KnotAli: Informed Energy Minimization Through the Use of Evolutionary Information

Mateo Gray (✉ mgray7@uvic.ca)

University of Victoria

Sean Chester

University of Victoria

Hosna Jabbari

University of Victoria

Research Article

Keywords: RNA secondary structure, MFE, Pseudoknot, sequence alignment, covariation, thermodynamic energy minimization

Posted Date: September 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-700965/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

KnotAli: Informed Energy Minimization Through the Use of Evolutionary Information

Mateo Gray^{*}
 , Sean Chester^{*}
 and Hosna Jabbari^{*}

^{*}Correspondence: mgray7@uvic.ca; schester@uvic.ca; jabbari@uvic.ca
 Department of Computer Science,
 University of Victoria, Victoria,
 Canada
 Full list of author information is
 available at the end of the article

Abstract

Background

Improving the prediction of structures, especially those containing pseudoknots (structures with crossing base pairs) is an ongoing challenge. Homology-based methods utilize structural similarities within a family to predict the structure. However, their prediction is limited to the consensus structure, and the quality of the alignment. Minimum free energy (MFE) based methods, on the other hand, do not rely on familial information and can predict structures of novel RNA molecules. Their prediction normally suffers from inaccuracies due to their underlying energy parameters.

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 a multiple 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 programs, 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.

Conclusions

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. KnotAli can be found online on github at <https://github.com/mateog4712/KnotAli>

Keywords: RNA secondary structure; MFE; Pseudoknot; sequence alignment; covariation; thermodynamic energy minimization

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 [1], translation [1, 2], splicing [3, 4], catalysis [1, 5] and regulating gene expression [1, 3, 6, 7]. 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 [8, 9]. 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 [10]. 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 [11]. These programs use a set of empirical parameters to define the folding energies of a structure, where every structural feature has been assigned a specific free 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 [12] uses thermodynamic energy minimization, Hxmatch [13] uses maximum weighted matching (MWM), and Cacofold [14] uses an RNA-based grammar. Despite their coupling, these programs still heavily rely on the quality of the alignment to make accurate predictions. These programs, in addition, only predict the consensus structure rather than the structures for all sequences. Within alignment-based programs, 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 [12]) and pseudoknotted (Hxmatch [13], and Cacofold [14]). 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.

2 RNA secondary structure

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, 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 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.

Algorithms We start with a high level definition of how the different types of algorithms work and their complexities.

Alignment based algorithms [12, 13, 14] 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.

Thermodynamics-based algorithms [16, 17, 18] 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 free energy is minimum. For pseudoknot-free structure prediction, the standard time and space complexity is $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$.

MFE pseudoknotted secondary structure prediction is NP-hard [19, 20] and inapproximable [21]. Polynomial-time algorithms require limiting the class of pseudoknotted structures as time complexity is traded off with generality [11]. The most general thermodynamics-based algorithm is PKnots [18] 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 [22, 23] within the remainder of RNA gives cause to creating algorithms which can predict them.

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 biologically important types of pseudoknots i.e. kissing hairpins and H-type pseudoknots with arbitrary 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.

3 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 columns. 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 maximum 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 maximum 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 maximum 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 is used. 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 [24, 25].

The resultant structure is given as guide to the thermodynamic portion of the algorithm. KnotAli follows the relaxed hierarchical folding hypothesis [17], 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. 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, as shown in Figure 3. 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 $W(i, j)$ 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 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 [26]. 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:

$$MI(a,b) = \sum_{x,y \in \{A,C,G,U\}} f_{a,b}(x,y) \cdot \log_2 \left(\frac{f_{a,b}(x,y)}{f_a(x) \cdot f_b(y)} \right) \quad (1)$$

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 inter-dependence 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 [27]. Average Product Correction [28] 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 2. Its components defined in Equation 3 and 4.

$$APC(a,b) = \frac{MI(a,\bar{z}) \cdot MI(b,\bar{z})}{MI_{\text{avg}}} \quad (2)$$

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

$$MI_{\text{avg}} = \frac{2}{n(n-1)} \sum_{w=0}^{n-1} \sum_{z=0}^{n-1} \begin{cases} MI(w,z) & \text{where } |w-z| > 3 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

thresh	tRNA	5s	SRP	tmRNA	RNaseP	16s	23s	Group I Intron	Group II Intron	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 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.

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

$$MI_p(a, b) = MI(a, b) - APC(a, b) \quad (5)$$

MI_p was found to be sensitive and selective compared to the original MI [28].

To determine at what point the MI_p 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 (See Table 1). All pairings with $MI_p > t$ 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.

3.3 Compared Algorithms

KnotAli was compared to three algorithms: RNAalifold [12], Hxmatch [13], and Cacofold [14]. Similar to KnotAli, both Hxmatch and Cacofold are alignment-based algorithms which return possibly pseudoknotted structures. RNAalifold is restricted to pseudoknot-free structures.

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. These options 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

$$h(i, j) = \begin{cases} 1 & \text{if } i \neq j \text{ and } i.j \in \{AU, CG, GC, GU, UA, UG\} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

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

$$R(i.j, i'.j') = \log \left(\frac{f(i.j, i'.j')}{f(i, i') \cdot f(j, j')} \right) \quad (7)$$

where $f(i.j, i'.j')$ is the frequency of base pairs $i.j$ and $i'.j'$ being aligned, and $f(i, i')$ (and $f(j, j')$) is the frequency of aligning nucleotides as positions i and i' (and j and j'). The full covariance term for columns a, b is then $\frac{1}{2} \sum_{\substack{a, b \in A \\ a \neq b}} x R(a_i.a_j, b'_i.b'_j)$. 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 sequence. The value is multiplied by -1 to make it positive and placed in a scoring matrix H .

$$C_{a,b} = \sum_{i.j, i'.j'} f(i.j) \cdot D_{i.j, i'.j'} \cdot f(i'.j') \quad (8)$$

where $f(i,j)$ is the frequency of base pair XY 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 . The helix and covariation scores are then combined into a matrix

$$\pi_{a,b} = H_{a,b} + \phi_{a,b} \cdot C_{a,b} \quad (9)$$

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 [29, 30, 14]. Cacofold uses E-value and covariation power in tandem to distinguish what are called positive and negative base pairs.

An *E-value* is an expectation value signifying the expected number of false positives [29]. E is defined as $E = NP(\text{score} > X)$ where N is the number of column pairs 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 [14, 30]. Covariation power is used to distinguish when a lack of structure is due to low sequence variation rather than low covariation.

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 secondary structure prediction algorithms. RNA basic grammar [31] 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 [32, 33], and are used to find additional helices. The structures formed from each subset are combined after filtering out redundancies without covariation support.

4 Experiment Design

In this section we provide the details of our experiment design.

Family	# of Sequences	Sequence length
5s	1053	103-135
16s	22	950-1995
23s	5	2904-2968
Group I Intron	89	210-736
Group II Intron	11	619-780
RNaseP	410	120-486
SRP	583	28-533
telomerase	37	382-559
tmRNA	363	102-437
tRNA	461	57-93

Table 2 List of families with their conservation level, corresponding number of sequences, and range of length

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 [34]. Table 2 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 2 were compiled from [35] where duplicates were removed. Sequence lengths vary between 28 nucleotides long at the minimum and 2,868 nucleotides at the maximum. We find that this dataset represents a wide degree of conservation ranging from tRNA which is highly conserved [36] to less conserved families such as Group I Intron and Group II Intron which lack sequence conservation [37, 38]. We modify the families to remove all hairpins with size < 3 [24, 25]. This removal occurred within the SRP family.

4.2 RNA Sequence Aligners

To evaluate 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 a sequence aligner therefore plays a fundamental role in the quality of its predicted alignments. In a previous benchmark study [39], 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 10 aligners either predict solely off of sequence similarity or by combining the sequence similarity with structure prediction. Of these 10 aligners, we chose MUSCLE [40] and MAFFT [41], as they solely use sequence similarity. 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).

$$Sensitivity = \frac{TP}{TP + FN} \quad (10)$$

$$PPV = \frac{TP}{TP + FP} \quad (11)$$

$$F_{measure} = \frac{2 \cdot PPV \cdot Sensitivity}{PPV + Sensitivity} \quad (12)$$

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 Significance Analysis

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 'perm' package 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 3

family	MUSCLE						MAFFT					
	RNAalifold			Riboalifold			RNAalifold			Riboalifold		
	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	.81	.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	.93	.829	.8	.931	.855*

Table 3 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

on our dataset. We find that for the majority of families, a switch to RIBOSUM matrices show a significant improvement in the score.

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.

5 Results

Preliminary We begin by evaluating how well APC increases the F-measures relative to just MI. Figure 4 illustrate the results on our dataset. We compare the two measures within KnotAli while keeping the thermodynamic portion the same.

Figure 5b 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 5a 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 [13], RNAalifold [12], and Cacofold [14]. Tables 4a and 4b demonstrate KnotAli’s prediction accuracy on our dataset against the other algorithms. The tables consist 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 prediction improvement in KnotAli’s F-measure compared to other algorithms across 6 of the 10 families using

MUSCLE as aligner and 7 of the 10 families using MAFFT as aligner. Within Table 4a, RNAalifold was found to be significant in the other four families. We place an asterisk 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 4b 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

family	KnotAli			RNAalifold			Hxmatch			Cacofold		
	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	.589	.545	.566	.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	.25	.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.

family	KnotAli			RNAalifold			Hxmatch			Cacofold		
	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	.329	.468	.361	.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	.25	.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	.8	.931	.855	.642	.940	.758	.853	.925	.882*

(b) Input alignment created through MAFFT.

Table 4 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

in Table 4a, 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 programs. Similarly, the three programs 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.

6 Discussion

We recognize that the reference structures for the sequences within the dataset were determined through comparative sequence analysis [34]. We noted in Section 1 that CSA has been shown to accurately predict secondary structures [10]. 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. From the current calculation of F-measure, predictions on these undetermined portions are treated as false

positives. We find that this may influence program accuracies. Other measures of accuracy (such as modified versions of Matthew Correlation Coefficient [42, 43]) that can evaluate the general shape of a structure compared to the reference structure should be investigated.

We see in Tables 4a, and 4b that all programs 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 programs 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 realigned 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.

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 [14] [44]. 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 [39]. 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 [39]). The alignment used was RF00177, bacteria small subunit ribosomal

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

Table 5 Comparison of KnotAli to Cacofold within a spurious alignment. Alignment was chosen based on the results of RNAconTest [39].

RNA, consisting of 32 sequences of average length: 1,476. All programs were run with the same setting as was discussed in Section 4.5. The results are shown in Table 5 showing that KnotAli obtained an F-measure of .564. In contrast, Cacofold obtained an F-measure of .093.

7 Conclusions

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.

Declarations

Acknowledgements

The authors thank the COBRA lab for their feedback on the manuscript.

Funding

Funding was provided through NSERC Discovery grants. Funding provided no role in the design of the study.

Availability of data and materials

The datasets generated and/or analysed during the current study are available in the KnotAli-RawData repository, <https://github.com/mateog4712/KnotAli-RawData>

Ethics approval and consent to participate

Not Applicable

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Authors' contributions

MG designed, developed, and assessed the algorithm, acquired and interpreted the data, and was the primary contributor in writing the manuscript. HJ first bridged the idea. HJ and SC supervised the research and participated in interpreting the data and revising the manuscript. All authors read and approved the final manuscript.

Author details

Department of Computer Science, University of Victoria, Victoria, Canada.

References

1. Cruz, J.A., Westhof, E.: The dynamic landscapes of RNA architecture. *Cell* **136**, 604–609 (2009). doi:10.1016/j.cell.2009.02.003
2. Kozak, M.: Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13–37 (2005). doi:10.1016/j.gene.2005.06.037
3. Mortimer, S.A., Kidwell, M.A., Doudna, J.A.: Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15**, 469–479 (2014). doi:10.1038/nrg3681
4. Warf, M.B., Berglund, J.A.: Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci.* **35**, 169–178 (2010). doi:10.1016/j.tibs.2009.10.004
5. Wilson, T.J., Lilley, D.M.J.: RNA catalysis—is that it? *RNA* **21**, 534–537 (2015). doi:10.1261/rna.049874.115
6. Holt, C.E., Bullock, S.L.: Subcellular mRNA Localization in Animal Cells and Why It Matters. *Science* **326**, 1212–1216 (2013). doi:10.1126/science.1176488
7. Martin, K.C., Ephrussi, A.: mRNA Localization: Gene Expression in the Spatial Dimension. *Cell* **136**, 719–730 (2009). doi:10.1016/j.cell.2009.01.044
8. Kirby, D.A., Muse, S.V., Stephan, W.: Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A.* **92**, 9047–9051 (1995). doi:10.1073/pnas.92.20.9047
9. Wilke, C.O., Lenski, R.E., Adami, C.: Compensatory mutations cause excess of antagonistic epistasis in RNA secondary structure folding. *BMC Evol Biol* **3** (2003). doi:10.1186/1471-2148-3-3
10. Gutell, R.R., Lee, J.C., Cannone, J.J.: The accuracy of ribosomal RNA comparative structure models. *Current Opinion in Structural Biology* **12**(3), 301–310 (2002). doi:10.1016/S0959-440X(02)00339-1
11. Mathews, D.H., Turner, D.H.: Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology* **16**(3), 270–278 (2006). doi:10.1016/j.sbi.2006.05.010
12. Bernhart, S.H., et. al, I.L.H.: RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9** (2008). doi:10.1186/1471-2105-9-474
13. Witwer, C., Hofacker, I.L., Stadler, P.F.: Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **1**(2), 66–77 (2004). doi:10.1109/TCBB.2004.22
14. Rivas, E.: RNA structure prediction using positive and negative evolutionary information. *PLoS Comput Biol.* **16**(10), 1–25 (2020). doi:10.1371/journal.pcbi.1008387
15. Darty, K., Denise, A., Ponty, Y.: VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009). doi:10.1093/bioinformatics/btp250
16. et. al, H.J.: Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics* **34**, 3849–3856 (2018). doi:10.1093/bioinformatics/bty420
17. Jabbari, H., Condon, A.: A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. *BMC Bioinformatics* **15** (2014). doi:10.1186/1471-2105-15-147
18. Rivas, E., Eddy, S.R.: A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J Mol Biol.* **285**, 2053–2068 (1999). doi:10.1006/jmbi.1998.2436
19. Akutsu, T.: Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics* **104**(1), 45–62 (2000). doi:10.1016/S0166-218X(00)00186-4
20. Lyngsø, R.B., Pedersen, C.N.: RNA pseudoknot prediction in energy-based models. *J Comput Biol* **7**, 409–427 (2000). doi:10.1089/106652700750050862
21. Sheikh, S., Backofen, R., Ponty, Y.: Impact of the energy model on the complexity of RNA folding with pseudoknots. In: *Combinatorial Pattern Matching*, pp. 321–333. Springer, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-31265-6_26. https://doi.org/10.1007/978-3-642-31265-6_26
22. et. al, T.U.: Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell* **75**, 982–995 (2019). doi:10.1016/j.molcel.2019.07.025
23. Staple, D.W., Butcher, S.E.: Pseudoknots: RNA Structures with Diverse Functions. *Plos Biology* **3**, 213 (2005). doi:10.1371/journal.pbio.0030213
24. et. al., P.D.: bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res* **46**, 5381–5394 (2018). doi:10.1093/nar/gky285
25. Groebe, D.R., Uhlenbeck, O.C.: Characterization of RNA hairpin loop stability. *Nucleic Acids Res* **16**, 11725–11735 (1988). doi:10.1093/nar/16.24.11725
26. Pocock, A., Brown, G., Zhao, M.-J., Lujan, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *J Mach Learn Res* **13**(1), 27–66 (2012). doi:10.5555/2503308.2188387
27. Lindgreen, S., Gardner, P.P., Krogh, A.: Measuring covariation in RNA alignments: physical realism improves information measures. *BMC Bioinformatics* **22**, 2988–2995 (2006). doi:10.1093/bioinformatics/btl514
28. S.D. Dunn, L.M.W., Gloor, G.B.: Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *BMC Bioinformatics* **24**, 333–340 (2008). doi:10.1093/bioinformatics/btm604
29. Rivas, E., Clements, J., Eddy, S.R.: A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat Methods* **14**, 45–48 (2017). doi:10.1038/nmeth.4066
30. Rivas, E., Clements, J., Eddy, S.R.: Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics* **36**, 3072–3076 (2020). doi:10.1093/bioinformatics/btaa080
31. Rivas, E., Lang, R., Eddy, S.R.: A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA* **18**, 193–212 (2012). doi:10.1261/rna.030049.111
32. Knudsen, B., Hein, J.: RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**, 446–454 (1999). doi:10.1093/bioinformatics/15.6.446
33. Dowell, R.D., Eddy, S.R.: Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics* **5** (2004). doi:10.1186/1471-2105-5-71
34. et. al, I.K.: Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**, 192–200 (2021). doi:10.1093/nar/gkaa1047

35. Sloma, M.F., Mathews, D.H.: Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA* **22**, 1808–1818 (2016). doi:10.1261/rna.053694.115
36. Pak, D., Root-Bernstein, R., Burton, Z.F.: tRNA structure and evolution and standardization to the three nucleotide genetic code. *Transcription* **8**(4), 205–219 (2017). doi:10.1080/21541264.2017.1318811
37. de Lencastre, A., Pyle, A.M.: Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA* **14**, 11–24 (2008). doi:10.1261/rna.774008
38. Nawrocki, E.P., Jones, T.A., Eddy, S.R.: Group I introns are widespread in archaea. *Nucleic Acids Research* **46**(15), 7970–7976 (2018). doi:10.1093/nar/gky414
39. Wright, E.S.: RNAconTest: Comparing tools for non-coding RNA multiple sequence alignment based on structural consistency. *RNA* **26**, 531–540 (2020). doi:10.1261/rna.073015.119
40. Edgar, R.C.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004). doi:10.1186/1471-2105-5-113
41. Katoh, K., Standley, D.M.: MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013). doi:10.1093/molbev/mst010
42. B.W.Matthews: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* **405**, 442–451 (1975). doi:10.1016/0005-2795(75)90109-9
43. Gardner, P.P., Giegerich, R.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000). doi:10.1093/bioinformatics/16.5.412
44. Rivas, E.: Evolutionary conservation of RNA sequence and structure. *WIREs RNA*, 1649 (2021). doi:10.1002/wrna.1649