An Energy Sensitive, Binding-site-aware and Longer-stem-oriented Nussinov Algorithm with Minimal Loop Length Thresholding

Sophie Cao, Yanis Bencheikh, and Vitoria Lara Soria

Keywords:
Long non-coding RNA
Nussinov algorithm
Secondary structure prediction

Abstract Accurately predicting RNA secondary structures and understanding the folding mechanism through which they are generated could allow us to accelerate at an unprecedented level the progress of molecular biology research. We wish to present in this report, our contribution to the democratization of access to in-silico technologies; notably, to offer a novel, preliminarily generalizable, and light dynamic programming framework for the accurate prediction of RNA secondary structures. Grounding the presented variant models in the classical Nussinov-Jacobson algorithm, we have integrated as restrictive conditions in their functioning naturally found trends in RNA secondary structure going from the prioritization of longer stems and loops in trace-back to the distinguishing of varying base-pair-associated stabilizing effects. Carefully benchmarking novel and classical models on a dataset of 12 RNA sequences whose secondary structures are known, we report significant performances for our BS&E and BS&E+AltTB models surpassing threefold the mean accuracy of the classical Nussinov-Jacobson algorithm, and that, with conserved running time of $O(n^3)$ where n represents input RNA sequence length.

 Submitted: December 15th 2023 as the COMP561 Final Project.

1. Introduction

In recent years, there has been substantial research associated with the RNA folding problem. A particular emphasis has been placed on the folding patterns of non-coding RNAs, pivotal regulators of gene expression and epigenetics. Long non-coding RNAs, specifically, have emerged as influencing factors in chromatin function, mRNA stability, translation dynamics, and signaling pathways [1]. These RNAs achieve their function by interacting with not only DNA and proteins, but also other RNAs. Indeed, some non-coding RNAs inhibit their target RNA function through complementary binding, which consequently affects their secondary structure.

The Nussinov-Jacobson (NJ) algorithm stands out as a prominent RNA prediction method that relies on a base pair maximization strategy; nevertheless, it is not without limitations. Notably, the algorithm fails to account for complex structural elements such as pseudoknots and neglects the stacking of base pairs, among other constraints. Its application in the domain of RNA-interacting molecules is particularly challenging due to the complete disregard for binding sites, a crucial factor in understanding the interactions involving RNA.

As a consequence, various dynamic programming (DP) algorithms have been developed to specifically predict joint secondary structures of interacting RNAs. For example, the PairFold algorithm, based on Zuker's algorithm, predicts the secondary structure of two interacting RNAs utilizing minimum free energy [2]. On the other hand, Poolsap et al. introduced DP algorithms addressing an alternate challenge: predicting the secondary structure of a target RNA given the binding sites of its antisense RNA. [3] With a $O(N^3n^3)$ runtime, where N denotes the number of binding sites and n is the sequence length, their algorithms, based on Nussinov and Zuker's, optimize for the sum of base pair energies while integrating stacking energy.

Despite these advances, to our knowledge, no algorithms have been developed for predicting the secondary structures of interacting RNAs with known binding sites, while simultaneously considering biological constraints.

Therefore, this research projects aims to develop a novel algorithm for predicting RNA-interacting structures with known binding sites. Our approach extends Nussinov's algorithm by incorporating constraints grounded in biological knowledge. These constraints include minimal loop length, energy-sensitive base pairing, stem lengths, and loop/bulge size. Notably, the last constraint is based on the premise that a single loop or bulge is more stable than one divided by a base pair [4].

Our proposed algorithms accurately predict the secondary structure for interacting RNAs, with the highest performing model achieving an accuracy of 89%. This specific model is binding-site aware, preventing the formation of base pairs that include bound nucleotides. It is also minimal-loop length aware, utilizing an energy model as a scoring scheme, and employs an alternative traceback method. All of our outlined algorithms have a comparable runtime to Nussinov's $O(n^3)$, where n is a sequence length.

In this paper, we expand on our approach, describing the derivation of variants from Nussinov's algorithm that account for these biological constraints. Furthermore, we showcase the performance of our proposed algorithms on a study-based dataset, highlighting their potential to contribute to the field of RNA folding prediction.

2. Methodology

2.1 The Nussinov Algorithm

The Nussinov-Jacobson (NJ) algorithm first presented in 1980 [5] features an innovative dynamic programming framework to rapidly predict the secondary structure of linear RNA molecules with an average run time of O(n^3) where n is the length of the sequence being predicted. With an identical scoring weight assigned to all classical Watson, Crick, and Franklin base pairings (i.e. A-U and C-G), the NJ algorithm features a trace-back procedure oriented towards the maximization of base pairs and thus, by proxy, the minimization of the ribonucleic acid's free energy.

While its dynamic programming table can feature many secondary structures with equal number of base pairs, it would not be possible to retrieve all of them in a single run as its original trace-back procedure is restricted to the return of a single optimal structure. In the interest of benchmarking the NJ algorithm with respect to our novel approaches, we have utilized implementations of both its original version and of an alternate one with minimal loop length (MLL) restriction. [6]

2.2 Minimal Loop Length

Due to its base pair maximization strategy, the classical Nussinov algorithm does not distinguish between loop sizes. However, literature has demonstrated the infrequency of short loop lengths in RNA structures [1]. Indeed, loops fewer than three bases do not form due to their high conformational instability, while optimal loop lengths tend to range between 4 to 8 bases [4]. To better align algorithmic predictions with these observed structural patterns, we integrated a constraint into the algorithm, such that base pairing is exclusively permitted between nucleotides separated by a distance equivalent to a given minimal loop length MLL.

In the implementation of this constraint, we computed the optimal MLL for each sequence in our database, subsequently deriving an average that excludes zero values. The resulting optimal loop lengths were later used during the testing of our developed algorithms.

2.3 Energy Sensitive Base Pairing

In the classical NJ algorithm, Watson, Crick, and Franklin base pairs are all assigned an identical positive score of +1 orienting the trace-back procedure towards retrieving the secondary structure path with maximal score. Since the discovery of the natural occurence of G-U wobble pairing in RNA molecules [7], variants of the NJ algorithm such as Poolsap et al.'s Stacking energy model (SEM) have integrated the scoring of guanosine-uracile base pairings. [3]

Following the same Clote and Backofen scoring scheme for base pairings [8], we have shifted our NJ variants from maximization oriented frameworks to minimization oriented frameworks whereby the dynamic programming tables used for trace-back contain negative values. These values represent the progressive (i.e. from diagonal entries to the top right entry) and potentially optimal diminishing of the molecules final free energy.

Base pair	Energy value
{G, C}	−5
{A, U}	−4
{G, U}	−1

FIGURE 1. Clote and Backofen Scoring of Base Pairings as Reported by Poolsap et al. in their 2009 paper

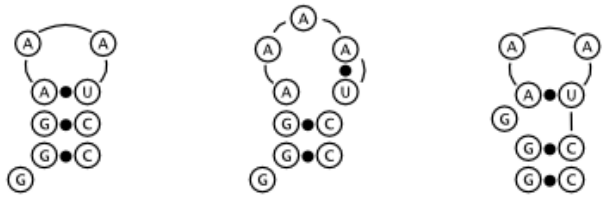
Quantifying their relative stability in function of molecular factors such as hydrogen bonding, G-C base pairings display the lowest free energy value indicating that its integration in our novel minimization framework would lead to a more stable and thus favorable structure. Following the four classical NJ cases dictating the assignment of scores to the entries of the DP table W, the third case to be considered for minimization will utilize an energy function $e(i,j)$ implementing the aforementioned scoring scheme as follows:

2.4 Binding Site Awareness

Now aware of the catalytic and regulatory roles of many documented RNAs, notably that of long-non coding RNAs (LNCRNA) experimentally associated to disease [9], resolving the structure function relationship of these molecules has gained significant interest. Indeed, creating and democratizing computational tools that could allow us to better understand

$$W(i,j) = \min \begin{cases} W(i+1,j), \\ W(i,j-1), \\ W(i+1,j-1) + e(i,j), \\ \min\limits_{i \le k < j} \{W(i,k) + W(k+1,j)\}, \end{cases}$$

FIGURE 2. Energy Sensitive Variation of the NJ Algorithm as Illustrated by Poolsap et al. in their 2009 paper



(a) (b) (c)

FIGURE 3. Illustrations[10] of three predicted RNA secondary structures generated by Nussinov's Algorithm, sharing an identical number of base pairs while displaying varying stabilities. (a) This secondary structure represents the most stable conformation among the three possibilities, characterized by an extended stem and a loop length of three base pairs. (b) This secondary structure exhibits base pairing between adjacent nucleotides, rendering it the least stable conformation among the presented possibilities. Such adjacent base pairings necessitate a highly unstable bending of the bases[11].(c) In this structure, a small bulge is introduced within the stem loop, decreasing its overall stability

RNA-folding could help the scientific community develop new therapeutic approaches in an era where, for example, RNA vaccines have already proven to be life-saving.

That is why we sought to make our NJ variants optionally aware of known binding sites in its retrieval of optimal secondary structure. To achieve this, we initiated a process wherein, given a sequence representing the binding site of an RNA, we systematically searched for a match within the RNA sequence. Given our knowledge that the binding site precisely corresponds to a sub-sequence of the RNA, we employed a window-sliding method with a length equivalent to that of the binding site sequence to search for a perfect match. This approach enabled us to extract the indices of the RNA that align with the binding site nucleotides.

Using the derived list of indices, we adapt the function to populate the Nussinov matrix. Specifically, when evaluating a nucleotide index that falls within the list corresponding to the binding site, we assign it a value of zero Thus, during trace-back, we prevent the recording of base pairs which are part of specified binding sites. While our most performant NJ variant is also capable of individually predicting secondary structure of the sense and antisense sequences of RNA pairs used in Poolsap et al. 2009 paper [3], it can also predict the secondary structure of any other single stranded RNA molecule with or without known binding sites.

2.5 Alternative Trace Back

A notable limitation of the Nussinov algorithm is its inability to distinguish between structures that have the same number of base pairs. Figure 3 illustrates that structures with equal base pair counts may differ in stability based on the arrangement of these pairings. Indeed, past research has demonstrated that the stability of a stem-loop structure depends largely on helix and loop lengths[4]. Longer stems generally contribute to higher stability, while single loops and bulges are preferred over split ones[4].

To overcome these limitations, we introduced an alternative traceback algorithm to Nussinov. This new iterative approach prioritizes the extension of base pairs and loops whenever possible. Similarly to the original traceback, we maintained the starting point and traced back through pairs that had the most significant contribution to our RNA structure, adding them to a stack.

In cases where multiple pairs were possible, rather than arbitrarily choosing one for traceback, we stored all potential pairs in a candidate stack. Within the candidate stack, we selected the pair with the same direction as the last traced-back pair. Here, a direction refers to the position within the matrix of the last visited entry and is categorized as diagonal, left, or down. Each iteration updates the direction based on the matrix recursion rule that led to the value of the last visited entry. Figure 4 provides an illustrative example of this process, and Appendix A (Algorithm 1) outlines the pseudocode for this traceback method.

3. Dataset

In the evaluation of our enhanced Nussinov algorithms, designed to predict RNA secondary structures more accurately, we meticulously selected datasets that would robustly test and validate our models. Recognizing the importance of RNA-binding sites, energy pair interactions, and an innovative traceback approach, our focus was on RNA molecules known for their interactive capabilities.

4. Pre-processing

To benchmark our models against each other and the traditional Nussinov algorithm, we sourced 14 distinct RNA sequences: Tar 16, Tar 16* [12], R1inv, R2inv [13], DIS [14], CopA, CopT [15], ATP Sensitive Ribozyme [16], lncRNA54, RepZ [17], RyhB, SodB [18], OxyS, and fhlA [15]. These sequences, obtained from various scientific articles, were manually annotated

	G	G	G	A	A	A	U	C	C	
	0	0	0	0	0	0	-4	-9	-14	G
	0	0	0	0	0	0	-4	-9	-14	G
		0	0	0	0	0	-4	-9	-9	G
			0	0	0	0	-4	-4	-4	A
				0	0	0	-4	-4	-4	A
					0	0	-4	-4	-4	A
						0	0	0	0	U
							0	0	0	C
								0	0	C

FIGURE 4. Traceback results of Nussinov's approach and our alternative traceback method for the sequence *GGGAAAUCC*, where Nussinov's algorithm was initially applied with an energy model. The path taken by the classic Nussinov traceback is shown in yellow, while the path from our alternate traceback is in red. The segments for which both algorithms overlap are shown in green.

to ensure precision in our computational analysis. The annotation process involved detailing the RNA molecules' explicit sequence, their binding sites, and their observed secondary structures in both dot bracket notation, also known as Vienna notation [19], and in sets of tuples representing all base pairs of the structure. This comprehensive annotation was crucial in providing a robust dataset for testing our models. This rigorous selection and preparation of RNA datasets were pivotal in providing a comprehensive and accurate assessment of our algorithm's performance in predicting RNA secondary structures, especially in the context of RNA-RNA interactions.

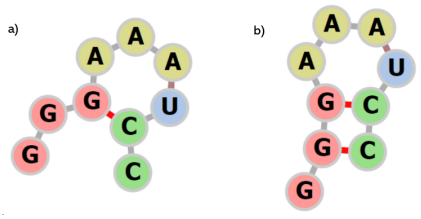


FIGURE 5. Vienna-notation-generated graphs of two predicted structures by the NJMLL model where **a)** was specified restrictive binding conditions for the first and last nucleotide of the input RNA sequence VS **b)** when all nucleotides are made available for base pairing.

5. Results

In this following first table of the report, we have compiled the prediction accuracies of three variant models, two of which have been implemented by us (i.e. the BS&E model & BS&E+AltTB model), in order to perform a benchmarking of algorithms designed to predict RNA secondary structure. The NJMLL model along with our two aforementioned novel approaches are all restricted by an MLL condition.

The accuracy was computed using the following equation:

$$\text{Prediction Accuracy} = \frac{\text{\# of Predicted Base Pairs}}{\text{\# of Observed Base Pairs}}$$

As detailed in the pre-processing section, the identified base pairs align with those extracted from figures in the literature. Each base pair has been represented as a tuple, with the elements denoting the indices of the involved nucleotides.

TABLE 1. Prediction Accuracy of RNA Secondary Structures for Each Base and Variant Models using Constant Minimal Loop Length of 5.

RNA	NJ	NJMLL	BS&E	BS&E+AltTB
Tar 16	0	0.6	0.6	1.0
Tar 16*	0.8	0.8	0.8	1.0
R1inv	0.429	0.857	1.0	1.0
R2inv	0	0.833	1.0	1.0
DIS	0	0.818	0.818	0.909
CopA	0.231	0.769	1.0	0.923
CopT	0	0.917	0.917	1.0
ATP-SR	0	0.6	0.6	0.6
lncRNA54	0.818	1.0	1.0	1.0
RepZ	0.636	0.909	0.636	0.682
OxyS	0.158	0.737	0.789	0.737
fhlA	0	0.548	0.774	0.871
Mean	**0.256**	**0.782**	**0.828**	**0.894**

5.1 The NJ and NJMLL models

With an average prediction accuracy of 25.6% for all RNAs of our test set, the basic NJ algorithm [5] [6] obtains significantly lower performance to that of our binding site and energy aware variant model (BS&E). In fact, when allowing the specification of MLL, the basic NJ algorithm with minimal loop awareness (NJMLL) [6] generates a threefold improvement in mean prediction accuracy going from 25.6 to 78.2 percent. Where the basic NJ algorithm has completely failed at predicting any base pairs part of the observed test RNA structures, its minimal loop

aware homolog, NJMLL, has always improved, notably bringing prediction accuracies of 0 to 83.3 or even 91.7 percent for the R2inv and CopT RNAs respectively.

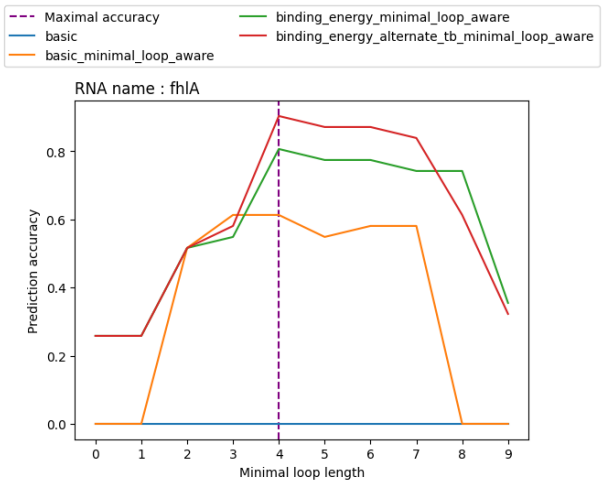


FIGURE 6. Minimal loop length optimization on fhlA RNA. The optimal minimum loop length is marked by the purple dotted line, at a value of four.

To facilitate the reproducibility of our results using Python, we offer potential future users a way to iteratively optimize accuracy in function of minimal loop length and visualize the resulting trends in order to select for any given model an optimal choice of hyperparameters as demonstrated in the following Figure 6 and 9. When performed on all test RNAs, the first minimal loop lengths to generate the maximal accuracy in one of four benchmarked models were compiled and are now presented in Table 2.

TABLE 2. First minimal loop length in range 0 to $\sqrt{n}$, where n is the length of the RNA sequence, in which maximal structure prediction accuracy was obtained.

	RNA Seq.	Optimal Minimal Loop	
Tar 16	0	CopT	3
Tar 16*	0	ATP Sensitive Rib.	5
R1inv	0	lncRNA54	4
R2inv	0	RepZ	5
DIS	3	OxyS	5
CopA	3	fhlA	4

5.2 The BS&E model

The binding site aware and energy sensitive NJ variant with specifiable MLL, which we named BS&E for short, displays even better performance than its NJMLL predecessor with a 4.6 percent increase in mean accuracy. Thanks to a minimization oriented trace-back and a scoring scheme emulating the varying effects of base pairings on free energy, BS&E resolves with greater accuracy secondary structures which have been naturally observed in experiments. We believe that such improvements originate from the joint modeling of G-U wobble base pairing and the quantitative distinguishing of the different stabilizing effects of base pairings using the Clote and Backofen scoring scheme [8]. Except for the RepZ RNA, our first variant model has consistently outperformed both NJ and NJMLL; our hypothesized reasons for such exceptions will be elaborated in the discussion section of this report.

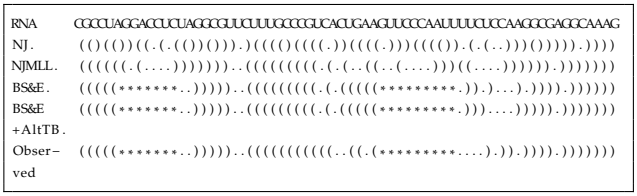


FIGURE 7. Dot notation representations of OxyS RNA obtained from algorithmic variants of Nussinov, compared to the observed structure. The asterisks (*) represent the binding sites.

5.3 The BS&E+AltTB model

Our latest algorithm combines key features from previously discussed models. The BS&E + AltTb algorithm incorporates binding and minimal loop awareness, integrates an energy model, and employs the alternate traceback function outlined earlier. The achieved average accuracy stands at 0.894, reflecting an 8% improvement compared to the BS&E model. Table 1 illustrates that this specific algorithm attains a perfect accuracy of 1.0 for six out of twelve RNAs (Tar 16, Tar 16*, R1inv, R2inv, CopT, and lncRNA54). Notably, instances like CopA and OxyS (Fig. 7) show better accuracy with the BS&E model. Additionally, the algorithm matches the accuracy of the BS&E algorithm for lncRNA54, R1inv, R2inv, and ATP-STP. In comparison to the classical Nussinov algorithm with an accuracy of 0.256, the BS&E + AltTB algorithm achieves a 249% increase, and compared to the minimal length loop-aware Nussinov model, it marks a 14% accuracy improvement.

Regarding computational complexity, our algorithm maintains a runtime of $O(n^3)$, consistent with the Nussinov algorithm. It's worth noting that in the alternate traceback, an extra loop is introduced, iterating over the candidate stack. However, this stack has a maximum length of three, as there are only three conditions for adding a matrix entry to the candidate traceback (diagonal, left, and down). Consequently, we can generalize the runtime to $O(n^3)$.

6. Discussion

Following the results reported in Table 1 and the mean accuracies which it displays for each models, we conclude that the Nussinov-Jacobson algorithm is the least performant of the benchmarked candidates. With half of its set of predicted base pairs obtaining an accuracy of 0%, this classical DP algorithm still matches in one instance the performance of our BS&E variant model, notably for the RepZ RNA. Lacking both a minimal loop length (MLL) restriction and binding site awareness, the NJ model still seeks to maximize the number of base pairs of its predicted secondary structure which allows it, just as in the aforementioned exceptional case, to still obtain an accuracy of over 63 percent. Nevertheless, not being restricted by an MLL, the NJ algorithm might more frequently output predicted base pairs which are biologically unlikely. Indeed, we consider adjacent (i.e. where the binding nucleotides are separated in sequence by less than 4 nucleotides) base pairs more unstable than stacked ones (i.e. which form long stems) separated by long loops. [1] [4] [10] Considering all these limitations and unmodeled experimental observations, we expected the utilized Python implementation of the NJ algorithm [6] to display low accuracies.

NJMLL also demonstrated a noteworthy accuracy of 0.782, signifying an 205% improvement compared to the base model. This underscores the critical importance of implementing a minimum loop length in secondary structure prediction. Indeed, this is because a minimum stem loop size contributes to the stability of the RNA structures. For instance, RNA hairpins with loops comprising four or five nucleotides exhibit the highest stability, with a loop size of 4 being the most common. As reiterated earlier, the stem loop length holds additional biological relevance. Beyond contributing to stability, these structures are prevalent in various interacting RNA molecules, including transfer RNA, pre-microRNA, and ribozymes, where they play a pivotal role in ensuring functionality.

Recognizing the significance of minimum loop length, all subsequent models we developed incorporated a dedicated parameter to account for this constraint.

For the BS&E model, we achieve an average accuracy of 0.828. This increase compared to the previous algorithms highlights the importance of considering binding sites while predicting secondary structures. The rationale behind this enhancement lies in the fact that these binding sites establish bonds with external molecules. Consequently, it becomes pertinent to exclude them from the potential base pairings during the prediction process.

For all the models and samples in our dataset, given

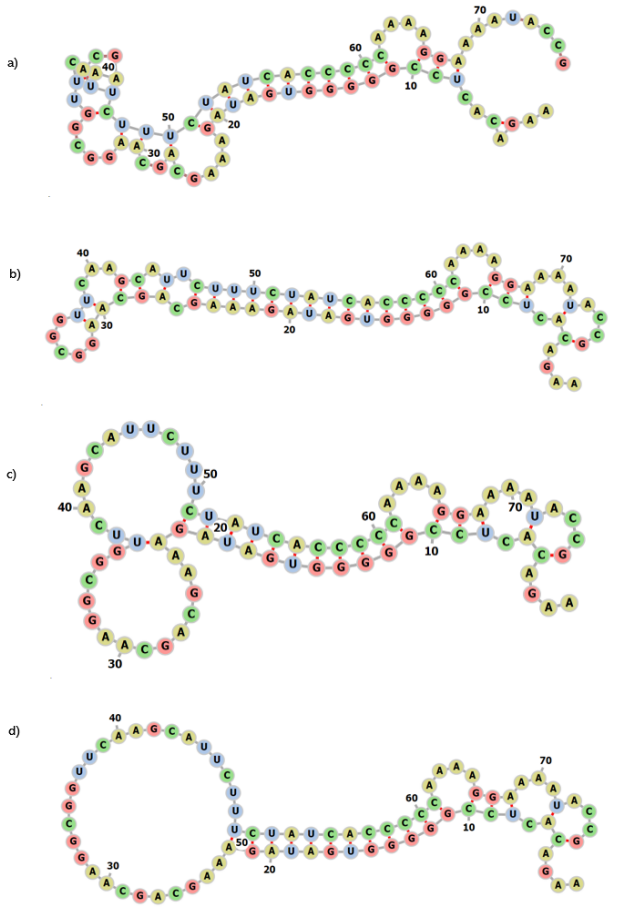


FIGURE 8. Vienna-notation-generated graphs of four predicted structures respectively predicted by **a)** the NJ model with accuracy of 0.636, **b)** the NJMLL model with accuracy of 0.909, **c)** the BS&E model with accuracy of 0.636, and **d)** the BS&E+AltTB model with accuracy of 0.682.

the nucleotide sequence of a binding site, we successfully extracted their indices within the RNA. In handling known binding sites, where we had certainty that the binding site is a subsequence of the RNA, our approach involved searching for perfect matches. However, this straightforward approach becomes more complex when confronted with a variation of the problem involving partial binding sites. In such cases, where we only possess knowledge of partial binding sites, incorporating a degree of flexibility for mismatches is essential. Instead of relying on the sliding window method, alignment algorithms such as BLAST and Smith-Waterman emerge as robust solutions for extracting this nuanced information.

Furthermore, when confronted with an even more complex scenario where the binding sites are entirely unknown, and we only have sequences of two RNA molecules known to interact, alignment algorithms once again prove to be useful. In such instances, these algorithms play a pivotal role in deciphering the interaction by aligning the sequences and highlighting potential regions of complementarity between the molecules.

Lastly, our highest performing model was the BS&E + AltTb one which achieved an accuracy of 89%. This model featured constraints that were grounded on biological knowledge, the main one being prioritizing base pairs that contribute to longer stems and longer loops. Nussinov only returns one structure, while there are many eligible structures, often with higher stability. To mediate this limitation, at each step of the traceback, we consider all possible paths, and then choose the one that aligns with the previously mentioned biological constraints. Since we see an improvement from the BS&E to this one, we can conclude that this consideration is important in predicting RNA structures.

However, we also see that for some RNAs, BS&E performed better. This includes for CopA and OxyS. LncRNA54, Tar 16, Tar16*, CopT and ATP-SR had accuracies that were equal to BS&E; it is to note that aside from ATP-SR, this is because the achieved accuracy was 100%.

For the RNAs that underperformed using the BS&E + AltTb algorithm, an analysis of the structures helped conclude some limitations of the latest model.

In the case of OxyS, noticeable bulges within the stem loop are formed by nucleotides at positions 33, 34, 56, and another set at 37 and 53. Unfortunately, our algorithm faces challenges in accurately identifying these bulges. During the traceback process, our algorithm tends to prioritize the formation of extended stem lengths and lengthy loops. Consequently, at the base of the binding site, where free nucleotides could potentially form, our algorithm tends to favor the creation of binding pairs instead. This behavior explains why the observed structure features larger loops at positions around 40-50 compared to our algorithm's predictions. In essence, when confronted with nucleotides in close proximity to binding sites, our algorithm encounters difficulty and may inadvertently add inappropriate base pairs due to its prioritization of stem elongation.

Interestingly, in the case of Repz, the performance of our binding sites aware models was outperformed by the Nussinov MLL (0.909). Refer to Figure 8, where the BS&E+AltTB model exhibits a large loop at the left extremity. However, a more precise examination of the structure (Fig. 8b) reveals that this segment more accurately resembles a conglomeration of smaller loops interconnected by short stem loops. This observation underscores the challenge in recognizing this pattern by our latest algorithm.

Furthermore, RepZ is known for its interaction with lncRNA54, such that it undergoes a conformational change upon binding to lncRNA54. Indeed, the

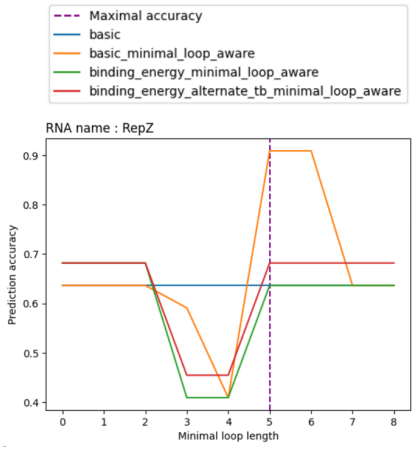


FIGURE 9. Minimal loop length optimization on the RepZ RNA. The optimal minimum loop length is marked by the purple dotted line, at a value of five.

structure at the top of the stem-loop is substantially affected. Although our algorithm is tailored to predict RNA structure independently of its binding conformation, understanding the structural implications of this interaction could potentially enhance our predictive capabilities.

Finally, additional limitations pertain to the selection of values for the energy model, minimal loop lengths, and other adjustable parameters. Zuker's algorithm, which incorporates diverse thermodynamic parameters associated with various structural motifs, serves as a promising foundation for developing a more biologically relevant scoring scheme. Regarding minimal loop lengths, our algorithmic pipeline was executed with a minimal loop of 5, aligning with the maximum optimal minimal loop observed. However, a more precise value would be 4, representing the average of these measurements. While we speculate that the impact on accuracy would be minimal, it is worth noting. An improved approach would have involved running the algorithms with the optimal minimal loop length determined for each RNA, potentially reducing the occurrence of the large loops observed in our predictions.

7. Future Direction

An obvious next step would be to consider the potential formation of pseudoknots in our variant algorithms although they are known to be less likely to form in natural RNA secondary structures. [20] To consider structures harboring pseudoknots might drastically increase both the search space and computational cost of our algorithm and thus reduce the usability of the presented variant models; if such consideration were to be added to them, a penality proportionnal to the natural rarity of pseudoknots will have to be included in our future models' scoring scheme to maintain high

prediction accuracy. [20]

When RNAs fold, the individual stability of each of its base pairs are not the only factors contributing to the overall free energy of the molecule. Modeling secondary structure in two dimensions, using biologically realistic ranges of bond angles, the possible distances between different parts of the RNA molecule (e.g. a hairpin with a neighboring flexible linear sequence that can interact with it) could also hint at the overall stability of the molecule. Compact structures, in comparison to more *airy* ones, could undergo more internal steric hindrance and could thus present itself as a less than conformationally optimal structure.

In comparison to Poolsap et al.'s [3] model predictions of the joint structure of known interacting sense and antisense RNA pairs, our novel variant models do not resolve secondary structures in the midst of interaction, but rather individually pre-interaction. Thus, our model does not consider the potential conformational changes which an RNA molecule can undergo upon interaction with other molecules at its binding site(s). We think that perhaps, an additional modeling of steric hindrance in function of in-sequence distance of unbound nucleotides with respect to those of the binding site could also help in further minimizing the free energy of predicted secondary structures. In essence, the closer a nucleotide is to the known binding site of the RNA sequence it belongs to, the less likely it is to bind (i.e. form base pairs) with other nucleotides of its own structure in interacting conformation due to hightened steric hindrance upon occupation of the binding site.

8. Conclusion

In conclusion, this study presents significant advancements in RNA secondary structure prediction through the development of enhanced Nussinov algorithms. Our findings demonstrate that the incorporation of energy sensitivity, binding site awareness, and minimal loop length constraints significantly improves prediction accuracy. The BS&E and BS&E+AltTB models, in particular, have shown remarkable performance, outperforming the classic Nussinov-Jacobson algorithm. These advancements not only contribute to the theoretical understanding of RNA folding mechanisms but also provide practical tools for molecular biology research. Future work could explore the integration of pseudoknot prediction and more complex RNA interactions, further enriching the field of RNA structural analysis.

9. Contributions

Sophie Cao implemented the novel BS&E and BS&E+AltTB variant models, wrote sections 1, 2.2, 2.4, 2.5, 5.3 and 6 of the report along with the supplemental material of the appendix, while also generating figures 4, 6 and 7. Yanis Bencheikh contributed to the implementation of variant models, MLL optimization framework, and pipeline used to run the algorithms. Yanis also wrote the abstract and sections 2.1, 2.3, 2.4, 4, 5.1, 5.2, 6 and 7 while also generating figures 5, 8, 9. Vitoria Lara performed the literature review and annotation of the RNA sequences which composed our test dataset used for the benchmarking of the four models presented in this report. Vitoria wrote sections 3, 4 and 8 of this report while also generating table 1 and 2.

References

[1] Schudoma, Christian *et al.*: *Sequence-structure relationships in rna loops: establishing the basis for loop homology modeling*. Nucleic Acids Research, 38(3):970–980, 2010.

[2] Andronescu, Mirela *et al.*: *Rnasoft: A suite of rna secondary structure prediction and design software tools*. Nucleic Acids Research, 31(13):3416–3422, 2003.

[3] POOLSAP, UNYANEE, YUKI KATO, and TATSUYA AKUTSU: *Dynamic programming algorithms for rna structure prediction with binding sites*. Biocomputing 2010, page 98–107, 2009.

[4] Ma, Hairong, David J. Proctor, Elzbieta Kierzek, Ryszard Kierzek, Philip C. Bevilacqua, and Martin Gruebele: *Exploring the energy landscape of a small rna hairpin*. Journal of the American Chemical Society, 128(5):1523–1530, 2006. https://doi.org/10.1021/ja0553856, PMID: 16448122.

[5] Nussinov, R and A B Jacobson: *Fast algorithm for predicting the secondary structure of single-stranded rna.* Proceedings of the National Academy of Sciences, 77(11):6309–6313, 1980.

[6] *Nussinov algorithm to predict secondary RNA fold structures &x2013; Bayesian Neuron — bayesianneuron.com*. https://bayesianneuron.com/2019/02/nussinov-predict-2nd-rna-fold-structure-algorithm/. [Accessed 30-11-2023].

[7] Varani, Gabriele and William H McClain: *The g·u wobble base pair*. EMBO reports, 1(1):18–23, 2000.

[8] Clote, P. and R. Backofen: *Computational Molecular Biology An Introduction*. John Wiley amp; Sons, 2000.

[9] Cui, Qinghua: *Lncrnadisease tutorial*. https://www.cuilab.cn/lncrnadisease.

[10] *An Alternative Traceback Method for Nussinov's RNA Folding Algorithm — datamech.com*. http://datamech.com/devan/nussinov-traceback.html. [Accessed 30-11-2023].

[11] Durbin, Richard, Sean Eddy, Anders Krogh, and Geoffrey Mitchinson: *Biological Sequence Analysis*. Cambridge University Press, New York, 1998.

[12] Chang, Kung Yao and Ignacio Tinoco: *The structure of an rna "kissing" hairpin complex of the hiv tar hairpin loop and its complement*. Journal of Molecular Biology, 269(1):52–66, 1997.

[13] Rist, M.: *Association of an rna kissing complex analyzed using 2-aminopurine fluorescence*. Nucleic Acids Research, 29(11):2401–2408, 2001.

[14] Paillart, J C, E Skripkin, B Ehresmann, C Ehresmann, and R Marquet: *A loop-loop "kissing" complex is the essential part of the dimer linkage of genomic hiv-1 rna.* Proceedings of the National Academy of Sciences, 93(11):5572–5577, 1996.

[15] Wagner, E.Gerhart H and Klas Flärdh: *Antisense rnas everywhere?* Trends in Genetics, 18(5):223–226, 2002.

[16] Tang, J: *Mechanism for allosteric inhibition of an atp-sensitive ribozyme*. Nucleic Acids Research, 26(18):4214–4221, 1998.

[17] Asano, Katsura and Kiyoshi Mizobuchi: *Structural analysis of late intermediate complex formed between plasmid colib-p9 inc rna and its target rna*. Journal of Biological Chemistry, 275(2):1269–1274, 2000.

[18] Geissmann, Thomas A and Danièle Touati: *Hfq, a new chaperoning role: Binding to messenger rna determines access for small rna regulator*. The EMBO Journal, 23(2):396–405, 2004.

[19] Kerpedjiev, Peter: *TBI - forna: RNA Secondary Structure Visualization Using a Force Directed Graph Layout — rna.tbi.univie.ac.at*. http://rna.tbi.univie.ac.at/forna/. [Accessed 30-11-2023].

[20] Rødland, Einar Andreas: *Pseudoknots in rna secondary structures: Representation, enumeration, and prevalence*. Journal of Computational Biology, 13(6):1197–1213, 2006.

A. Supplemental Material

Algorithm 1: Alternate Traceback (Adapted from Lai et al.[10])

Data: Matrix nm, RNA sequence rna
Result: List of base pairs $fold$
$n \leftarrow$ length of rna;
$fold \leftarrow []$;
$tb\_stack \leftarrow [(0, n-1)]$;
$candidate\_stack \leftarrow []$;
$last\_direction \leftarrow$ None;
while *tb_stack is not empty* **do**
  $candidate\_stack \leftarrow []$;
  $(i, j) \leftarrow$ pop from tb_stack;
  if $i < j$ **then**
    if $nm[i][j] = nm[i+1][j-1] + energy((rna[i], rna[j]))$ ***and*** $couple((rna[i], rna[j]))$ **then**
      append $("DIAGONAL", i+1, j-1)$ to $candidate\_stack$;
    if $nm[i][j] = nm[i][j-1]$ **then**
      append $("LEFT", i, j-1)$ to $candidate\_stack$;
    if $nm[i][j] = nm[i+1][j]$ **then**
      append $("DOWN", i+1, j)$ to $candidate\_stack$;
    if ***not*** $candidate\_stack$ **then**
      $last\_direction \leftarrow$ None;
      for $k \leftarrow i+1$ **to** $j-1$ **do**
        if $nm[i][j] = nm[i][k] + nm[k+1][j]$ **then**
          append (i, k) to tb_stack;
          append $(k+1, j)$ to tb_stack;
          break;
    else
      $added \leftarrow$ False;
      while $candidate\_stack$ **do**
        $(d, m, n) \leftarrow$ pop from $candidate\_stack$;
        if ***not*** $last\_direction$ **then**
          $last\_direction \leftarrow d$;
        if $d = last\_direction$ **then**
          append (m, n) to tb_stack;
          $added \leftarrow$ True;
          if $d = "DIAGONAL"$ **then**
            append (i, j) to $fold$;
          break;
      if ***not*** $added$ **then**
        append (m, n) to tb_stack;
        $last\_direction \leftarrow d$;
        $added \leftarrow$ True;
        if $d = "DIAGONAL"$ **then**
          append (i, j) to $fold$;
return $fold$;