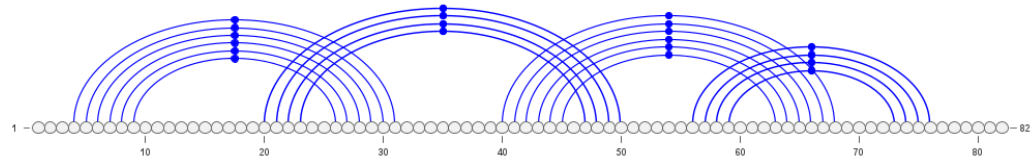


Figure 2.3: A pseudoknot-free structure containing all basic substructures.



(a) A kissing hairpin



(b) A chain of interleaving stems

Figure 2.4: Arc representation of the pseudoknot substructures covered by KnotAli: (a) A kissing hairpin loop, (b) A chain of interleaving stems. An H-type psuedoknot can be visualized in Figure 2.1.

Chapter 3

Materials and Methods

A description of our algorithm is given in Section 3.1. We demonstrate KnotAli's ability to work on a set of families containing both pseudoknotted and pseudoknot-free structures. We look first at the description of our algorithm and end on a description of currently available algorithms.

3.1 KnotAli Algorithm

We defined *intermediary pairs* as the set of non-static pairs found through an alignment. Precise definition of methods for identifying intermediary pairs is given in Section 3.2.2. Here, we also define the *Intermediary structure* to be the pseudoknot-free non-static structure made up of all intermediary pairs.

We first provide a high-level view of the concepts used in our algorithm and later define them in Section 3.2. KnotAli finds the mean mutual information for the alignment and each column by calculating the mutual information (`MI`) between every two column. KnotAli uses the column and alignment means to find the Average Product correction (`APC`) for each pair of bases tested. Non-conflicting base pairs with high mutual information after subtracting the average product correction are selected as intermediary pairs to be used in the final step. Alongside this, columns whose max mutual information is less than the alignment mean are restricted to not being considered for pairing.

The restriction of bases, corresponding to columns of the alignment where the score is low, was used to control base pairing within the algorithm. We denote that within our algorithm a '_' character is used to signify a base that is free/available to

pair with another freely available base, and '.' is used to signify bases that cannot form a base pair. By using the max of each column, found during the previous step, and comparing this value to the mean for the alignment, a measure was made on the probability of that base forming a pair in the following part. When the column max is lower than the mean for the alignment, the character is changed from a '_' to a '.' as bases are considered unpaired when denoted by '.'.

To find the individual structure for each sequence, the intermediary structure, at each iteration, is compared to the gapped sequence. Pairing between a base and a gap, a gap and a gap, or bases that do not form canonical base pairs are reverted to unpaired bases. All bases corresponding to gaps in the sequence are then removed, and hairpins with a size < 3 following the removal of gaps have the inner pair reverted to unpaired bases. This follows the evidence that base pairs with < 3 bases are unlikely to form [21, 31].

The resultant structure is given as guide to the thermodynamic portion of the algorithm. KnotAli follows the relaxed hierarchical folding hypothesis [36], in which base pairs of the input structure may be partially unfolded to allow more thermodynamically stable base pair formations. We provide a general overview of the recursions in the thermodynamic portion in Figure 3.1. For predicting the structure of region $[i, j]$ of the sequence S, the pairing is calculated as $W(i, j)$. $W(i, j)$ is then decomposed as four cases. The first case corresponds to the scenario where j is unpaired thus shifting to $W(i, j-1)$. The second case is where the bases at positions i and j pair to form a loop and $W(i+1, j-1)$ is then considered for the bases within. The third case is where the region within $[i, j]$ can be decomposed into two non-overlapping substructures — one in $[i, k-1]$ and one in $[k, j]$. The final case is where $i.j$ forms a pseudoknot due to the presence of pairs from the intermediary structure on the top plane.

3.2 Intermediary Structure

3.2.1 MI

Mutual Information or `MI` is the reduction in uncertainty of one position given another. It can be thought of as a measure of mutual dependence between two columns in an alignment. Measured in bits, the range of `MI` is between 0 and 2 where 0 suggests that there is no detectable dependency between the two positions and 2 suggests

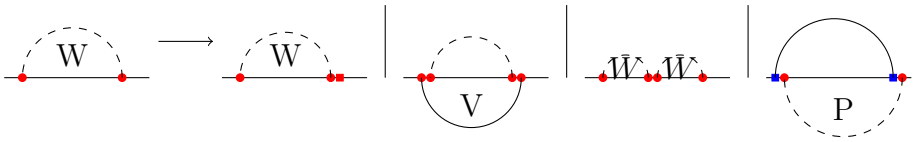


Figure 3.1: Illustration of Recursions within KnotAli. We have divided the structure into two planes, top and bottom, to distinguish between pairs from the intermediary structure (top) and those being predicted in the thermodynamic step (bottom). Likewise, blue is used to denote these pairs on the top, and a solid arc is used to distinguish a fixed pairing. Red is used to denote pairs predicted on the bottom plane. A square is used to show a previous base while a circle shows the current pair being observed. Similar to the top plane, a solid arc distinguishes a fixed pairing while a dashed arc is unfixed.

a high dependency. Due to the effect of compensatory mutations, positions with conserved base pairings have a higher dependency on each other than positions which are independent. Using this fact, `MI` can be used to find these conserved base pairings.

Our mutual information calculation is adapted from the MIToolbox [48]. In a standard mutual information calculation, 4 bases and a gap would allow for 25 possible pairs. Only 6 of these pairs form valid base pairs — either Watson-Crick pairs ('A' with 'U' or 'G' with 'C') or Wobble pairs ('G' with 'U'). When calculating the measure, we ignore non-valid pairs. Let $f_{a,b}(x, y)$ denote the joint frequency of bases x, y at column a, b; similarly, let $f_a(x)$ denotes the frequency of base x at column a and $f_b(y)$, the frequency of base y at column b. We thus define the *mutual information* between column a and column b of an alignment, denoted $MI(a, b)$, as follows:

$$MI(a,b) = \sum_{x,y \in \{A,C,G,U\}} f_{a,b}(x,y) \cdot \log_2 \left(\frac{f_{a,b}(x,y)}{f_a(x) \cdot f_b(y)} \right) \tag{3.1}$$

3.2.2 MIp

Similar to `MI`, `MIp` is the reduction of uncertainty of one position given another but takes into account the effect of noise. While `MI` works well at finding column interdependence in an alignment, it suffers from noise due to random and phylogenetic sources. Moreover, the reduction of noise has been shown to improve measures of covariation [41]. Average Product Correction [56] is a measure that was previously applied to remove background noise in protein structure prediction.

We define $APC(a, b)$, at column a and b, in Equation 3.2. Its components are

defined in Equation 3.3 and 3.4.

$$APC(a,b) = \frac{MI(a,\bar{z}) \cdot MI(b,\bar{z})}{MI_{\text{avg}}} \tag{3.2}$$

$$MI(a,\bar{z}) = \frac{1}{n-1} \sum_{z=0}^{n-1} \begin{cases} MI(a,z) & \text{where } |a-z| > 3 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

$$MI_{\text{avg}} = \frac{2}{n(n-1)} \sum_{w=0}^{n-1} \sum_{z=0}^{n-1} \begin{cases} MI(w,z) & \text{where } |w-z| > 3 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Mutual Information with this background correction is termed `MIp` and is the difference between `MI` and `APC`:

$$MIp(a,b) = MI(a,b) - APC(a,b) \tag{3.5}$$

`MIp` was found to be sensitive and selective compared to the original `MI` [56].

To determine at what point the `MIp` score demonstrates enough interdependence to detect a base pair correctly, we performed a grid search.

By comparing the F-measure for each family at every threshold, a threshold t was picked which maximized the F-measure across all groups. All pairings with `MIp` $> t$ are compiled into a vector. As pairings with repeated positions are possible, i.e. i.j and i.j', the pairings are sorted by score. These are then applied to an empty structure of length n. Pairings which form pseudoknots with the current structure or whose base is already part of a pairing are skipped.

We build an example for our algorithm in Figure 3.2.2. Five sequences are shown from the full alignment of tRNA sequences. Arcs are made between columns of the alignment to simulate the `MIp` calculations made during the comparison step. Each arc contains its corresponding score and arcs extending from the same column represent how higher scores translate to the chosen intermediary structure. We do not show all of the arcs but, instead, a subset of the possible arcs predicted. We also show below the determined intermediary structure for the alignment. Pairs are chosen from non-intersecting arcs which give the highest scores for their columns.

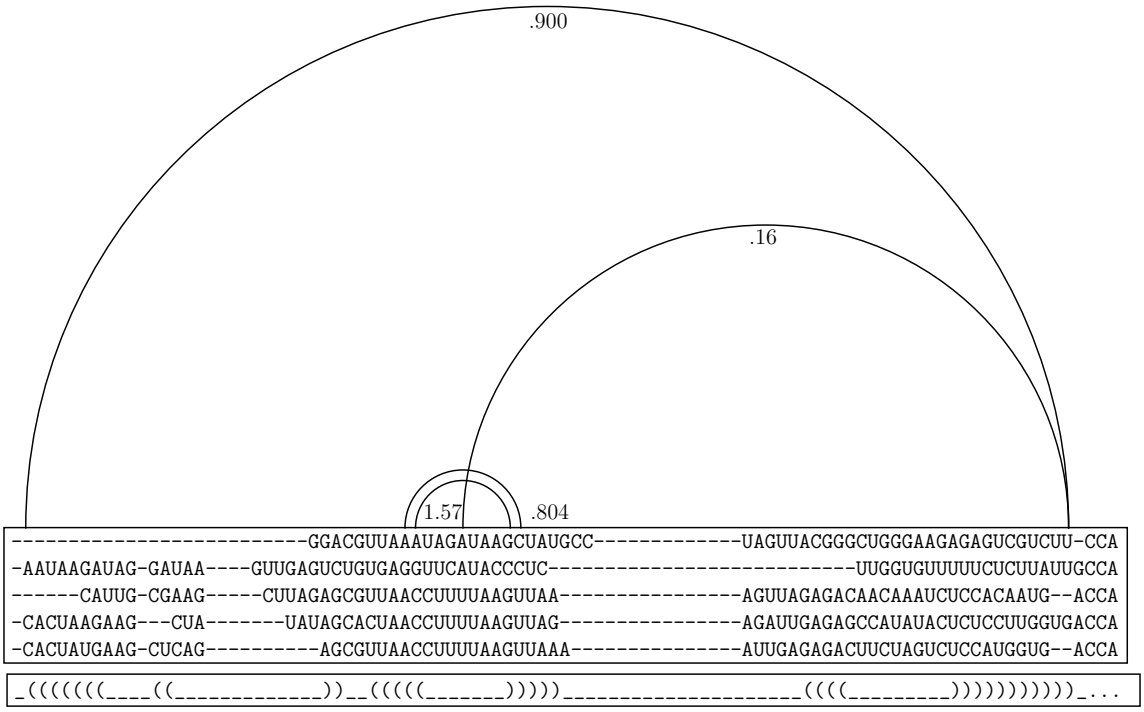


Figure 3.2: Outline of the process for determining the intermediary structure from an alignment. The alignment is built from 462 tRNA sequences. In this example, the first five are shown. Arcs represent the MIp calculations between columns, where higher scores confer a more likely pairing. Pairs are ordered by their score and chosen for the intermediary structure. Base pairs cannot intersect thus pairs which would intersect with previously chosen base pairs are skipped. Pairs which contain a column which have already paired with another column are also skipped. We show a subset of the arcs corresponding to the structure.

3.3 Compared Algorithms

KnotAli was compared to three algorithms: RNAalifold [4], Hxmatch [64], and Cacofold [50]. Similar to KnotAli, both Hxmatch and Cacofold are alignment-based algorithms which return possibly pseudoknotted structures. RNAalifold is restricted to pseudoknot-free structures. We chose these three algorithms as they do not find the alignment with the structure, are not limited in their number of sequences, and due to their similarity to KnotAli.

3.3.1 RNAalifold

RNAalifold is a pseudoknot-free consensus structure prediction algorithm which takes an alignment as an input. There are two versions of covariation measures that RNAalifold uses within its algorithm which can be chosen by the user when running. The first, which appears in the base version, is the use of Hamming distance as a means of distinguishing possible pairings. RNAalifold's Hamming distance is defined as

$$h(i,j) = \begin{cases} 1 \texttt{ if i} \neq \texttt{j and i.j} \in \texttt{ \{AU,CG,GC,GU,UA,UG\}} \\ \texttt{0 otherwise} \end{cases} \tag{3.6}$$

This scoring evaluates the alignment's columns by looking at all i.j within columns a,b using the Hamming distance and then penalizing based on the number of non-pairable $i.j$'s.

The second covariation measure, replacing the Hamming distance, is RIBOSUM matrices. The score is defined as

$$R(i.j, i'.j') = \log\left(\frac{f(i.j, i'.j')}{f(i.i') \cdot f(j.j')}\right) \tag{3.7}$$

where $f(ac)$ is defined as frequency of nucleotide a and c being aligned, and $f(i.j; i'.j')$ is the frequency of base pair i.j and i'.j' being aligned. The full covariance term for columns a,b is then $\frac{1}{2}\sum_{\substack{a,b\epsilon A \\ a\neq b}} xR(a_i.a_j, b'_i.b'_j)$ The Hamming distance and RIBOSUM score are treated differently moving into the final thermodynamic step of RNAalifold. For Hamming distance, the covariation step accounts for about 7% of the total score for determining base pairs. In contrast RIBOSUM scores account for 44% of the score. The remaining 93% and 56%, respectively, comes from the thermodynamic energy minimization.

3.3.2 Hxmatch

Hxmatch is a possibly pseudoknotted alignment based consensus structure prediction algorithm. Hxmatch starts by defining a base pair scoring method which combines a *helix score* and a *covariation score*. The helix score considers all possible base pairs for each sequence in the alignment and calculates the energy of the largest helix containing the base pair. This is done for each position a,b of the alignment. The value is multiplied by -1 to make it positive and placed in a scoring matrix H. The covariation score at position a,b is

$$C_{a,b} = \sum_{i.j,i'.j'} f_{a,b}(i.j) \cdot D_{i.j,i'.j'} \cdot f_{a,b}(i'.j') \tag{3.8}$$

where $f(i.j)$ is the frequency of base pair $i.j$ and $D_{i.j,i'.j'}$ is 0 when $i.j = i'.j'$ or either pair is an invalid pairing, 1 when a single base is different and 2 if both bases are different. A penalty is applied to this score based on the number of invalid pairings at position a,b $\cdot$ a scaling value. The helix and covariation score are then combined into a matrix

$$\pi_{a,b} = H_{a,b} + \phi_{a,b} \cdot C_{a,b} \tag{3.9}$$

with $\phi_{a,b}$ corresponding to another scaling value.

A *maximum weighted matching* (MWM) approach uses the base pair scores found before and builds a set of vertices and edges where vertices are positions 1 to n and the edges are all pairings with a $score > 0$. This step finds the matching which maximizes the sum of edge weights.

3.3.3 Cacofold

Cacofold is a possibly pseudoknotted alignment-based method which uses probabilistic folding methods and positive and negative covariation scores to find a consensus structure. Cacofold is part of the package R-scape [52, 53, 50]. An *E-value* is an expectation value signifying the expected number of false positives [52]. E is defined as $E = NP(score > X)$ where N is the number of column pairs looked at and $P(\text{score} > x)$ is the probability that the column pair would give a covariation score greater than x.

Covariation power is an estimate of the expected ability to detect covariations [50, 53]. Covariation power is used to distinguish when a lack of structure is due to low

sequence variation rather than low covariation. Covariation power and E-value are used in tandem to distinguish what are called positive and negative base pairs. A *positive base pair* is a base pair which reports high covariation (a low e-value). In contrast, a *negative base pair* is a base pair which reports low covariation but high covariation power. Negative base pairs are forbidden to appear in the final structure.

Cacofold groups positive base pairs into nested subsets. The first subset is made up of the maximal number of positive pairings such that there are no crossing base pairs or triplets, and succeeding sets are made up of the remaining positive base pairs. The subsets are used as constraints for the folding algorithms. RNA basic grammar [54] is used on the first subset to find the main nested structure. Later subsets use a simplified grammar called G6X, an extension of the G6 model [39, 10], and are used to find additional helices. The structures formed from each subset are combined after filtering out redundancies without covariation support.

Chapter 4

Experiment Design

4.1 Dataset

All algorithms were tested on ten (pseudoknotted and pseudoknot-free) RNA families whose reference structures were previously determined by comparative sequence analysis [16]. Table 4.1 summarizes these families. The pseudoknot-free families are made up of 5s, SRP, Group II Intron, and tRNA while the remaining families contain at least one sequence whose structure is pseudoknotted. Sequences from families in Table 4.1 were compiled from [58] where duplicates were removed. Sequence lengths vary between 28 nucleotides long at the minimum and 2868 nucleotides at the maximum. We find that this dataset represents a wide degree of conservation ranging from tRNA which is highly conserved [47] to less conserved families such as Group I Intron and Group II Intron which lack sequence conservation [9, 46]. We modify the families to remove all hairpins with size < 3 [21, 31]. This removal occurred within the SRP family.

4.2 RNA Sequence Aligners

To look at the structural similarities within differently-sized sequences, the sequences first have to be aligned and gaps placed such that they all have the same size. The strength of the sequence aligner therefore plays a fundamental role in aligning such that their structures align. In a previous benchmark study [65], 10 different aligners were evaluated. The study sought to score the alignments generated by evaluating the consistency of the secondary structure to the aligned reference sequences. These

Family	# of Sequences	Sequence length	Conservation
5s	1053	103-135	High [17]
16s	22	950-1995	Medium [18, 20]
23s	5	2904-2968	Medium [12]
Group I Intron	89	210-736	Low [46]
Group II Intron	11	619-780	Low [9]
RNaseP	410	120-486	Medium [33]
SRP	583	28-533	Medium [14]
telomerase	37	382-559	Low [25]
tmRNA	363	102-437	Medium [67]
tRNA	461	57-93	High [47]

Table 4.1: List of families with their sequence conservation level, corresponding number of sequences, and range of length

10 aligners either predict solely off of sequence similarity or by combining it with structure prediction. Of these 10 aligners, we chose MUSCLE [11] and MAFFT [37]. Within our benchmarking, we sought to minimize the effect of aligners on the results presented in this work. Within this group, MUSCLE and MAFFT were the best performing aligners that progressively predict a structure and tune the alignment to the structure.

MUSCLE tends to reduce the number of gaps within the alignment, whereas MAFFT tends to add an increased amount of gaps, especially in instances where there is higher variation within the alignment. Both programs only require a FASTA file as an input. No additional parameters were used to affect the alignment.

4.3 Accuracy Measures

The number of *true positives* (TP) is defined as the number of correctly predicted base pairings within the structure. The number of *false positives* (FP), similarly, is the number of predicted base pairs that do not exist in the reference structure. Any base missed in the prediction that corresponds to a pairing in the reference structure is a *false negative* (FN).

We evaluate the performance of algorithms based on three measures: sensitivity, positive predictive value (PPV), and their harmonic mean (F-measure).

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.1}$$

$$PPV = \frac{TP}{TP + FP} \tag{4.2}$$

$$F_{measure} = \frac{2 \cdot PPV \cdot Sensitivity}{PPV + Sensitivity} \tag{4.3}$$

These unitless measures range between 0 and 1. When the predicted structure is the same as the reference structure, the value of F-measure, as well as PPV and sensitivity, is 1. In contrast, when PPV and/or sensitivity is 0, there are no common base pairs between the reference and predicted structure and F-measure is 0. High PPV describes an algorithm which predicts a small number of false base pairs relative to its total predicted. This signifies an ability to differentiate positives from false positives. In contrast, high sensitivity shows an algorithm's ability to overall find base pairs from a sequence. Algorithms seek to maximize both. Therefore, combining both sensitivity and PPV helps to better describe the different strengths of algorithms.

4.4 Bootstrapping and Permutation Test

A bootstrap confidence interval is used to show the dependence of the measured results on the set of RNA from the family. This is performed in the following way. First, we take the list of F-measures for our algorithm on a specific family. Next, 10^4 re-samplings are taken with replacement. A 95% confidence interval is constructed by placing the 2.5^{th} and 97.5^{th} percentiles of the ranked differences as the boundaries of the confidence interval. From this, we can assess that we are 95% certain that the value of the mean is within the interval.

We consider the performance of an algorithm to be superior or inferior to another one if the difference in their accuracy is considered significant based on a two sided permutation test. We define f_1 and f_2 to be the vectors of F-measures obtained by algorithms 1 and 2, and $\bar{f}_1$ and $\bar{f}_2$ to be the mean of the F-measures of algorithm 1 and 2, respectively. We term our test statistic $t_s = \bar{f}_1 - \bar{f}_2$. We permute the vectors f_1 and f_2 in 10^4 different ways where the size of the groups remain the same as the size of f_1 and f_2. For each permutation, we recalculate the difference of means between these permuted groups and compare it to t_s. The p-value is then the proportion of these values whose difference $\geq t_s$. A p-value of less than .05 is used to reject the null hypothesis, where the null hypothesis is that the two algorithms have the same mean performance and that the variance is due to statistical randomness. This was

accomplished using the 'boot' and 'perm' packages in R.

4.5 Configuration

The default settings of Hxmatch and Cacofold were used when testing. RNAalifold uses the '-r' setting indicating the use of RIBOSUM matrices as the covariation metric instead of Hamming distance. We show the result of switching RNAalifold's covariation measure from Hamming distance to RIBOSUM matrices in Table 4.2 on our dataset. We find that for the majority of families, a switch to RIBOSUM matrices show a significant improvement in the score.

	MUSCLE						MAFFT					
	RNAalifold			Riboalifold			RNAalifold			Riboalifold		
family	sen	ppv	F	sen	ppv	F	sen	ppv	F	sen	ppv	F
5s	.561	.929	.698	.761	.884	**.817**	.419	.979	.586	.644	.843	**.729**
16s	.323	.818	.462	.505	.748	**.602**	.336	.852	.481	.548	.794	**.647**
23s	.805	.815	**.810**	.798	.767	.782	.807	.806	**.807**	.895	.771	.783
Group I Intron	0	0	0	.047	.588	**.087**	.0	0	0	.036	.719	**.068**
Group II Intron	0	0	0	0	0	0	.105	.992	.189	.106	.773	.186
RNaseP	0	0	0	.235	.602	**.334**	0	0	0	.206	.758	**.321**
SRP	.124	.897	.198	.165	.764	**.241**	.079	.883	.135	.078	.661	.129
telomerase	.223	.793	.348	.508	.636	**.563**	.225	.869	**.356**	.25	.361	.294
tmRNA	.147	.978	.254	.255	.803	**.386**	.123	.955	.216	.196	.799	**.313**
tRNA	.857	.973	.909	.931	.972	**.949**	.757	.930	.829	.800	.931	**.855***

Table 4.2: Comparison of RNAalifold with and without the use of RIBOSUM matrices as a covariation measure. Riboalifold is used to denote RNAalifold with RIBOSUM matrices. Results are shown across both MUSCLE and MAFFT. **BOLD** is used to show if there was a significant difference in the results. An * is added when the the significance is close enough to warrant distinction but not fully crossing a p-value of .05

We note the difference in output between our algorithm and the others — namely individual structures versus a consensus structure. For comparison, the consensus structure, from the other three algorithms, is applied to all individual structures. Similar to KnotAli, non-canonical base pairs are removed. Loops of size < 3 after the removal of gaps have inner pairs reverted to unpaired bases until the minimum loop size is satisfied. Single base pairs are not removed within this step.

Chapter 5

Results

Preliminary We begin by evaluating how well `APC` increases the F-measures relative to just `MI`. Figure 5.1 illustrate the results on our dataset. We compare the two measures within KnotAli while keeping the thermodynamic portion the same.

Figure 5.2b shows a predicted structure for a member of the tRNA family. We note, unlike a true tRNA structure, the absence of the cloverleaf shape as well as the presence of pseudoknots. The structure has a lower free energy (-12.4 vs -11.88) than the true structure for the tRNA. Figure 5.2a shows the true structure which coincides with the predicted structure when including restricted bases. The restricted bases occur in the last three bases of the sequence — base 74,75, 76. Observe that when these last three bases change from pseudoknotted base pairs to unpaired bases with the addition of restricted bases, the predicted structure forms into the reference structure.

Benchmark We gauged KnotAli's ability to predict both pseudoknotted and pseudoknot-free structure. We use 10 different families. Two of the families, 5s and tRNA, are known to be highly conserved while the other six are variable in their conservation. The performance of our proposed algorithm, KnotAli, was compared to three alternative algorithms: Hxmatch [64], RNAalifold [4], and Cacofold [50]. Tables 5.2a and 5.2b demonstrate KnotAli's prediction accuracy on our dataset against the other algorithms. The table consists of the sensitivity, PPV, and F-measures of the algorithms where F-measure is used as the primary scoring metric. Bold values denote values significant to the others through the methods described in Section 4.4. Our results demonstrate a significant improvement in F-measure across 6 of the 10 families using MUSCLE and 7 of the 10 families using MAFFT. Within Table 5.2a, RNAalifold was found to be significant in the other four families. We place an asterisk

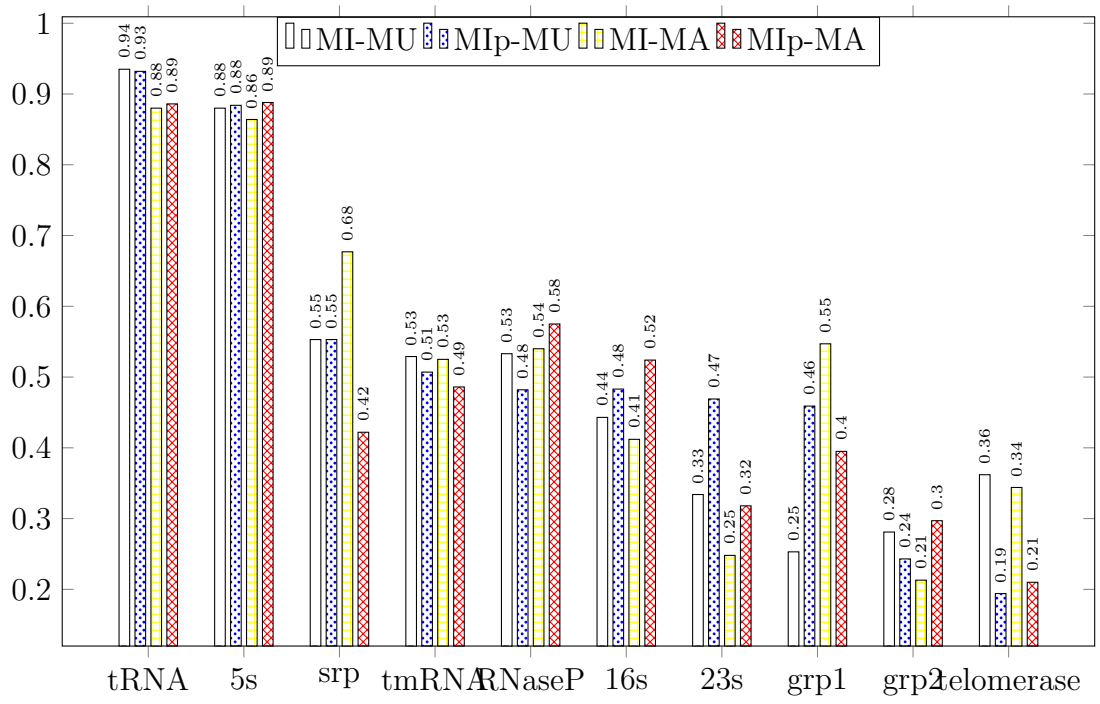
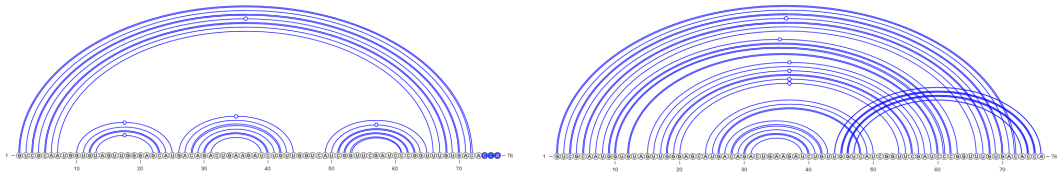


Figure 5.1: We compare the use of `MI` versus `APC` in KnotAli. The set of F-measures obtained for each family are averaged to find the mean F-measure which is then used as the basis of comparison. We show the results on both MUSCLE and MAFFT denoted by MU and MA in the legend respectively.

on RNAalifold's F-measure for tRNA to denote that the p-value did not cross .05. The permutation test between RNAalifold and KnotAli gave a p-value of .061. We decided this was close enough to warrant distinction. F-measures for Group II Intron were found to be low across all families. Group II Intron represents one of two families with low sequence conservation. Two of algorithms, RNAalifold and Cacofold, found no true pairs within their predicted structure.

Table 5.2b shows the significance of KnotAli within 7 of the 10 families. An asterisk was placed on the F-measure obtained by Cacofold for tRNA. Similar to RNAalifold in Table 5.2a, KnotAli obtained a p-value of .058 when compared to Cacofold. As both algorithms were significant to the other two, we have bolded them both. KnotAli saw a significant decrease in F-measure on 23s when moving from MUSCLE to MAFFT. There was an increase in Group II Intron when shifting to MAFFT, especially for the other three algorithms, but F-measures were still low.

KnotAli suffers from a noticeably lower F-measure on telomerase for both MUSCLE and MAFFT against the other three algorithms. Similarly, the three algorithms



(a) The predicted structure with restricted bases (colored)

(b) The predicted structure without restricted bases

Figure 5.2: The effect of restricted bases is shown within a case of tRNA. Pseudoknotted bases were added to the tRNA structure before the use of restricted bases. The cloverleaf shape known within the tRNA group is also lost. The addition of these restricted bases, which restricted three bases at the end of the sequence, supports the prediction to that of the true structure.

have a noticeably lower score on Group I Intron and SRP for both MUSCLE and MAFFT compared to KnotAli. Group I Intron being the second family with low sequence conservation.

Using the confidence intervals from Section 4.4, we generate box and whisker plots for each family for each algorithm and alignment. Each plot shown is for one family. We split the results from each aligner into two colors: black (MAFFT) and red (MUSCLE). Shortened versions of the algorithm name are also used — detailed more in Figure 5.3. We separate outliers from our data by calculating the whiskers as the IQR (`interquartile range`) $\cdot$ 1.5 added to the third quartile and subtracted from the first quartile. As F-measure is on the range of 0 to 1, whiskers are limited to between 0 and 1 likewise.

We note that the results of the boxes for 5s, Group I Intron, SRP, tmRNA, and tRNAsupplement the results obtained in Tables 5.2a and 5.2b. The boxes, representing the confidence intervals, remain small and close to the mean values for KnotAli.

Areas which deviate from expected include Group II Intron. KnotAli was found to have the largest mean F-measure for Group II Intron on both MUSCLE and MAFFT. While the edges of the box still support this position, KnotAli's box is much larger than the other algorithms. 23s shows the result of larger boxes for both KnotAli and Cacofold when on the MAFFT alignment. This contrasts with both RNAalifold and Hxmatch whose mean is higher and boxes smaller.

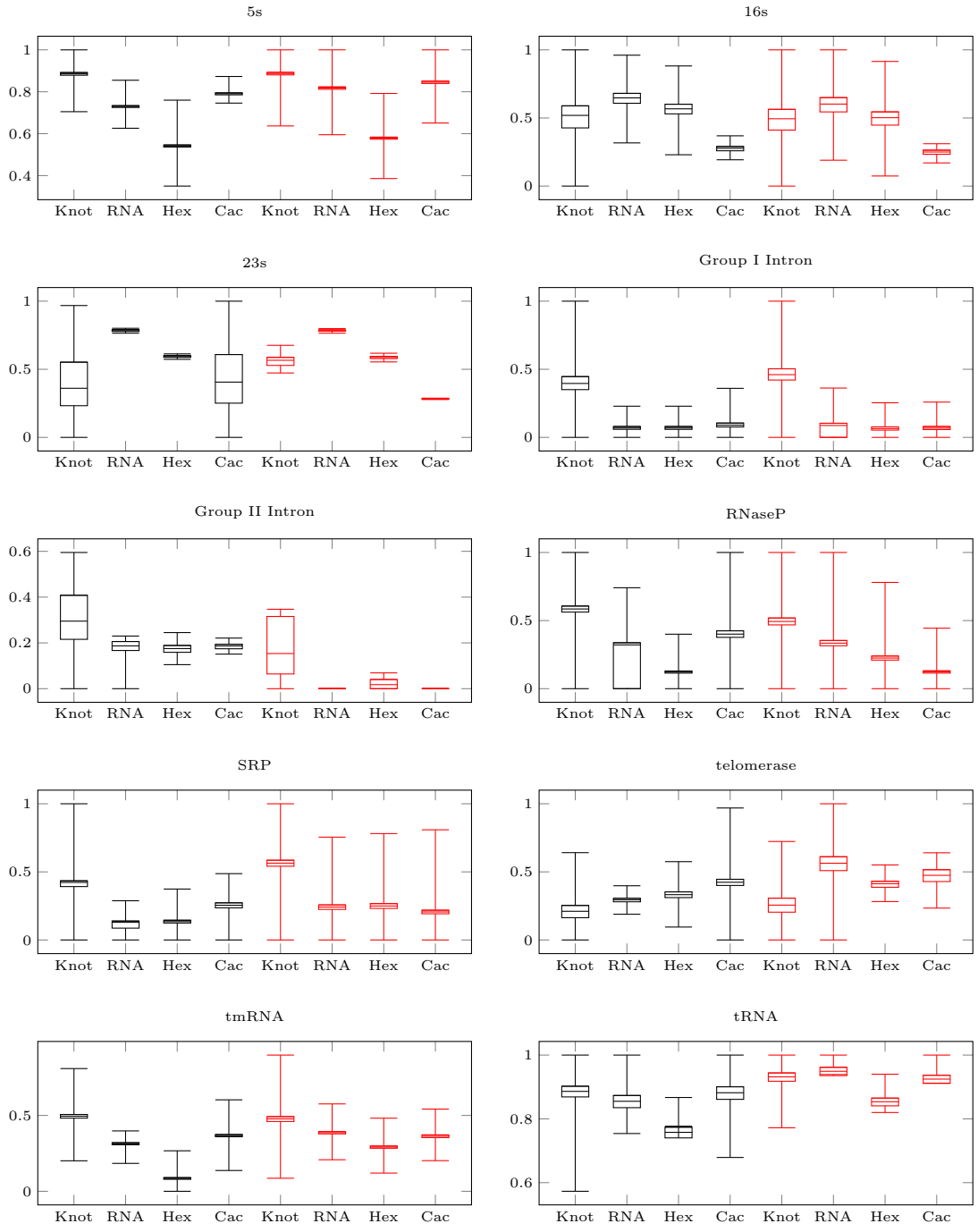


Figure 5.3: We show the results of the four algorithms in a box and whisker plot. Black denotes the results on the MAFFT alignment while red denotes MUSCLE. Algorithms names are shortened to Knot (KnotAli), RNA (RNAalifold), Hex (Hxmatch) and Cac (Cacofold).

thresh	tRNA	5s	SRP	tmRNA	RNaseP	16s	23s	Group I Intr	Group II Intr	telomerase
0	.843	.821	.266	.31	.413	.439	.192	.155	.051	.066
0.1	.867	.809	.245	.301	.418	.432	.19	.15	.05	.07
0.2	.866	.775	.247	.305	.394	.433	.182	.136	.047	.04
0.3	.865	.735	.248	.277	.347	.424	.179	.108	.039	.035
0.4	.851	.727	.221	.24	.267	.41	.169	.072	.035	.039
0.5	.851	.703	.201	.188	.243	.378	.163	.037	.032	.018
0.6	.851	.67	.193	.17	.182	.338	.149	.016	.023	.018
0.7	.813	.622	.18	.125	.121	.284	.126	.007	.017	.019
0.8	.785	.545	.151	.111	.079	.244	.116	.007	.014	.009
0.9	.754	.393	.151	.102	.044	.198	.087	.007	.015	0
1	.727	.31	.092	.075	.007	.167	.071	.007	.015	0
1.1	.693	.257	.071	.066	0	.121	.052	.007	.015	0
1.2	.676	.238	.047	.049	0	.08	.011	0	.015	0
1.3	.595	.219	.025	.04	0	.047	.005	0	.008	0
1.4	.368	.105	0	.031	0	.03	0	0	0	0
1.5	.196	.027	0	.021	0	.01	0	0	0	0
1.6	.086	0	0	0	0	0	0	0	0	0
1.7	.045	0	0	0	0	0	0	0	0	0
1.8	0	0	0	0	0	0	0	0	0	0
1.9	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0

Table 5.1: The heatmap illustrates the results of a grid search across 21 different possible thresholds on the 10 families. The values of the heatmap represent the mean F-measure for the family at the specific threshold and were averaged across MUSCLE and MAFFT.

	KnotAli			RNAalifold			Hxmatch			Cacofold		
family	sen	ppv	F	sen	ppv	F	sen	ppv	F	sen	ppv	F
5s	.899	.876	**.887**	.761	.884	.817	.573	.911	.701	.835	.859	.846
16s	.494	.501	.494	.506	.748	**.602**	.385	.724	.502	.172	.455	.250
23s	.481	.460	.469	.798	.767	**.782**	.552	.625	.587	.242	.341	.283
Group I Intron	.490	.444	**.461**	.047	.588	.087	.034	.529	.064	.039	.300	.068
Group II Intron	.177	.139	**.154**	0	0	0	.010	.108	.018	0	0	0
RNaseP	.498	.491	**.493**	.235	.602	.334	.135	.699	.225	.069	.578	.122
SRP	.580	.556	**.564**	.165	.764	.241	.166	.897	.250	.186	.496	.255
telomerase	.289	.233	.256	.508	.636	**.563**	.412	.707	.412	.380	.480	.423
tmRNA	.491	.468	**.477**	.255	.803	.386	.176	.852	.291	.234	.868	.367
tRNA	.950	.917	.932	.931	.972	**.949***	.764	.974	.854	.886	.970	.925

(a) Input alignment created through MUSCLE.

	KnotAli			RNAalifold			Hxmatch			Cacofold		
family	sen	ppv	F	sen	ppv	F	sen	ppv	F	sen	ppv	F
5s	.902	.871	**.885**	.644	.843	.729	.321	.922	.473	.739	.852	.790
16s	.548	.495	.519	.548	.794	**.647**	.440	.802	.567	.198	.467	.277
23s	.583	.545	.563	.795	.771	**.783**	.563	.627	.593	.404	.449	.406
Group I Intron	.424	.378	**.396**	.036	.719	.068	.036	.719	.069	.051	.384	.090
Group II Intron	.382	.244	**.295**	.106	.773	.186	.103	.580	.175	.105	.920	.187
RNaseP	.592	.583	**585**	.206	.758	.321	.067	.700	.122	.301	.639	.400
SRP	.428	.424	**.423**	.078	.661	.129	.078	.885	.134	.148	.377	.206
telomerase	.243	.186	.211	.250	.361	.294	.255	.483	.333	.442	.517	**.475**
tmRNA	.504	.492	**.495**	.196	.799	.313	.047	.435	.084	.255	.650	.362
tRNA	.898	.878	**.886**	.800	.931	.855	.642	.940	.758	.853	.925	**.882***

(b) Input alignment created through MAFFT.

Table 5.2: Comparison of KnotAli with RNAalifold, Hxmatch, and Cacofold. Each column corresponds to algorithm used and each sub-column represents the metric being looked at: F-measure, Sensitivity or PPV. **BOLD** represents the highest accuracy whose values are significant to the rest. In the case of two algorithms whose accuracy outperformed the rest while not significantly better than each other, both were represented in bold. An accompanying $*$ is then used to denote a p-value close to but not below .05

Chapter 6

Conclusions

6.1 Discussion

We recognize that the reference structures for the sequences within the dataset were determined through comparative sequence analysis [16]. We noted in Section 1 that CSA has been shown to accurately predict secondary structures [32]. Structures predicted are not guaranteed to contain all base pairs from the true structure. Within the reference structures of some families provided by CSA, there are large loops indicating a lack of determined structure for the segment. We thereby introduce the alternative scoring metric for grading our structures called Matthew Correlation Coefficient (MCC) [6, 28]. MCC, unlike F-measure, was found to be more reliable on binary classifications compared to F-measure [30]. We modify the MCC metric using the changes done in [29] which changes the distinction for false positives. False positives are considered *inconsistent* when for a predicted pair $i.j$, there is a pair in the reference structure where either $i.k$ or $h.j$ and $h \neq i$ and $k \neq j$. False positives are also considered *contradicting* when there is a predicted pair $i.j$ where $k < i < l < j$. Pairs neither contradicting nor inconsistent are termed *compatible* and do not contribute to the number of false positives. Within our scoring, due to the presence of pseudoknots, pairs are compatible as long as they are not inconsistent.

We note in Table 6.1 that the results from KnotAli and the other three algorithms are overall much lower in many of the families compared to when using F-measure – seen in Tables 5.2a and 5.2b. There is a clear shift in the results from when using F-measure. Counter to the results in Tables 5.2a and 5.2b, KnotAli was only significant in 3 of the 10 families when using MCC. We do note that RNAalifold for tmRNA on

	MUSCLE				MAFFT			
family	Knot	RNA	Hex	Cac	Knot	RNA	Hex	Cac
tRNA	**.934**	**.921**	.765	.875	**.855**	.784	.654	**.831**
5s	**.832**	.700	.436	.777	**.833**	.605	.435	.692
SRP	.223	.246	.246	.005	-.003	**.165**	**.165**	-.076
tmRNA	.185	**.302**	.255	**.312**	.233	**.255***	.077	.180
RNaseP	.153	.180	.194	.096	**.323**	.236	.142	.282
16s	.199	**.457**	.364	.041	.237	**.504**	.432	.054
23s	.263	**.695**	.378	-.124	.098	**.696**	.378	-.148
Group I Intron	**.236**	.105	.085	.003	.117	.111	.111	.065
Group II Intron	-.051	0	.016	0	.189	.216	.190	.258
telomerase	-.041	**.432**	.326	.230	-.148	.005	.162	**.301**

Table 6.1: Comparison of KnotAli with RNAalifold, Hxmatch, and Cacofold. Abbreviations are made for each family and can be referenced against Table 4.1. Each column represents the aligner used and each subcolumn represents the algorithm where the values represent the MCC. Abbreviations are made for each algorithm: Knot (KnotAli), RNA (RNAalifold), Hex (Hxmatch), and Cac (Cacofold). **BOLD** represents an MCC value significant to the rest for the aligner being looked at.

the MAFFT alignment had a p-value of .055 when compared to KnotAli. This does not constitute significance, but, due to the relative closeness to a p-value of .05, we determined it to be worth acknowledging. We have denoted this with an asterisk.

$$MCC = \frac{TP \cdot TN - (FP - \gamma) \cdot FN}{\sqrt{(TP + FP - \gamma)(TP + FN)(TN + FP - \gamma)(TN + FN)}} \tag{6.1}$$

The multiplicative nature of FP and FN is thought to play a role in the change in scores. MCC treats the a reduction of one false metric as an improvement over the averaging of both — see Equation 6.1. We find this to be true in the calculation of MCC for RNaseP on the MUSCLE alignment. KnotAli's sensitivity is substantially higher, with a difference of .263 and .363 for RNAalifold and Hxmatch, respectively, than comparative results but has a lower PPV by .111 and .208. This change in PPV led to an overall lower MCC for KnotAli compared to these algorithms. Furthermore, we believe the difference in score to be due to the multifaceted nature of scoring pairings. When referring back to [30], MCC was found to be better on specifically binary classifications. In the case where we are looking at simply whether a base is paired or not paired, the view of this being a binary classification would hold true. This is, however, not how we score whether we have a TP or a FP. Rather, to have a

TP, the base must be paired correctly to another base. This shifts the classification to a non-binary classification rather than a binary one. The effect from using MCC is that the relative weight of TP is not appropriate for the difficulty in finding them. For this reason, F-measure is a better classification method when evaluating structures in RNA. From this initial test, we further investigated the cases where predictive ability was lower in a subset of the families.

We see in Tables 5.2a, and 5.2b that all algorithms perform poorly on Group I Intron and SRP — noted in Section 5. Both SRP and Group I Intron are families whose varying sizes cause the sequences to be heavily gapped compared to other families. To study this, we evaluated how the algorithms perform when the range in length is decreased. We were interested in seeing how the output from RNAalifold, whose output includes a heavily gapped consensus sequence for these families, change in response. The sequence lengths were restricted for both families, for SRP between 200 and 350, and for Group I Intron between 325 and 450 resulting in 285 and 33 sequences, respectively. After restricting the number of sequences, sequences were re-aligned using MAFFT. Results were then compared to the F-measure of structures from the previous prediction. The resulting alignment added a significant number of base pairs compared to the original prediction for both SRP and Group I Intron.

For RNAalifold's prediction, SRP had 23 more predicted base pairs, changing the F-measure from .19 to .315. Group I Intron changed from 5 base pairs predicted to 39 base pairs. This resulted in an increase of F-measure from .068 to .238. Hxmatch saw an increase from .134 to .300 for SRP and an increase from .069 to .339 for Group I Intron in F-measure when using the constrained set. Cacofold saw an increase in SRP from .206 to .419 and from .09 to .167 in Group I Intron. Using KnotAli with the constrained set saw the same result as with the other three algorithms. There is a change in SRP from .423 to .581 and from .396 to .546 in Group I Intron. These changes were found to be significant for all algorithms.

	Pre-shortened				Post-shortened			
family	Knot	RNA	Hex	Cac	Knot	RNA	Hex	Cac
SRP	.423	.190	.134	.206	.581	.315	.300	.419
Group I Intron	.396	.068	.069	.090	.546	.238	.339	.167

Table 6.2: Comparison of KnotAli with RNAalifold, Hxmatch, and Cacofold on two families with large variation in sequence length. Scores are compared between pre-shortened versions and post-shortened versions.

	KnotAli			Cacofold		
family	sen	ppv	F	sen	ppv	F
RF00177	.507	.481	.486	.067	.154	.093

Table 6.3: Comparison of KnotAli to Cacofold within a spurious alignment. Alignment was chosen based on the results of RNAconTest [65].

We note that the effect of having sequences of highly varying sizes had a lesser effect on KnotAli compared to the other algorithms. We further study this attribute by looking at the effect of poor alignment quality on prediction. Cacofold, which is more affected by spurious covariations in an alignment, is of interest. In Section 4.5, we observed that the Cacofold uses no additional settings. This sets the minimum loop size to 1 — the standard minimum loop size is 3 as smaller loops are less stable. We found that, within Cacofold, spurious alignments resulted in more loops of size < 3.

Cacofold, similarly to KnotAli, uses `APC` as a form of background correction on its covariation measure. [50] [51]. Cacofold differs in its approach by its use of a G-test covariation measure rather than `MI`. In addition, Cacofold makes use of positive and negative base pairs — discussed in Section 3.3.3. KnotAli finds similarity to negative base pairs in its objective of limiting improbable base pairs, but it solves the issue differently. KnotAli focuses on individual bases rather than pairings within its algorithm. KnotAli restricts a base from forming a pair with any other base rather than limiting singular base pairs.

Continuing our study, we assessed KnotAli and the alternative algorithms on a known spurious alignment [65]. We measured on a family with a moderate number of sequences containing structures which include pseudoknots. When choosing an alignment, we moved toward one that had a lower alignment score (scoring metric found in [65]). The alignment used was RF00177, bacteria small subunit ribosomal RNA, consisting of 32 sequences of average length: $1,476$. All algorithms were run with the same setting as was discussed in Section 4.5. The results showed KnotAli obtained an F-measure of .564. In contrast, Cacofold obtained an F-measure of .093.

Our results in both the Table 6.2 and 6.3 show that KnotAli's ability to predict was superior as alignment quality decreases. It also shows the significance of alignment quality on predictive ability. From this, we assert that alignment quality plays a significant role in the ability of alignment-based algorithms to predict structures.

6.2 Conclusion

In this work we introduced KnotAli, a pseudoknotted secondary structure prediction algorithm that utilizes evolutionary information to make informed energy minimization. We evaluated KnotAli's performance against three competing algorithms (RNAalifold, Hxmatch and Cacofold) and we found KnotAli's performance on a large dataset to be more accurate in majority of the cases. In addition, KnotAli is more resilient to changes in alignment quality compared to the other algorithms. Other comparison-based approaches rely on a substantial number of homologous sequences and a non-spurious alignment in order to make accurate structure predictions. Implementation of thermodynamic energy minimization that is informed, but not restricted, by comparison allows KnotAli to find structures within families with less conservation and within more spurious alignments.

6.3 Shortcomings of KnotAli

KnotAli struggles with its time efficiency when it comes to large sequences. Current alignment-based algorithms find the consensus structure for an alignment. KnotAli, in comparison, finds the structure for each individual sequence within its thermodynamic portion. KnotAli's time complexity is $\mathcal{O}(Nn^3)$ where N is the number of sequences and n is the length of the sequence. This is in line with other alignment-based algorithms. KnotAli's practical time, however, is much longer than comparable algorithms for longer sequences.

6.4 Future Work

One of the future goals for this work is to improve the thermodynamic prediction time and accuracy. We would accomplish this through the integration of SparseMFEFold [62] into KnotAli. SparseMFEFold is a pseudoknot-free prediction algorithm which use sparsification to improve space efficiency. As well, the algorithm is more time efficient than simfold, the algorithm it will replace. Within the thermodynamic portion of KnotAli is the algorithm simfold which works as a pseudoknot-free structure prediction by which the pseudoknotted structure prediction works off of.

The integration of SparseMFEFold would decrease the running time of KnotAli significantly compared to current running times. The introduction of this would also

lead to a sparsified pseudoknotted structure prediction algorithm which could be used in single sequence and multi sequence prediction. Likewise, benefiting from this improvement would be our second future direction.

This future direction is the windowing of the algorithm on sufficiently long sequences. Prediction time and space become prohibitive as the length of the sequence increases. To offset the cost of long RNA, windowing would reduce the effective length of the sequence. Identification of window locations would occur during the covariation portion of the algorithm. By finding low covarying positions outside of loops, the positions would be marked within the intermediary structure. Subsequences and substructures would be later used in determining the MFE structure.

Bibliography

[1] Tatsuya Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1):45–62, 2000.

[2] S. L. Alam, J. F. Atkins, and R. F. Gesteland. Programmed ribosomal frameshifting: Much ado about knotting! *PNAS*, 96:14177–14179, Dec 1999.

[3] Mirela S. Andronescu, Cristina Pop, and Anne E. Condon. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, 16:26–42, Jan 2010.

[4] Stephan H Bernhart and Ivo L Hofacker et. al. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9, Nov 2008.

[5] Eckart Bindewald and Bruce Shapiro. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, 12:342–352, Mar 2006.

[6] B.W.Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451, Oct 1975.

[7] José Almeida Cruz and Eric Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136:604–609, Feb 2009.

[8] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25:1974–1975, Aug 2009.

[9] Alexandre de Lencastre and Anna Marie Pyle. Three essential and conserved regions of the group II intron are proximal to the 5‘-splice site. *RNA*, 14:11–24, Jan 2008.

[10] Robin D Dowell and Sean R Eddy. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, 5, Jun 2004.

[11] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, Aug 2004.

[12] Chad R. Bernier et. al. Translation: The Universal Structural Core of Life. *Mol Biol Evol*, 1:2065–2076, Aug 2018.

[13] Daniel Sundfield et. al. Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, 32, Apr 2016.

[14] Ebbe Sloth Andersen et. al. The tmRDB and SRPDB resources. *Nucleic Acids Res*, 34:D163—-D168, Jan 2006.

[15] Hosna Jabbari et. al. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*, 34:3849—-3856, Jun 2018.

[16] Ioanna Kalvari et. al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49:D192—-D200, Jan 2021.

[17] J. Vierna et. al. Systematic analysis and evolution of 5S ribosomal DNA in metazoans. *Heredity*, 111:410–421, Nov 2013.

[18] Marcel Martinez-Porchas et. al. How conserved are the conserved 16S-rRNA regions? *Heredity*, 5:e3036, Feb 2017.

[19] Martin Raden et. al. Freiburg RNA tools: a central online resource for RNA-focused research and teaching. *Nucleic Acids Research*, 46:W25—-W29, Jul 2018.

[20] Nilay Peker et. al. A comparison of Three Different Bioinformatics Analyses of the 16S-23S rRNA encoding Region for Bacterial Identification. *Front Microbiol*, 10:620, Apr 2019.

[21] Padideh Danaee et. al. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res*, 46:5381—-5394, Jun 2018.

[22] Ronny Lorenz et. al. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6, Nov 2011.

[23] Sebastian Will et. al. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *Plos Computational Biology*, 3:900—-914, Apr 2007.

[24] Sebastian Will et. al. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18:900—-914, Mar 2012.

[25] Stanislava Gunisova et. al. Identification and comparative analysis of telomerase RNAs from Candida species reveal conservation of functional elements. *RNA*, 15:546—-559, Apr 2009.

[26] Tina Uroda et. al. Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell*, 75:982–995, Aug 2019.

[27] W J Melchers et. al. Kissing of the two predominant hairpin loops in the coxsackie B virus 3' untranslated region is the essential structural feature of the origin of replication required for negative-strand RNA synthesis. *J Virol.*, 71:686—-696, Jan 1997.

[28] Paul P Gardner and Robert Giegerich. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, May 2000.

[29] Paul P Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5, Sep 2004.

[30] Paul P Gardner and Robert Giegerich. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, Sep 2020.

[31] D R Groebe and O C Uhlenbeck. Characterization of RNA hairpin loop stability. *Nucleic Acids Res*, 16:11725—-11735, Dec 1988.

[32] Robin R Gutell, Jung C Lee, and Jamie J Cannone. The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology*, 12(3):301–310, 2002.

[33] E S Haas and J W Brown. Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res.*, 26:4093—-4099, Sep 1998.

[34] Ivo Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, Jul 2003.

[35] Christine E. Holt and Simon L. Bullock. Subcellular mRNA Localization in Animal Cells and Why It Matters. *Science*, 326:1212–1216, Sep 2013.

[36] Hosna Jabbari and Anne Condon. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. *BMC Bioinformatics*, 15, May 2014.

[37] Kazutaka Katoh and Daron M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol*, 30:772–780, Apr 2013.

[38] D A Kirby, S V Muse, and W Stephan. Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A.*, 92:9047–9051, Sep 1995.

[39] B Knudsen and J Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15:446—-454, Jun 1999.

[40] Marilyn Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, Nov 2005.

[41] S Lindgreen, P.P. Gardner, and A. Krogh. Measuring covariation in RNA alignments: physical realism improves information measures. *BMC Bioinformatics*, 22:2988—-2995, Dec 2006.

[42] Rune B Lyngsø and Christian N Pedersen. RNA pseudoknot prediction in energy-based models. *J Comput Biol*, 7:409–427, 2000.

[43] Kelsey C. Martin and Anne Ephrussi. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*, 136:719–730, Feb 2009.

[44] David H Mathews and Douglas H Turner. Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology*, 16(3):270–278, Jun 2006.

[45] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15:469—479, May 2014.

[46] Eric P Nawrocki, Thomas A Jones, and Sean R Eddy. Group I introns are widespread in archaea. *Nucleic Acids Research*, 46(15):7970—-7976, Sep 2018.

[47] Daewoo Pak, Robert Root-Bernstein, and Zachary F. Burton. tRNA structure and evolution and standardization to the three nucleotide genetic code. *Transcription*, 8(4):205–219, Jun 2017.

[48] Adam Pocock, Gavin Brown, Ming-Jie Zhao, and Mikel Lujan. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J Mach Learn Res*, 13(1):27—66, Jan 2012.

[49] E Rivas and S R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol.*, 285:2053–2068, Feb 1999.

[50] Elena Rivas. RNA structure prediction using positive and negative evolutionary information. *PLoS Comput Biol.*, 16(10):1–25, Oct 2020.

[51] Elena Rivas. Evolutionary conservation of RNA sequence and structure. *WIREs RNA*, page e1649, Mar 2021.

[52] Elena Rivas, Jody Clements, and Sean R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods*, 14:45–48, Jan 2017.

[53] Elena Rivas, Jody Clements, and Sean R. Eddy. Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics*, 36:3072—3076, May 2020.

[54] Elena Rivas, Raymond Lang, and Sean R Eddy. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA*, 18:193–212, Feb 2012.

[55] David Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *Siam J Appl. Math.*, 45:810–825, 1985.

[56] L.M. Wahl S.D. Dunn and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *BMC Bioinformatics*, 24:333—-340, Feb 2008.

[57] Saad Sheikh, Rolf Backofen, and Yann Ponty. Impact of the energy model on the complexity of RNA folding with pseudoknots. In *Combinatorial Pattern Matching*, pages 321–333, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[58] Michael F Sloma and David H. Mathews. Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA*, 22:1808–1818, Aug 2016.

[59] David W Staple and Samuel E Butcher. Pseudoknots: RNA Structures with Diverse Functions. *Plos Biology*, 3:E213, Jun 2005.

[60] M. Bryan Warf and J. Andrew Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci.*, 35:169–178, Mar 2010.

[61] Claus O Wilke, Richard E Lenski, and Christoph Adami. Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. *BMC Evol Biol*, 3, Feb 2003.

[62] Sebastian Will and Hosna Jabbari. Sparse RNA Folding Revisited: Space-Efficient Minimum Free Energy Prediction. In *Algorithms in Bioinformatics*, pages 257—-270, Aug 2015.

[63] Timothy J. Wilson and David M.J. Lilley. RNA catalysis—is that it? *RNA*, 21:534–537, Apr 2015.

[64] Christina Witwer, Ivo L. Hofacker, and Peter F. Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(2):66–77, Apr 2004.

[65] Erik S. Wright. RNAconTest: Comparing tools for non-coding RNA multiple sequence alignment based on structural consistency. *RNA*, 26:531–540, Jan 2020.

[66] Michael Zuker and Peter Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133—-138, Jan 1981.

[67] Christian Zwieb, Iwona Wower, and Jacek Wower. Comparative sequence analysis of tmRNA. *Nucleic Acids Res.*, 27:2063—-2071, May 1999.