applied sciences

MDPI

Article

Syntactic Pattern Recognition for the Prediction of L-Type Pseudoknots in RNA

Christos Koroulis [1,†], Evangelos Makris [1,†], Angelos Kolaitis [1], Panayiotis Tsanakas [1] and Christos Pavlatos [2,*,†]

[1] School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou St., 15780 Athens, Greece; el17159@mail.ntua.gr (C.K.); vmakris@mail.ntua.gr (E.M.)
[2] Hellenic Air Force Academy, Dekelia Air Base, Acharnes, 13671 Athens, Greece
* Correspondence: christos.pavlatos@hafa.haf.gr; Tel.: +30-210-7722541
† These authors contributed equally to this work.

Abstract: The observation and analysis of RNA molecules have proved crucial for the understanding of various processes in nature. Scientists have mined knowledge and drawn conclusions using experimental methods for decades. Leveraging advanced computational methods in recent years has led to fast and more accurate results in all areas of interest. One highly challenging task, in terms of RNA analysis, is the prediction of its structure, which provides valuable information about how it transforms and operates numerous significant tasks in organisms. In this paper, we focus on the prediction of the 2-D or secondary structure of RNA, specifically, on a rare but yet complex type of pseudoknot, the L-type pseudoknot, extending our previous framework specialized for H-type pseudoknots. We propose a grammar-based framework that predicts all possible L-type pseudoknots of a sequence in a reasonable response time, leveraging also the advantages of core biological principles, such as maximum base pairs and minimum free energy. In order to evaluate the effectiveness of our methodology, we assessed four performance metrics: precision; recall; Matthews correlation coefficient (MCC); and F1-score, which is the harmonic mean of precision and recall. Our methodology outperformed the other three well known methods in terms of Precision, with a score of 0.844, while other methodologies scored 0.500, 0.333, and 0.308. Regarding the F1-score, our platform scored 0.671, while other methodologies scored 0.661, 0.449, and 0.449. The proposed methodology surpassed all methods in terms of the MCC metric, achieving a score of 0.521. The proposed method was added to our RNA toolset, which aims to enhance the capabilities of biologists in the prediction of RNA motifs, including pseudoknots, and holds the potential to be applied in a multitude of biological domains, including gene therapy, drug design, and comprehending RNA functionality. Furthermore, the suggested approach can be employed in conjunction with other methodologies to enhance the precision of RNA structure prediction.

Keywords: syntactic pattern recognition; context-free grammar; RNA; L-type pseudoknot

Citation: Koroulis, C.; Makris, E.; Kolaitis, A.; Tsanakas, P.; Pavlatos, C. Syntactic Pattern Recognition for the Prediction of L-Type Pseudoknots in RNA. *Appl. Sci.* **2023**, *13*, 5168. https://doi.org/10.3390/app13085168

Academic Editor: Jorge Igual

Received: 10 March 2023
Revised: 18 April 2023
Accepted: 19 April 2023
Published: 21 April 2023

1. Introduction

Many studies have been conducted through the past decades focusing on structural bioinformatics. The RNA molecule is crucial in protein expression, regulating gene expression, and other vital functions. Hence, predicting RNA's 3-D or tertiary structure is a pivotal task in the domain. Tasks such as the relationship between this 3-D folding and its biological functions, the detection of similar characteristics, and the understanding of dynamics, are of significant importance. All the above-mentioned challenging tasks are closely related to the 3-D structure, which is rather difficult to predict accurately. Considering this impediment, with the aim of providing an efficient and performant toolset to experts, our research focuses on the prediction of the 2-D or secondary structure of RNA as an intermediate step. Complex motifs, such as pseudoknots, can be predicted in

2-D, making them available to scientists for initial conclusions and rapid decision making. Thereafter, it is easier to map this secondary structure to the 3-D space for further analysis and experimentation.

Various motifs, such as stems, hairpins, bulges, interior loops, and multi-branch loops, can be found in the RNA secondary structure, and there are several methodologies that can accurately predict these motifs using dynamic programming, thermodynamic models, stochastic methods, artificial intelligence, syntactic pattern recognition techniques, or a combination of these. Section 3 provides a detailed overview of the relevant research, including the state-of-the-art works. However, predicting a pseudoknot is a more complex task, and only a few existing algorithms are designed to handle the interconnection of pseudoknots. In that context, we introduced, in our previous research, two robust frameworks that are capable of predicting H-type pseudoknots, the most common type of this motif. Firstly, we presented Knotify [1] and an optimized version of this framework in terms of execution time, which predict H-type pseudoknots using syntactic pattern recognition, minimum free energy, and a parallel architecture to achieve high performance. These approaches show a high accuracy level in terms of core stems prediction and total base pair prediction, better than or comparable with state-of-the-art frameworks, depending on the metrics evaluated. Secondly, we enhanced the previous implementation by creating Knotify+ [2], which is capable of predicting more complex motifs of H-type pseudoknots that include bulges. Knotify+ maintains execution speed at a low level, contrary to the most well-known platforms, and has, on average, a level of accuracy equal to that of Knotify. In our endeavor to help biologists, in this work, we provide a framework that predicts a very rare and rather complex type of pseudoknot, the L-type. The methodology retains all the valuable and optimized parts of our previous work, and encapsulates a context-free grammar that is able to reveal these complex L-type motifs.

The paper is organized with the following structure. Section 2 offers the required theoretical background, while Section 3 covers the relevant research. Section 4 introduces the proposed L-type grammar, accompanied by an illustrative example. The methodology's implementation details are explored in Section 5, and performance evaluation is analyzed in Section 6. Section 7 provides a discussion of the proposed platform's performance evaluation. Finally, Section 8 presents the study's conclusions, as well as potential areas for improvement and expansion.

2. Conceptual Framework

In this section, we examine the basic theoretical concepts of the pseudoknot structure and parser implementations to provide the reader with a fundamental understanding. The nitrogenous bases A, C, G, and U (adenine, cytosine, guanine, and uracil), sugars, and a phosphate backbone are combined to form RNA. RNA plays a crucial role in many biological functions, such as carrying genetic information, regulating gene expression, and transcribing the genomes of mammals. To carry out these tasks, RNA forms various bonds and creates a set of nitrogenous base pairs, including the Watson–Crick [3] pairs (A–U and G–C) and the less common wobble base pair G–U. The arrangement of these base pairs, along with the unpaired regions, is known as the secondary structure, and is important for several processes.

2.1. The Pseudoknot Motif

Pseudoknots are of significance in this research, and can be found in different organisms. They consist of two helices that are connected by one or more single-stranded sections, known as loops. Although pseudoknots are a rare pattern in RNA sequences, they are challenging to predict due to the structure of the two intersecting base pairs. The initial identification of this pattern occurred in the Turnip Yellow Mosaic virus [4]. Pseudoknots come in various forms [5], such as H, K, L, and M. The H-type pseudoknot [6] comprises two stems and two loops with flexible lengths. Its creation results from the intersection of a pair of base pairs, or core stems, as per our notation. The current research is centered

specifically on L-type pseudoknots [5], which can form when a stem has its two bases situated in each of the two loops of an H-type pseudoknot.

2.2. Syntactic Pattern Recognition—The Core Concept

The suggested framework for pseudoknot prediction in RNA is based on recognizing syntactic patterns. This approach involves defining a language using a set of syntax rules that can generate strings belonging to that language [7]. These syntax rules are part of a grammar that dictates how precise sequences of symbols, constituting the defined language, can be generated [7]. According to Noam Chomsky's hierarchy [8], all grammars can be classified into one of four categories. Context-free grammar (CFG) is one such category, which is commonly utilized for implementing programming languages and recognizing human languages [9].

Context-Free Grammars

CFGs are mathematical models used in formal language theory and computer science for defining and generating formal languages. A CFG consists of a set of production rules that specify how strings of symbols can be generated from a given starting symbol, known as the axiom or start symbol. The production rules define how symbols can be transformed into other symbols, including themselves, until a string that belongs to the language defined by the grammar is obtained.

In a CFG, symbols can be either terminals or non-terminals. Terminals are the actual symbols that appear in the strings of the language, while non-terminals are symbols used to generate the terminals through the application of the production rules. The production rules in a CFG must satisfy certain restrictions, such as being context-free, meaning that the rule can be applied regardless of the context in which the non-terminal appears. This is in contrast to context-sensitive grammars, where the rule for a non-terminal may depend on the context in which it appears.

Consequently, the definition of a CFG is a set of four components: $\langle NT, T, R, S \rangle$. The start symbol S (which is also known as the root of the grammar) is part of the set of non-terminal symbols NT. The set T contains all terminal symbols, while NT includes all non-terminal symbols. The set R holds all the production rules. The production rules are written in the form $D \rightarrow \gamma$, where D is a non-terminal symbol from NT, and γ is a string made up of symbols from T and/or NT.

The proposed methodology for predicting pseudoknots in RNA structures is founded on recognizing syntactic patterns. This methodology requires the formulation of a language using a set of syntax regulations that can generate strings that are part of that language [7]. These syntax rules form a grammar that specifies how specific sequences of symbols that make up the defined language can be generated [7]. Noam Chomsky's hierarchy [8] divides all grammars into four categories, with context-free grammars (CFGs) being one of them. CFG is often used for programming language implementation and recognition of human languages [9].

CFGs are widely used in computer science for tasks such as parsing and compiling computer programs, recognizing and generating human language, and for modeling RNA structures, among others. The concept of a CFG and its production rules can be extended to more powerful models, such as context-sensitive grammars and unrestricted grammars, allowing for the definition and generation of more complex languages.

Many parsing algorithms have been suggested for CFG grammars because of their significant expressive capabilities. Two widely used parsing algorithms are the CYK parser [10] and the Earley parser [11]. There are also several extensions [12–14] and parallel versions [15,16] of these two algorithms in the literature. The proposed implementation chose Earley's parser owing to its competence in handling ambiguous grammars and effectiveness. The implementation relies on the Yet Another Earley Parser (YAEP) [17], a proficient Earley parser able to parse ambiguous grammars with efficiency.

3. Related Work

To the best of the authors' knowledge, this is the first attempt at a platform that incorporates a dedicated module to handle the prediction of RNA L-type pseudoknots. The proposed methodology aims to address this identified research gap. While there are several computational platforms available that can efficiently predict RNA pseudoknots, they mainly focus on H-type pseudoknots. Therefore, this section aims to present the proposed platforms in the literature, even if they are only capable of predicting H-type pseudoknots.

Most RNA secondary structure prediction algorithms utilize dynamic programming as a core concept of their methodologies, attempting to calculate the structure with the lowest free energy. In the special case of pseudoknot prediction, in addition to the minimum free energy, the proposed frameworks also consider factors such as stability and entropy in their calculation pipeline, as in [18]. Despite the fact that this problem has been proven to be NP (nondeterministic polynomial time)-complete [19], researchers have developed stochastic and heuristic approaches [20–22]. Knotty [23], for example, is a highly efficient framework for pseudoknot prediction, which uses a CCJ (Chen–Condon–Jabbari) algorithm [24] with sparsification. ProbKnot [25] predicts the secondary structure of an RNA molecule by combining base pair probabilities of non-pseudoknotted structures and maximum expected accuracy. IPknot [26] and its extension [27] also take advantage of base pair probabilities, but they leverage integer programming and the LinearPartition model, combined with pseudo-expected accuracy, for further optimization in terms of accuracy.

Various machine learning algorithms have also been proposed in the literature. These algorithms aim to identify underlined patterns in training datasets through supervised and unsupervised methods. Most of them rely on deep learning techniques, as in [28], where a deep learning method is employed with tertiary constraints, while in [29], a bidirectional LSTM network combined with IBPMP to select the correct base pairs and predict the optimal structure, was proposed. In 2dRNA [30], a bidirectional LSTM encodes the data, which is then decoded by a fully connected network to produce the dot–bracket structure. The ATTfold method, capable of predicting pseudoknots [31], incorporates deep learning models with an attention mechanism as an encoder. The base pairing score matrix is encoded and then decoded by a convolutional neural network (CNN) into an appropriate format. UFold [32] is also a deep learning framework, and is trained directly on annotated data and base pairing rules. It utilizes an image-like representation of sequences, appropriate for processing by fully convolutional networks (FCNs).

In addition to the thermodynamic models, there are also implementations using stochastic context-free grammar (SCFG). Their accuracy varies depending on the SCFG chosen. Pfold [33,34], for example, receives RNA alignments as inputs and predicts a secondary structure. A multi-threaded version of Pfold, called PPfold [35], has also been developed for execution time optimization. RNA-Decoder [36] also predicts secondary structure using an SCFG, considering the known protein coding context of RNAs. Similar SCFG-based approaches include Contrafold [37], Evfold [38], Infernal [39], Oxfold [40], and Stemloc [41]. Previous studies have demonstrated the possibility of translating Zuker's thermodynamic model into an SCFG by calculating production probabilities from thermodynamic constants [22]. These techniques all strive to optimize an objective function, with thermodynamic methods endeavoring to minimize the free energy of a structure, deep learning approaches optimizing their loss function by adjusting their weights and trainable parameters, while SCFG-based methods aim to maximize the corresponding probability of a structure. The abundance of research on SCFG-based methods emphasizes the necessity for the effective amalgamation of all the above-mentioned methods. Therefore, it is crucial to discover the most suitable combination of these concepts to address the prediction of RNA secondary structures. To this end, we propose a grammar-centered framework for L-type pseudoknots, enhancing our existing set of tools for RNA secondary structure prediction.

4. The Proposed Methodology—A Demonstration

This section outlines the proposed methodology, which builds upon the Knotify and Knotify+ platforms introduced in [1,2], respectively, and includes the pruning technique from [42]. The proposed platform is an extension of Knotify, and is designed to predict L-type pseudoknots in RNA sequences. The Knotify platform consists of three main tasks, wherein the RNA sequence is first analyzed by a CFG parser that generates trees containing a pseudoknot pattern. The generated trees are then parsed to identify the core stems forming the pseudoknot and the potential base pairs around them. Finally, the optimal tree is selected using two established criteria: the maximum number of base pairs and the minimum free energy of the sequence. In this paper, we introduce a CFG that is integrated into the first task of the Knotify platform, enabling it to predict L-type pseudoknots. The proposed implementation takes the primary structure of a sequence—a string—as the input, and outputs the secondary structure in extended dot–bracket notation. The platform comprises several software modules, with each task being implemented through a separate module. In Section 5, we provide explicit implementation details and a thorough analysis of each task.

4.1. Proposed CFG to Detect L-Type Pseudoknots

The proposed approach for detecting L-type pseudoknots in RNA sequences relies on syntactic pattern recognition techniques and a CFG parser. The selection of appropriate primitive patterns is emphasized, as it plays a significant role in accurate recognition. In RNA recognition, the nitrogenous bases adenine (“A”), cytosine (“C”), guanine (“G”), and uracil (“U”) typically form the RNA representation. Hence, the suggested grammar’s vocabulary includes these four terminal symbols, and any RNA sequence can be linguistically represented as a string of these symbols, such as “UCACAACGAACCU”.

In order to recognize a given pattern syntactically, an appropriate pattern grammar is used to parse the linguistic representation of the original pattern. The pattern grammar’s design is critical in achieving accurate recognition, as it can significantly impact the results. Therefore, forming an efficient CFG to describe the pseudoknot syntactically is crucial. CFGs are widely known to be suitable for representing structural features, and in this study, the $G_{Lpseudo}$ presented in Table 1 is used for the task of L-type pseudoknot prediction.

Table 1. Syntactic rules $G_{Lpseudo}$.

Enumeration	Syntactic Rules
0	S → “A” L “A” L “A” D “U” L “U” L “U”
1	S → “A” L “A” L “U” D “U” L “U” L “A”
2	S → “A” L “A” L “G” D “U” L “U” L “C”
3	S → “A” L “A” L “C” D “U” L “U” L “G”
4	S → “A” L “U” L “A” D “U” L “A” L “U”
5	S → “A” L “U” L “U” D “U” L “A” L “A”
6	S → “A” L “U” L “G” D “U” L “A” L “C”
7	S → “A” L “U” L “C” D “U” L “A” L “G”
8	S → “A” L “G” L “A” D “U” L “C” L “U”
9	S → “A” L “G” L “U” D “U” L “C” L “A”
10	S → “A” L “G” L “G” D “U” L “C” L “C”
11	S → “A” L “G” L “C” D “U” L “C” L “G”
	.
	.
	.
60	S → “C” L “A” L “C” D “G” L “G” L “U”
61	S → “C” L “U” L “C” D “G” L “G” L “A”
62	S → “C” L “G” L “C” D “G” L “G” L “C”
63	S → “C” L “C” L “C” D “G” L “G” L “G”

Table 1. *Cont.*

Enumeration	Syntactic Rules
64	L → "A" L
65	L → "U" L
66	L → "C" L
67	L → "G" L
68	L → "A"
69	L → "U"
70	L → "C"
71	L → "G"
72	D → K K
73	K → "A"
74	K → "U"
75	K → "C"
76	K → "G"
77	K → ϵ

In the $G_{Lpseudo}$ grammar presented in Table 1, the second column displays the syntactic rules. The grammar comprises four non-terminal symbols, namely, NT = S, L, D, K, with S as the start symbol. All the syntactic rules with S on the left-hand side (i.e., rules 0 to 63) aim to identify a potential L-type pseudoknot in the input sequence. An L-type pseudoknot is defined as having at least three core stems, as depicted in Figure 1c. To recognize an L-type pseudoknot syntactically, an appropriate pattern grammar must be employed to parse the linguistic representation of the RNA sequence. The $G_{Lpseudo}$ CFG is designed to efficiently describe the syntax of an L-type pseudoknot within an RNA sequence, with the core stems forming the intercalated base pairs. For instance, the rule S → "U" L "G" L "G" D "A" L "C" L "C" denotes the presence of a pseudoknot of the form U..G..G..A..C..C, where the base pairs U–A, G–C, and G–C are intercalated, and the colors indicate the base pairs. These intercalating base pairs are referred to as **core stems** throughout the article, and are represented by punctuation marks "(", ")", "[", "]", "{" and "}" at the dot–bracket notation (see Tables 2 and 3). Since there are four possible base pairs (A–U, U–A, C–G, G–C) and three possible intercalations, there are sixty-four (4^3) syntax rules with S on the left-hand side, with rules 12 to 59 omitted as they are easily understood. Figure 2 provides an example of how the core stems intercalate each other to form an L-type pseudoknot.

Table 2. Locating base pairs around the core stems of a pseudoknot.

String Enumeration	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
String	U	U	U	A	C	G	G	C	C	G	A	A	U	U	C	G	C	G	G	A
Parser output:	.	.	(	.	.	[	.	.	.	{	)	.	.	.	]	.	}	.	.	.
Step 1	.	(	(	.	.	[	.	.	.	{	)	)	.	.	]	.	}	.	.	.
Step 2	.	(	(	.	.	[	.	{	{	{	)	)	.	.	]	.	}	}	}	.
Step 3	.	(	(	.	[	[	.	{	{	{	)	)	.	.	]	]	}	}	}	.

Table 3. L-type pseudoknot prediction.

Platform	**RNA/Dot Bracket**
	UGUGCGUUUCCAUUUGAGUAUCCUGAAAACGGGCUUUUCAGCAUGGAACGUACAUAUUAAAUAGUAGGAGUGCAUAGUGGCCCGUAUAGCAGGCAUUAACAUUCCUGA
Ground truth	(((((((.(((((........[[[[....[[[[....{{{{.)))))))))))).....................]]]].....]]]]..........}}}}
Knotify	.......((((((..............[[[[[[...{{{{.))))))).............................]]]]]]..................}}}}
IPknot	.((((((.(((((.[[[[...((.......))((..]]]]))))))))))))).....((((((((...(((.(((.........)))))))..)))))))).
Knotty	(((((((.((((((((((...[[[[...[[[[[[[.)))))).))))))))))))).....((((((((..((((]]]]]]]....]]]].)))))..)))))))).
ProbKnot	(((((((.((((((((((...((((....))))...))))))..)))))))))))))(((...))).((((((((..((((.(((.........)))))))..)))))))).

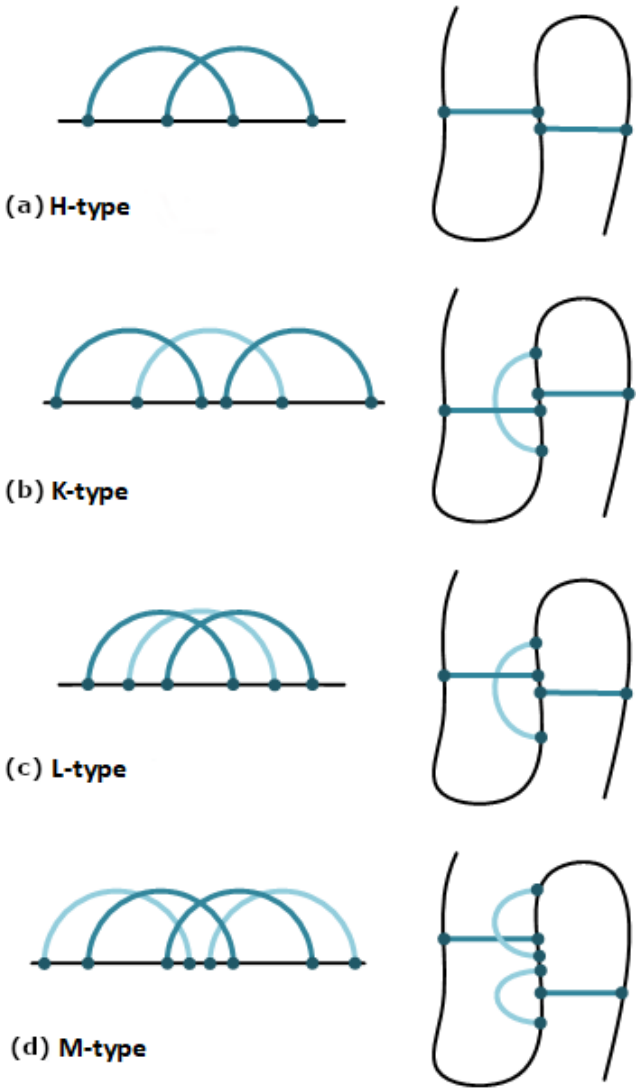

Figure 1. H-, K-, L-, and M-type pseudoknots.

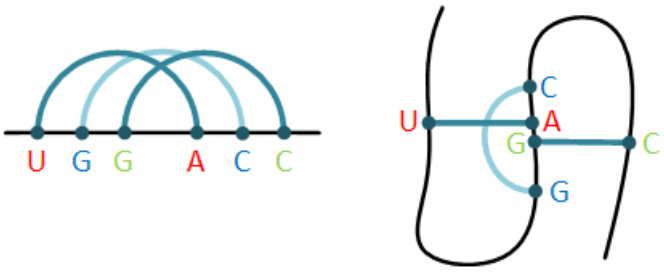

Figure 2. The rule S → "U" L "G" L "G" D "A" L "C" L "C" that detects the presence of an L-type pseudoknot.

The proposed approach for detecting L-type pseudoknots in RNA sequences involves utilizing syntactic pattern recognition techniques and a CFG parser. To represent RNA, the four standard nucleotide bases, i.e., adenine, cytosine, guanine, and uracil, are represented as single characters "A", "C", "G", and "U", respectively, and any RNA can be linguistically represented as a string of these symbols. The proposed grammar, $G_{Lpseudo}$, consists only of these four terminal symbols, and it includes four non-terminal symbols in the set NT = $\{S, L, D, K\}$.

The syntactic rules for the $G_{Lpseudo}$ grammar are presented in the second column of Table 1, and all syntactic rules that have the start symbol S on the left-hand side aspire to

identify a potential pseudoknot in the input string. The L-type pseudoknot is defined as having at least three base pairs, forming intercalated core stems. The non-terminal symbol L generates sequences of bases that make up the four interior loops of the pseudoknot, while the non-terminal symbol D generates sequences of bases between the two main crossing base pairs.

The CFG parser can identify pseudoknots in strings that have their initial and final symbols belonging to the core stems group, but it can also operate effectively on substrings using the sliding windows technique. The approach involves parsing all substrings of the sequence, starting from the shortest one that commences with the first symbol of the sequence and increasing the length by one symbol until the entire RNA sequence is included. The parsing operation stops when the substring length falls below a specified threshold, corresponding to the minimum length of the pseudoknot.

To handle the grammatical ambiguity, YAEP, a highly efficient parser that utilizes Earley's algorithm, is selected for the CFG parser. A context-free grammar is used to enable the expansion with attributes, which can store probabilities and facilitate the pruning of parse trees while constructing them, as a forthcoming extension. An alternate method for the initial task using a brute force approach was also proposed to improve the system's performance.

The proposed system allows the user to adjust the grammar to choose whether or not to include U–G base pairs in the loops of the pseudoknot. Furthermore, the proposed approach can be extended to incorporate longer substrings between the crossing base pairs with appropriate modifications to the grammar. Section 4.1.1 explains how this substring is incorporated into the original RNA sequence, and how extra base pairs are added to the pseudoknot. The optimal tree selection process for the generated parse trees is discussed in Section 4.2.

4.1.1. Decorate Core Stems

After creating the parse trees as described in the previous subsection, the pseudoknot is decorated with additional base pairs by exploring all generated trees. To improve the efficiency of the CFG parser, only the essential stems of the pseudoknot are recognized by the parser. Although this approach reduces the number of syntactic rules in the CFG and enhances its performance, it mandates traversing all parse trees to detect the base pairs flanking the essential stems. The parser sequentially examines each base within the pseudoknot loops to determine whether it can form a base pair with another base situated in the correct position. The algorithm's decoration is shown in Table 2. After identifying the core stems at positions 2–10, 9–16, and 5–14 for U–A, G–C, and G–C, respectively, bases in the loop at positions 11–13 and the loop at positions 6–8 are examined for potential base pairing with bases outside the loops (positions 0–1 and 17–19, respectively). Then, the bases in the internal loop at positions 3–4 and the internal loop at position 15 are examined for potential base pairing.

Sequentially, base pairs at positions 1–11 (step 1), 8–17 and 7–18 (step 2), and 4–15 (step 3) are identified. The complete process is described in detail in Table 2.

4.2. Choosing the Optimal Pseudoknot

Various techniques have been proposed in the scientific literature to predict RNA base pairing, including (i) the minimum free energy (MFE) method [43], based on the second law of thermodynamics, which identifies the RNA sequence with the lowest free energy, though it may not be often found in nature; (ii) the maximum pairing principle [44], which relies on the count of base pairs around the essential stems of the pseudoknot, and the dot–bracket representation with the highest count of base pairs surrounding the pseudoknot usually corresponds to the minimum free energy; (iii) the partition function method [45], which assumes that the true base pairs should have a high probability of forming in the estimated minimum free energy distribution, thus improving accuracy by considering the free energy of their nearest neighbors at a given temperature; (iv) comparative sequence analysis [46], which involves analyzing the substitution pattern in a pairwise alignment

of two homologous sequences; (v) physical experiments [47], which entail conducting wet experiments.

In this research, we propose a hybrid optimal tree selection model that combines principles from the two most common methods, maximum pairing and MFE methods, to accurately and efficiently predict the RNA secondary structure, including the complex motif of the L-type pseudoknot. The MFE approach is computationally efficient. All trees are initially ranked by the count of base pairs surrounding the recognized pseudoknot, and MFE is applied only to the trees with the highest base pair count. This heuristic surpasses the original MFE approach.

To identify the optimal secondary structure, our approach utilizes the minimum free energy as a criterion. We integrated a module from HotKnots [48] that calculates the energy of each structure and provides it to our framework for the ultimate choice. It uses the energy calculation algorithm proposed by Mathews [49] and modified for pseudoknots by Dirks [50].

5. Implementation Details

The current study proposes a new method for identifying L-type pseudoknots, a rare type of pseudoknot. The proposed method involves creating a set of all potential pseudoknots according to a proposed grammar, and selecting the best option based on the principles of maximum base pairs around the pseudoknot and minimum free energy. The prediction of pseudoknots for an arbitrary RNA sequence is an NP-complete problem, and current algorithms, such as free energy minimization algorithms, become less precise as the sequence length increases. In addition, heuristic approaches lack generalization capabilities when tested under different datasets. To address these challenges, we propose a hybrid strategy that selects the RNA sub-sequence with the most probable pseudoknot expression. This involves creating a set of all L-type pseudoknot structures and then solving an optimization problem to select the pseudoknot expression with the maximum number of base pairs around the pseudoknot and the minimum free energy. The code routines were implemented in C, in Python, for further execution time optimization during the parsing task. The input sequence is sliced into multiple sub-sequences to parallelize the workload, and a parallel CFG parser is employed to evaluate all sub-sequences in parallel by spawning a pool of tasks. Each parser instance produces a pseudoknot structure that describes potential pseudoknots within the CFG domain. All pseudoknot structures are stored in a data structure, and the most likely solution is selected based on the least free energy. To overcome the computational and memory-intensive task of computing the free energy for each potential RNA folding, a maximum stem count lookup is performed, with a time and space complexity of O(n) proportional to the input sequence's length. The source code for the implementation is publicly available under the *L-type-knotify* GitHub repo [51].

6. Predicting a Well-Known L-Type Pseudoknot

In order to validate the efficiency of such platforms, it is common practice to construct a suitable dataset and compare the predicted secondary structures using specific metrics against the ground truth structure observed by biologists as well as other well-known platform predictions. Unfortunately, in the case of L-type pseudoknots, the dataset available is extremely limited. To date, only one L-type pseudoknot has been observed, which was initially presented in [52]. This RNA sequence was used as the input to the proposed platform as well as three other state-of-the-art methods: IPknot [26], ProbKnot [25], and Knotty [23]. The resulting dot–bracket notations are presented in Table 3, demonstrating that our platform, referred to as "Knotify", accurately predicted the core stems of the L-type pseudoknot. In contrast, Knotty and IPknot predicted H-type pseudoknots, which possess hairpins in the pseudoknot loops, while ProbKnot predicted a structure containing only hairpins, thus failing to predict the core stems of the L-type pseudoknot.

It can be readily comprehended from Figure 3, which presents the ground truth in Subfigure a, our platform's prediction in Subfigure b, and Knotty, IPknot, and ProbKnot

predictions in Figure 3c,d,e, respectively, that the proposed system's prediction of an L-type pseudoknot closely approximates the actual pseudoknot structure. The process of visualizing the RNA pseudoknot structures shown in Figure 3 was executed through the utilization of the pseudoviewer tool [53].

In addition, we compared our platform's performance to other state-of-the-art methods, i.e., IPknot, ProbKnot, and Knotty, using the following four fundamental metrics. Table 4 displays the results for each method.

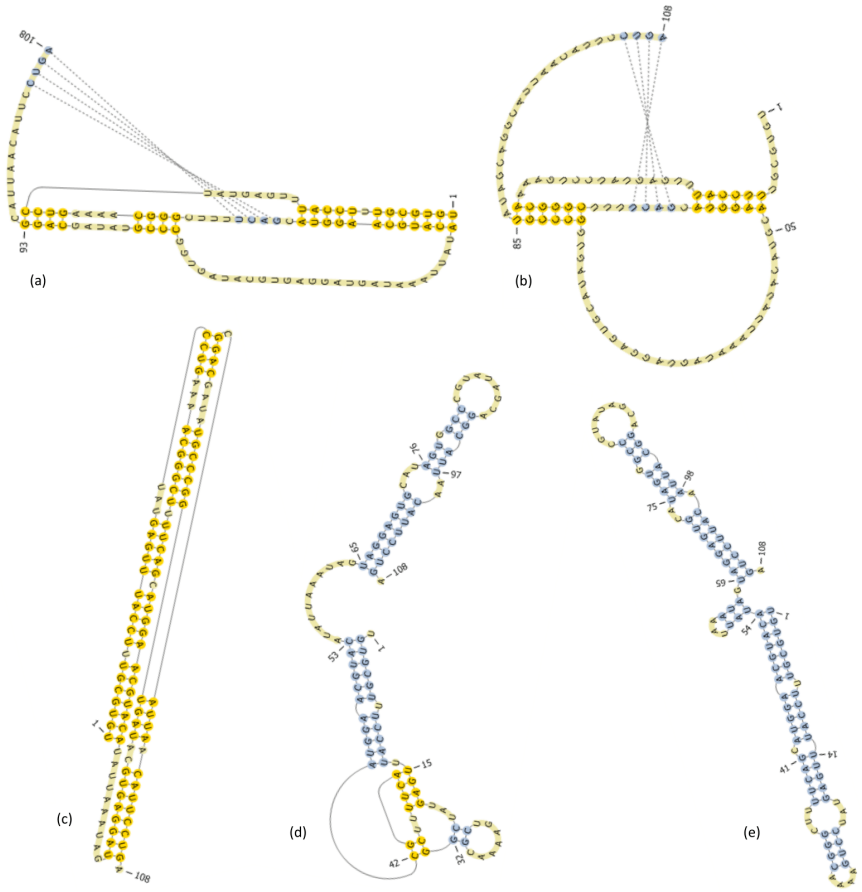

Figure 3. Ground truth (**a**), our platform's prediction (**b**), Knotty's prediction (**c**), IPknot's prediction (**d**), and ProbKnot's prediction (**e**) of L-type pseudoknot presented in [52].

- Positive predictive value (PPV): PPV is a metric that measures the proportion of true positives among the positive predictions made by a classification model. It is calculated as the ratio of true positives (TPs) to the sum of true positives and false positives (FPs):
  $PPV = TP/(TP + FP)$
- Recall: Recall is a metric that measures the proportion of true positives that were correctly identified by a classification model. It is calculated as the ratio of TPs to the sum of true positives and false negatives (FNs):
  $Recall = TP/(TP + FN)$
- F1-score: The F1-score is a metric that balances precision (PPV) and recall. It is calculated as the harmonic mean of precision and recall:
  $F1\text{-}score = 2 * (PPV * Recall)/(PPV + Recall)$
- Matthews correlation coefficient (MCC): MCC is a metric that takes into account all four outcomes of a binary classification model (true positives, false positives, true negatives, and false negatives). It is calculated as follows:

$$MCC = (TP * TN - FP * FN)/\sqrt{((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))}$$

In order to assess the overall performance of our methodology, we focused on precision; recall; Matthews correlation coefficient (MCC); and F1-score, which is the harmonic mean of precision and recall. Our methodology surpassed the other two methods in terms of precision, with a score of 0.844, while Knotty had a score of 0.500, IPknot a score of 0.333, and ProbKnot a score of 0.308. In terms of the F1-score, our platform achieved a score of 0.675, while Knotty achieved a score of 0.661, which is very close to our platform, and both other platforms scored 0.449. The proposed methodology outperformed all methods in terms of the MCC metric, with a score of 0.521. Finally, Knotty achieved a very high recall score (0.976) by having only one false negative, which may have been due to the fact that large RNA sequences often contain multiple structures (such as bulges) that do not necessarily relate to the pseudoknot directly. This may have increased the overall true positive score and reduced the false negative score. In our future work, we plan to further improve our methodology to handle even more complex patterns, such as L-type pseudoknots containing bulges or hairpins.

Table 4. Four fundamental metrics per platform.

Platform	tp	tn	fp	fn	Precision	Recall	F1-Score	MCC
Knotify	27	55	5	21	0.844	0.563	0.675	0.521
IPknot	22	32	44	10	0.333	0.688	0.449	0.102
ProbKnot	24	25	54	5	0.308	0.828	0.449	0.143
Knotty	40	27	40	1	0.500	0.976	0.661	0.419

Unfortunately, to the best of authors' knowledge, no other L-type pseudoknot structure has been reported to date, thus preventing the further evaluation of the proposed platform. The above-mentioned results are also shown in Figure 4.

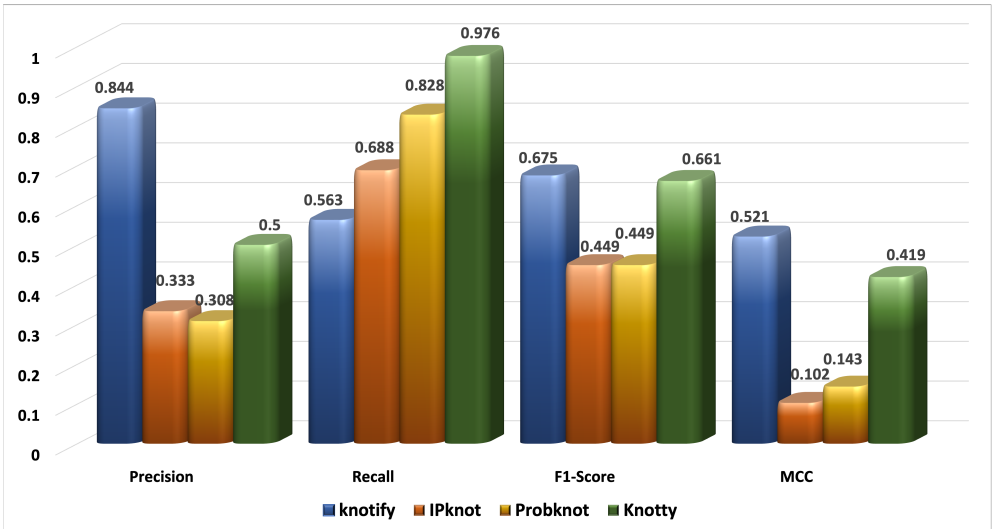

Figure 4. Precision, recall, F1-score, and MCC per platform.

7. Discussion

The structure of RNA molecules plays a vital role in understanding their functions and operations in organisms. For many years, scientists have used experimental methods to analyze RNA structures, but these methods are time-consuming, and may not be accurate. With the recent advancements in computational methods, predicting RNA structures has become faster and more precise. One of the challenging tasks in RNA structure prediction is the identification of pseudoknots, which are complex RNA structures that play critical roles in many biological processes. In this paper, we present a grammar-based framework for predicting a rare but complex type of pseudoknot, the L-type pseudoknot, extending our previous work on H-type pseudoknots. We evaluate the effectiveness of our methodology

by comparing it with three other well-known methods, using four performance metrics: precision, recall, Matthews correlation coefficient (MCC), and F1-score. The proposed methodology uses a grammar-based approach to predict all possible L-type pseudoknots of a given RNA sequence. We leverage core biological principles, such as maximum base pairs and minimum free energy, to ensure the accuracy of our predictions. We compare the performance of our methodology with three other methods using four performance metrics: precision, recall, MCC, and F1-score. Our methodology outperformed the other three methods in terms of precision, with a score of 0.844, while the other methods scored 0.500, 0.333, and 0.308. Regarding the F1-score, our methodology scored 0.671, while the other methods scored 0.661, 0.449, and 0.449. The proposed methodology surpassed all methods in terms of the MCC metric, achieving a score of 0.521. The Knotty platform achieved a very high recall score.

Our methodology provides an accurate and fast way to predict L-type pseudoknots in RNA sequences. The use of a grammar-based approach, combined with biological principles, ensures the accuracy of our predictions. The proposed method is added to our RNA toolset, which aims to enhance the capabilities of biologists in the prediction of RNA motifs, including pseudoknots. The use of computational methods in RNA structure prediction is an active area of research, and our methodology contributes to the development of accurate and fast methods for predicting RNA structures. Our methodology can be used in many biological applications, such as drug design, gene therapy, and understanding RNA function. Moreover, the proposed method can be used in combination with other methods to increase the accuracy of RNA structure prediction.

8. Conclusions and Future Work

The observation and analysis of RNA molecules have proved crucial for the understanding of various processes in nature. The prediction of the secondary structure in RNA, and especially for pseudoknots, is of utmost importance. In response to this need, a new methodology has been introduced to accurately and effectively detect the rare variation of L-type pseudoknot. The innovative method is based on Earley's parser, which generates a set of all possible parse trees for an RNA sequence, where each tree represents a potential pseudoknot structure. A well-known RNA sequence containing an L-type pseudoknot was applied as the input for the proposed platform, as well as for three other state-of-the-art methods, namely, IPknot, ProbKnot, and Knotty. It was observed that our platform succeeded in accurately predicting the core stems of the L-type pseudoknot. Conversely, Knotty and IPknot predicted H-type pseudoknots, which possess hairpins in the pseudoknot loops, while ProbKnot predicted a structure containing only hairpins. To identify the optimal structure, the method combines pairing maximization and free energy minimization for each structure, using a hybrid model. We intend to extend the methodology by designing cutting edge algorithms and incorporating machine learning techniques for more intricate patterns in RNA structures. In future work, different variations of the methodology with various execution parameters will be available on a web platform with a modern graphical user interface.

Author Contributions: Conceptualization, C.P.; methodology, C.P. and E.M.; software, C.K. and A.K.; validation, A.K.; formal analysis, E.M.; investigation, E.M. and C.P.; resources, E.M.; data curation, E.M. and C.K.; writing—original draft preparation, C.P. and E.M.; writing—review and editing, C.P. and E.M.; visualization, C.P. and E.M.; supervision, C.P. and P.T.; project administration, C.P. and P.T.; funding acquisition, P.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFG	Context-free grammar
CNN	Convolutional neural network
CPU	Central processing unit
CYK	Cocke–Younger–Kasami
DAG	Direct acyclic graph
FCN	Fully convolutional network
IBPMP	Improved base pair maximization principle
LSTM	Long short-term memory
MCC	Matthews correlation coefficient
MFE	Minimum free energy
NMR	Nuclear magnetic resonance
RNA	Ribonucleic acid
SCFG	Stochastic context-free grammar
YAEP	Yet Another Earley Parser

References

1. Andrikos, C.; Makris, E.; Kolaitis, A.; Rassias, G.; Pavlatos, C.; Tsanakas, P. Knotify: An Efficient Parallel Platform for RNA Pseudoknot Prediction Using Syntactic Pattern Recognition. *Methods Protoc.* **2022**, *5*, 14. [CrossRef]
2. Makris, E.; Kolaitis, A.; Andrikos, C.; Moulos, V.; Tsanakas, P.; Pavlatos, C. Knotify+: Toward the Prediction of RNA H-Type Pseudoknots, Including Bulges and Internal Loops. *Biomolecules* **2023**, *13*, 308. [CrossRef]
3. Watson, J.; Crick, F. Molecular Structure Of Nucleic Acids. *Am. J. Psychiatry* **2003**, *160*, 623–624. [CrossRef] [PubMed]
4. Rietveld, K.; Van Poelgeest, R.; Pleij, C.W.; Van Boom, J.; Bosch, L. The tRNA-Uke structure at the 3′ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Res.* **1982**, *10*, 1929–1946. [CrossRef]
5. Kucharík, M.; Hofacker, I.L.; Stadler, P.F.; Qin, J. Pseudoknots in RNA folding landscapes. *Bioinformatics* **2016**, *32*, 187–194. [CrossRef] [PubMed]
6. Staple, D.W.; Butcher, S.E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* **2005**, *3*, e213. [CrossRef]
7. Hopcroft, J.E.; Ullman, J.D. *Formal Languages and their Relation to Automata*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1969.
8. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **1956**, *2*, 113 –124. [CrossRef]
9. Sipser, M. *Introduction to the Theory of Computation*; Thomson Course Technology: Boston, MA, USA, 2006; Volume 2.
10. Younger, D.H. Recognition and parsing of context-free languages in n^3. *Inf. Control.* **1967**, *10*, 189–208. [CrossRef]
11. Earley, J. An efficient context-free parsing algorithm. *Commun. ACM* **1970**, *13*, 94–102. [CrossRef]
12. Graham, S.L.; Harrison, M.A.; Ruzzo, W.L. An improved context-free recognizer. *ACM Trans. Program. Lang. Syst.* **1980**, *2*, 415–462. [CrossRef]
13. Ruzzo, W.L. General Context-Free Language Recognition. Ph.D. Thesis, University of California, Berkeley, CA, USA, 1978.
14. Geng, T.; Xu, F.; Mei, H.; Meng, W.; Chen, Z.; Lai, C. A practical GLR parser generator for software reverse engineering. *J. Netw.* **2014**, *9*, 769–776. [CrossRef]
15. Pavlatos, C.; Dimopoulos, A.C.; Koulouris, A.; Andronikos, T.; Panagopoulos, I.; Papakonstantinou, G. Efficient reconfigurable embedded parsers. *Comput. Lang. Syst. Struct.* **2009**, *35*, 196–215. [CrossRef]
16. Chiang, Y.; Fu, K. Parallel parsing algorithms and VLSI implementations for syntactic pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 302–314. [CrossRef]
17. Available online: https://github.com/vnmakarov/yaep (accessed on 25 March 2020).
18. Cao, S.; Chen, S. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* **2009**, *4*, 696–706 [CrossRef]
19. Akutsu, T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discret. Appl. Math.* **2000**, *104*, 45–62. [CrossRef]
20. Meyer, I.M.; Miklos, I. SimulFold: Simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.* **2007**, *3*, e149. [CrossRef]
21. Van Batenburg, F.; Gultyaev, A.P.; Pleij, C.W. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **1995**, *174*, 269–280. [CrossRef]
22. Isambert, H.; Siggia, E.D. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6515–6520. [CrossRef] [PubMed]
23. Jabbari, H.; Wark, I.; Montemagno, C.; Will, S. Knotty: Efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics* **2018**, *34*, 3849–3856. [CrossRef]

24. Chen, H.L.; Condon, A.; Jabbari, H. An O(n^5) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *J. Comput. Biol.* **2009**, *16*, 803–815. [CrossRef]
25. Bellaousov, S.; Mathews, D.H. ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA* **2010**, *16*, 1870–1880. [CrossRef]
26. Sato, K.; Kato, Y.; Hamada, M.; Akutsu, T.; Asai, K. IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **2011**, *27*, 85–93. [CrossRef] [PubMed]
27. Sato, K.; Kato, Y. Prediction of RNA secondary structure including pseudoknots for long sequences. *Briefings Bioinform.* **2021**, *23*, bbab395. [CrossRef]
28. Singh, J.; Hanson, J.; Paliwal, K.; Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **2019**, *10*, 5407. [CrossRef] [PubMed]
29. Wang, L.; Liu, Y.; Zhong, X.; Liu, H.; Lu, C.; Li, C.; Zhang, H. DMfold: A novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair Maximization Principle. *Front. Genet.* **2019**, *10*, 143. [CrossRef]
30. Kangkun, M.; Jun, W.; Yi, X. Prediction of RNA secondary structure with pseudoknots using coupled deep neural networks. *Biophys. Rep.* **2020**, *6*, 146–154.
31. Wang, Y.; Liu, Y.; Wang, S.; Liu, Z.; Gao, Y.; Zhang, H.; Dong, L. ATTfold: RNA secondary structure prediction with pseudoknots based on attention mechanism. *Front. Genet.* **2020**, *11*, 1564. [CrossRef]
32. Fu, L.; Cao, Y.; Wu, J.; Peng, Q.; Nie, Q.; Xie, X. UFold: Fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Res.* **2021**, *50*, e14. [CrossRef]
33. Knudsen, B.; Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **1999**, *15*, 446–454. [CrossRef] [PubMed]
34. Knudsen, B.; Hein, J. Pfold: RNA Secondary Structure Prediction Using Stochastic Context-Free Grammars. *Nucleic Acids Res.* **2003**, *31*, 3423–3428. [CrossRef] [PubMed]
35. Sukosd, Z.; Knudsen, B.; Vaerum, M.; Kjems, J.; Andersen, E.S. Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. *BMC Bioinform.* **2011**, *12*, 103. [CrossRef]
36. Pedersen, J.S.; Meyer, I.M.; Forsberg, R.; Simmonds, P.; Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **2004**, *32*, 4925–4936. [CrossRef] [PubMed]
37. Do, C.B.; Woods, D.A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98. [CrossRef] [PubMed]
38. Pedersen, J.S.; Bejerano, G.; Siepel, A.; Rosenbloom, K.; Lindblad-Toh, K.; Lander, E.S.; Kent, J.; Miller, W.; Haussler, D. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2006**, *2*, e33. [CrossRef]
39. Nawrocki, E.P.; Kolbe, D.L.; Eddy, S.R. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* **2009**, *25*, 1335–1337. [CrossRef]
40. Anderson, J.W.; Haas, P.A.; Mathieson, L.A.; Volynkin, V.; Lyngsø, R.; Tataru, P.; Hein, J. Oxfold: Kinetic folding of RNA using stochastic context-free grammars and evolutionary information. *Bioinformatics* **2013**, *29*, 704–710. [CrossRef]
41. Bradley, R.K.; Pachter, L.; Holmes, I. Specific alignment of structured RNA: Stochastic grammars and sequence annealing. *Bioinformatics* **2008**, *24*, 2677–2683. [CrossRef]
42. Makris, E.; Kolaitis, A.; Andrikos, C.; Moulos, V.; Tsanakas, P.; Pavlatos, C. An intelligent grammar-based platform for RNA H-type pseudoknot prediction. In *Artificial Intelligence Applications and Innovations, Proceedings of the AIAI 2022 IFIP WG 12.5 International Workshops: IFIP Advances in Information and Communication Technology, Crete, Greece, 17–20 June 2022*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 652.
43. Trotta, E. On the normalization of the minimum free energy of RNAs by sequence length. *PLoS ONE* **2014**, *9*, e113380. [CrossRef]
44. Nussinov, R.; Jacobson, A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 6309–6313. [CrossRef] [PubMed]
45. Mathews, D.H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **2004**, *10*, 1178–1190. [CrossRef]
46. Rivas, E.; Eddy, S.R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform.* **2001**, *2*, 8. [CrossRef] [PubMed]
47. Chu, Y.; Corey, D.R. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* **2012**, *22*, 271–274. [CrossRef]
48. Ren, J.; Rastegari, B.; Condon, A.; Hoos, H.H. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **2005**, *11*, 1494–1504. [CrossRef] [PubMed]
49. Mathews, D.; Sabina, J.; Zuker, M.; Turner, D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure1. *J. Mol. Biol.* **1999**, *288*, 911–940. [CrossRef]
50. Dirks, R.; Pierce, N. Introduction A Partition Function Algorithm for Nucleic Acid Secondary Structure Including Pseudoknots. *J. Comput. Chem.* **2003**, *24*, 1664–1677. [CrossRef]
51. Available online: https://github.com/chriskor1 (accessed on 9 March 2023).

52. Bon, M.; Vernizzi, G.; Orland, H.; Zee, A. Topological classification of RNA structures. *J. Mol. Biol.* **2008**, *379*, 900–911. [CrossRef]
53. Byun, Y.; Han, K. PseudoViewer3: Generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics* **2009**, *25*, 1435–1437. [CrossRef] [PubMed]