B.Sc. in Computer Science and Engineering Thesis

RNA Motif Identification : A Graph Traversal Approach

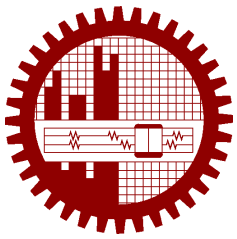
Submitted by

Badhan Das
201205023

Zarin Tasnim Promi
201205047

Supervised by

Dr. Md. Abul Kashem Mia



Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Dhaka, Bangladesh

September 2017

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis, titled, "RNA Motif Identification : A Graph Traversal Approach", is the outcome of the investigation and research carried out by us under the supervision of Dr. Md. Abul Kashem Mia.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Badhan Das
201205023

Zarin Tasnim Promi
201205047

CERTIFICATION

This thesis titled, **"RNA Motif Identification : A Graph Traversal Approach"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in September 2017.

Group Members:

Badhan Das

Zarin Tasnim Promi

Supervisor:

Dr. Md. Abul Kashem Mia
Professor
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

ACKNOWLEDGEMENT

First and foremost, we are grateful to the God for the good health and wellbeing that were necessary to complete this book.

Secondly, we have to thank my research supervisors, Dr. Md. Abul Kashem Mia. Without his assistance and dedicated involvement in every step throughout the process, this paper would have never been accomplished. We would like to thank him very much for his support and understanding over this past one year.

We wish to express our sincere thanks to Dr. M. Sohel Rahman, Head of the Dept., CSE, for providing us with all the necessary facilities for the research.

We take this opportunity to express gratitude to all of the Department faculty members for their help and support. We also thank our parents for the unceasing encouragement, support and attention.

Finally,

Dhaka
September 2017

Badhan Das

Zarin Tasnim Promi

Contents

List of Figures

ABSTRACT

RNA structural motifs are recurrent three-dimensional (3D) components found in the RNA architecture which exhibits highly conserved 3D geometries and base-interaction patterns. RNA motifs are key regulators of gene expression. These motif segement's arrangement, abundance and interaction largely determine the folding behaviors and functionalities of different RNA structures. Analysis of the RNA 3D structures and their molecular functions depends on efficient and accurate identification of these motifs. However, constructing efficient RNA structural motif identification tools are not easy due to high complexity of these motifs. In this thesis work we present a motif identification algorithm based on a graph traversal technique. Our approach enables automatic identification of both partially and fully matched motif instances. This approach represents the RNA sequence as a graph where each nitrogen base as a vertex and each base pair and base stacking bond as an egde. From the graph representation our algorithm finds the motif using a depth first search (DFS) traversal based on the degree of the edges. The best known efficient algorithms uses alignment approach constructing matching graph. We have worked with a different approach to reduce both space and time complexity.

Chapter 1

Introduction

1.1 Background of the Study

Our rapidly expanding knowledge of the characteristics of RNAs illustrates that, like proteins, RNA assumes complex three-dimensional (3D) structures to perform specific roles based on that structure. Unlike proteins however, RNA forms more locally stable structures, called structural motifs, that are combinatorially linked and constrained by tertiary interactions to create a 3D structure. So RNA motifs identification has become a popular research topic in recent times which can patronize further research in bioinformatics. Many processes have been adapted to find a searching tool for motifs. Some adapts different heuristics. Some tools have graph matching approach.So the methodology of different processes varies rapidly . But only a few have been done using graph traversal approach. In our research work we tried to explore ways to use a graph traversal approach.

1.2 Motivation

RNAs play a large variety of roles inside a cell, and recent discoveries point to many of their novel cellular functions. The varieties of functionality of non-coding RNA is determined by their complex structural motifs . These motif segments arrangement, abundance and interaction largely determine the folding behaviors and functionality of the different RNA structures.

The identification and analysis of these motifs have largely enriched our experiences in RNA studies. But it still remains a challenging task to identify motif properly. Many search tolls needs manual inspection which is unfeasible for large scale data. Some aims at detecting a similar geometry. Some tools try to model variations using heuristics but none of them solve the

problem optimally . Hence we tried to find an approach to reduce polynomiality of the problem.

1.3 Objective of the Study

In our research work the main aim is to find a graph representation of the RNA motif sequence. To reduce the search space we to put different constraint and find out how effective they are. The main objectives of our study are as below:

- Understanding different terms of and related to RNA motif.
- Studying the challenges and potential obstacles that may happen during the recognition of a RNA motif.
- Studying the existing approach for recognition of RNA motif.
- Propose a novel approach for the recognition RNA motif sequence.
- Find an effective representation of input data that helps faster operations.

1.4 Contribution

RNA motifs play a significant role in determining RNA functionality and characteristics of the structures. Our aim was to find an way to identify motif using a graph like representation. In our works we used DFS graph traversal process based on the degrees of the edges where each edge represents a base paring or a base stacking bond. We mainly focused on putting a constraint on the search space by using degree of each node.

1.5 Thesis Organization

Chapter 1 discusses preliminary ideas about the topic revealing the motivation behind our work. As to understand our thesis work one needs certain background knowledge about different biological terms , we provided them in chapter 2. All the necessary definition for better understanding were given. In chapter 3 some related works in the field were discussed which includes summery of the methodology. Our proposed methodology to identify RNA Motif was discussed elaborately in chapter 4 .Chapter 5 finally mentions our findings, summarizes our

work and discusses the future works regarding the RNA Motif recognition approach. RNA Motif Identification is a popular research topic in bioinformatics, aims to exploit biological data to understand biological processes through computational approach.

Chapter 2

Preliminaries

2.1 RNA

Ribonucleic acid (RNA) is a polymeric molecule essential in various biological roles in coding, decoding, regulation, and expression of genes. RNA and DNA are nucleic acids, and, along with it lipids, proteins and carbohydrates, constitute the four major macromolecules essential for all known forms of life.

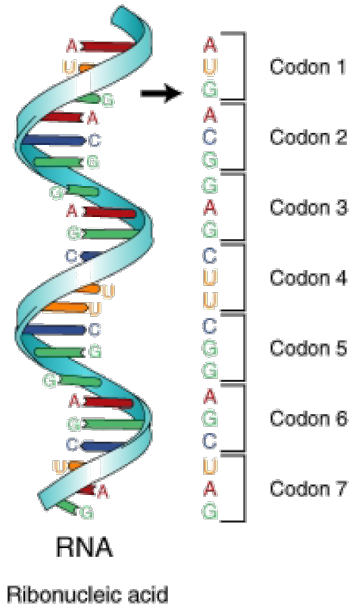


Figure 2.1: RNA and codon

Like DNA, RNA is assembled as a chain of nucleotides, but unlike DNA it is more often found in nature as a single-strand folded onto itself, rather than a paired double-strand. Cellular organisms use messenger RNA (mRNA) to convey genetic information (using letters G,U, A, and C

to denote the nitrogenous bases Guanine, Uracil, Adenine, and Cytosine) that directs synthesis of specific proteins. Many viruses encode their genetic information using an RNA genome.

Some RNA molecules play an active role within cells by catalyzing biological reactions, controlling gene expression, or sensing and communicating responses to cellular signals. One of these active processes is protein synthesis, a universal function where RNA molecules direct the assembly of proteins on ribosomes. This process uses transfer RNA (tRNA) molecules to deliver amino acids to the ribosome, where ribosomal RNA (rRNA) then links amino acids together to form proteins.

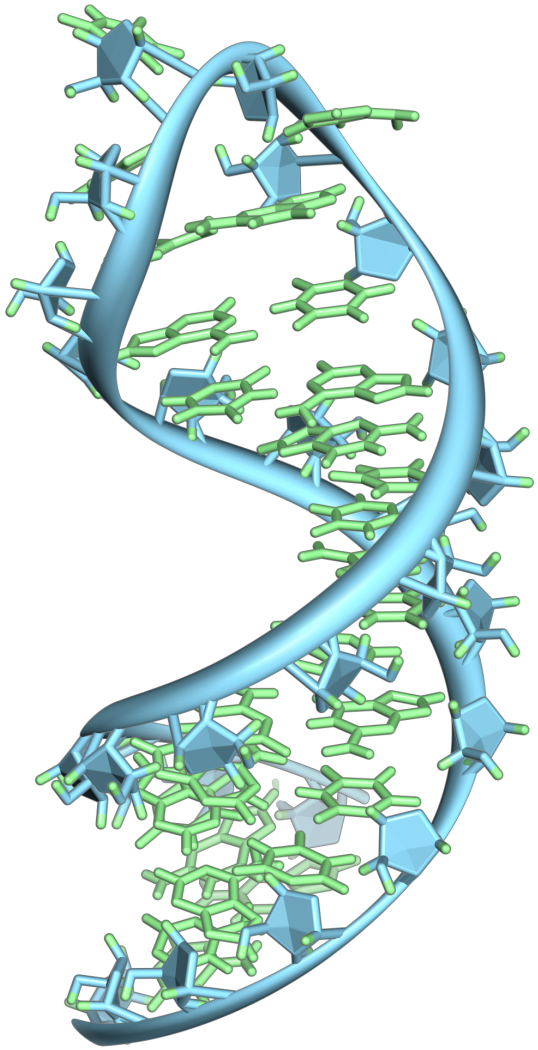


Figure 2.2: RNA

2.2 Non-coding RNA

A non-coding RNA (ncRNA) is an RNA molecule that is not translated into a protein. Less-frequently used synonyms are non-protein-coding RNA (npcRNA), non-messenger RNA (nmRNA), or functional RNA (fRNA). The DNA sequence from which a functional non-coding RNA is transcribed is often called an RNA gene.

Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), as well as RNAs such as snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs and scaRNAs and the long ncRNAs that include examples such as Xist and HOTAIR (see here for a more complete list of ncRNAs). The number of ncRNAs encoded within the human genome is unknown; however, recent transcriptomic and bioinformatic studies suggest the existence of thousands of ncRNAs. Many of the newly identified ncRNAs have not been validated for their function. It is also likely that many ncRNAs are non functional (sometimes referred to as Junk RNA), and are the product of spurious transcription.

2.3 Backbone Chain

In polymer science, the backbone chain of a polymer is the longest series of covalently bonded atoms that together create the continuous chain of the molecule. This science is subdivided into the study of organic polymers, which consist of a carbon backbone, and inorganic polymers which have backbones containing only main group elements.

In biochemistry, organic backbone chains make up the primary structure of macromolecules. The backbones of these biological macromolecules consist of central chains of covalently bonded atoms. The characteristics and order of the monomer residues in the backbone make a map for the complex structure biological polymers. The backbone is, therefore, directly related to biological molecules function.

The macromolecules within the body can be divided into four main subcategories, each of which are involved in very different and important biological processes: Proteins, Carbohydrates, Lipids, and Nucleic acids. Each of these molecules has a different backbone and consists of different monomers each with distinctive residues and functionalities. This is the driving factor of their different structures and functions in the body. Although lipids have a ”backbone,” they are not true biological polymers as their backbone is a three carbon molecule, glycerol, with longer substituent ”side chains.” For this reason, only proteins, carbohydrates, and nucleic

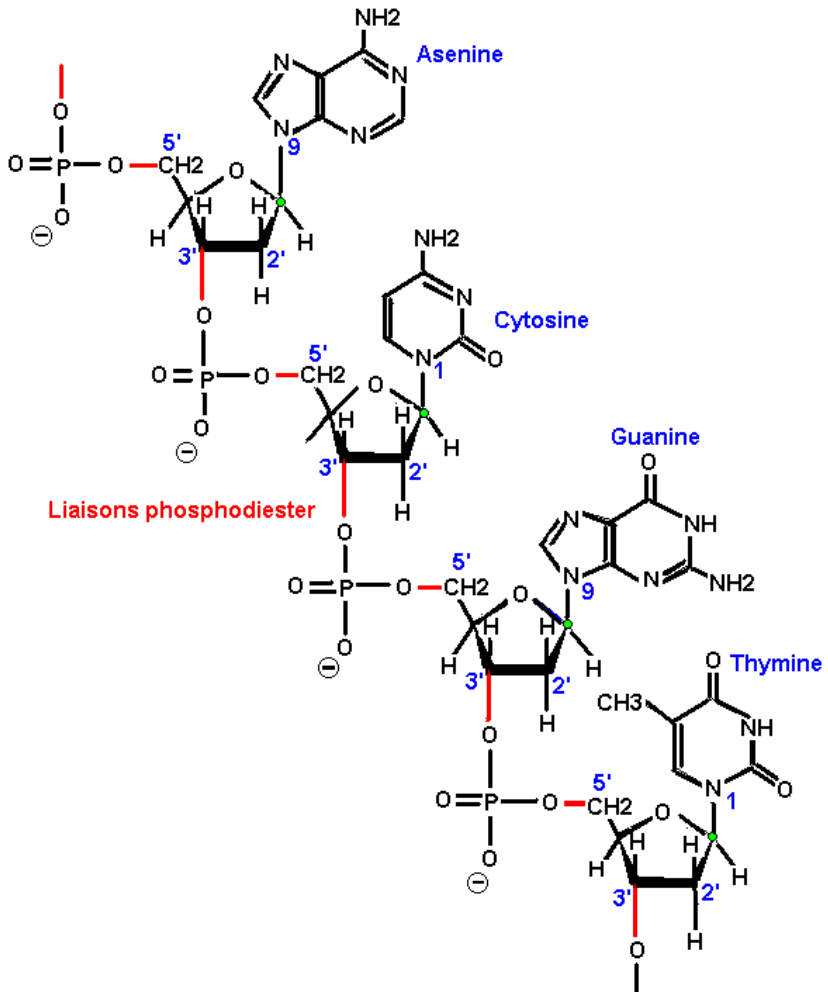


Figure 2.3: backbone chain and bases(RNA has Uracil instead of Thymine)

acids should be considered as biological macromolecules with polymeric backbones.

2.4 Base Pair

A base pair (bp) is a unit consisting of two nucleobases bound to each other by hydrogen bonds. They form the building blocks of the DNA double helix, and contribute to the folded structure of both DNA and RNA. Dictated by specific hydrogen bonding patterns, Watson-Crick base pairs (guanine-cytosine and adenine-thymine) allow the DNA helix to maintain a regular helical structure that is subtly dependent on its nucleotide sequence. [1] The complementary nature of this based-paired structure provides a backup copy of all genetic information encoded within double-stranded DNA. The regular structure and data redundancy provided by the DNA double helix make DNA well suited to the storage of genetic information, while base-pairing between DNA and incoming nucleotides provides the mechanism through which DNA polymerase replicates DNA, and RNA polymerase transcribes DNA into RNA. Many DNA-binding proteins can

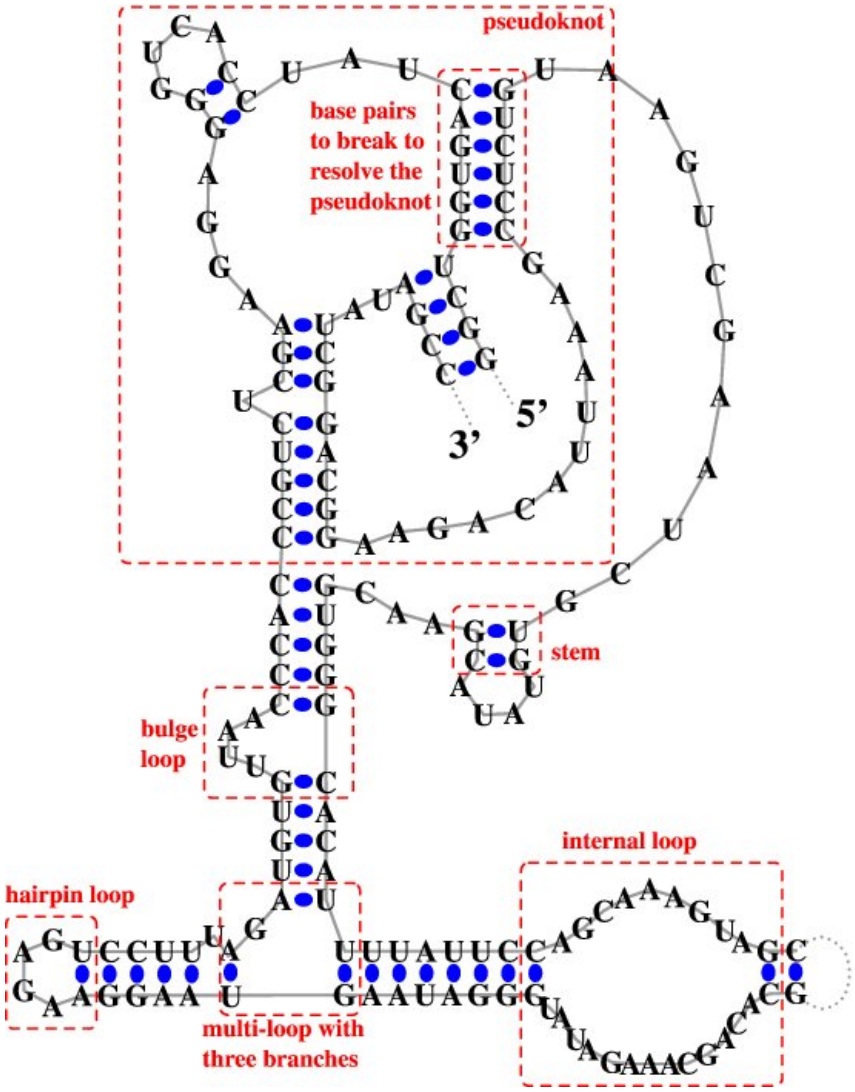


Figure 2.4: RNA strand

recognize specific base pairing patterns that identify particular regulatory regions of genes.

Intramolecular base pairs can occur within single-stranded nucleic acids. This is particularly important in RNA molecules (e.g., tRNA), where Watson-Crick base pairs (guanine-cytosine and adenine-uracil) permit the formation of short double-stranded helices, and a wide variety of non-Watson-Crick interactions (e.g., G-U or A-A) allow RNAs to fold into a vast range of specific three-dimensional structure. In addition, base-pairing between transfer RNA (tRNA) and messenger RNA (mRNA) forms the basis for the molecular recognition events that result in the nucleotide sequence of mRNA becoming translated into the amino acid sequence of proteins via the genetic code.

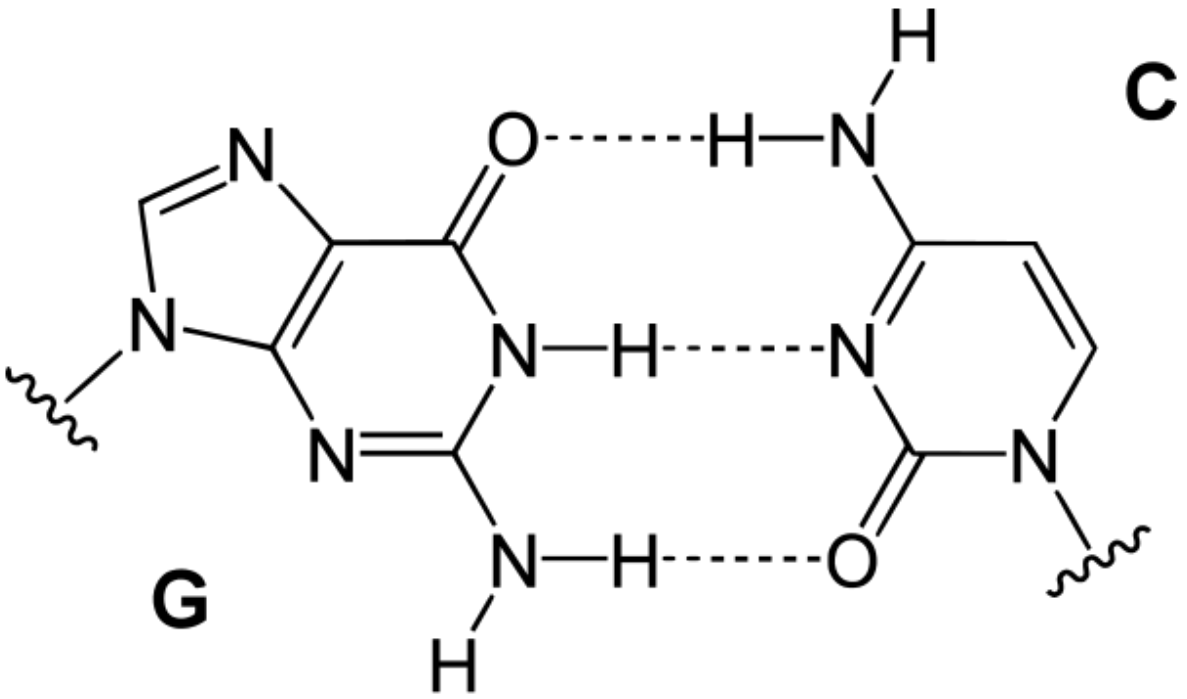


Figure 2.5: base pair

2.5 RNA Structural Motif

RNA structural motifs are recurrent three-dimensional (3D) components found in the RNA architecture. These RNA structural motifs play important structural or functional roles and usually exhibit highly conserved 3D geometries and base-interaction patterns.

RNA motifs have been defined variously as directed and ordered stacked arrays of non-Watson Crick base pairs forming distinctive foldings of the phosphodiester backbones of the interacting RNA strands (Leontis Westhof, 2003) and a discrete sequence or combination of base juxtapositions found in naturally occurring RNAs in unexpectedly high abundance (Moore, 1999). A comprehensive definition of an RNA structural motif should be based on and consist of not only base-pairing or secondary structure constraints, but a complete 3D description, including backbone conformation, all hydrogen-bonding and base-stacking inter-actions, and sequence preferences. In addition, a RNA structural motif may have co-factors such as bound waters, metals or other ions, to support its conformation ; it may have a specific functional role or a primary role in tertiary structure formation ; and it may be subject to evolutionary constraints. In order to determine these characteristics, it is necessary that the motif be frequently observed among known RNA structures. The overall 3D structure of such a recurrent motif is largely independent of the context in which it is found. Moreover, RNA structural motifs are truly structural, and there may be several sequences, seemingly unrelated, that obtain the same 3D structure. Examples include the tetraloop with sequence UMAC (M is A or C) which forms a structure almost identical to that of the GNRA tetraloop , and a tetraloop of sequence GUUA which has a fold like UNCG. We denote these as the GNRA fold and the UNCG fold. [2]

Motifs characterized by sequence or predicted secondary structure alone are not discussed here, although they may ultimately be structurally characterized as further RNA structures are determined.

2.6 Functionalities of RNA Motif

RNA is involved in a wide range of cellular activities e.g. translating genetic information, serving as a structural scaffold, catalysing biological reactions.This often require the molecule to fold into motifs in order to perform its targeted function.

The variety of functionalities of RNA is determined by their structural motifs. RNA motifs can exert their role in several different ways, particularly in dictating the interaction with RNA-

binding proteins, and acting in the regulation of a large number of cellular processes. RNA motifs determines interacion with other molecules.

Recent studies have shown that RNA structural motifs play essential roles in RNA folding by guiding the RNA folding process. These motifs often served as nucleation sites for RNA folding. RNA motifs can help stabilise a global RNA structure. RNA structural motifs works as the building blocks of the RNA architecture. These structural motifs comprise the secondary structure of RNA. These RNAs fold into characteristic secondary structures and perform specific-structure dependent biological functions.

2.7 RNA Secondary Structure

RNA molecules are polymers consisting of the four nucleotides Adenine , Guanine, Cytosine, Uracil on a sugar-phosphate backbone. The RNA linear chain of nucleotides in general folds into a three-dimensional structure, both for the purpose of molecular stability and to perform specific biological functions.

However, the RNA molecule can be viewed in terms of what is called a Secondary structure. The single stranded linear RNA molecule first folds onto itself and forms double-stranded regions by additional hydrogen bonds. These double stranded regions are complementary regions as per mainly the Watson-Crick base pairings (A-U and G-C) and possible G-U pairing. These complementary base pairings lend energic stability to the molecule. RNA Secondary structure is basically a 2-D representation of this self-folding.

Fig 2.6 shows the secondary structure of a yeast tRNA. This particular tRNA is involved in the transfer of the amino-acid phenyl-alanine to the translation site by attaching it to its 3-terminal end. The double stranded regions formed by the stacking of two or more (maximally) consecutive base-pairs are referred to as stems. The single stranded regions of the secondary structure form a variety of patterns or motifs, all of which are some kind of loops. The hairpin loop is a single-stranded region that a stem ends into. A bulge loop can be viewed as a single stranded region which interrupts a stem on one side. An interior loop on the other hand interrupts a stem on either side. A single stranded region into which more than two stems meet is calls multijunction loop. Single stranded regions in the secondary structure are in general important because they could serve as potential sites for protein/RNA binding and also be involved in additional bonding in the final tertiary structure. An RNA secondary structure is said to contain a pseudo-knot if there exist two stems related. In a pseudo-knot, the 5 segment of a stem S2 is between the 5 and 3 segments of S1, whereas its 3 segment is not. Essentially, pseudo-knot

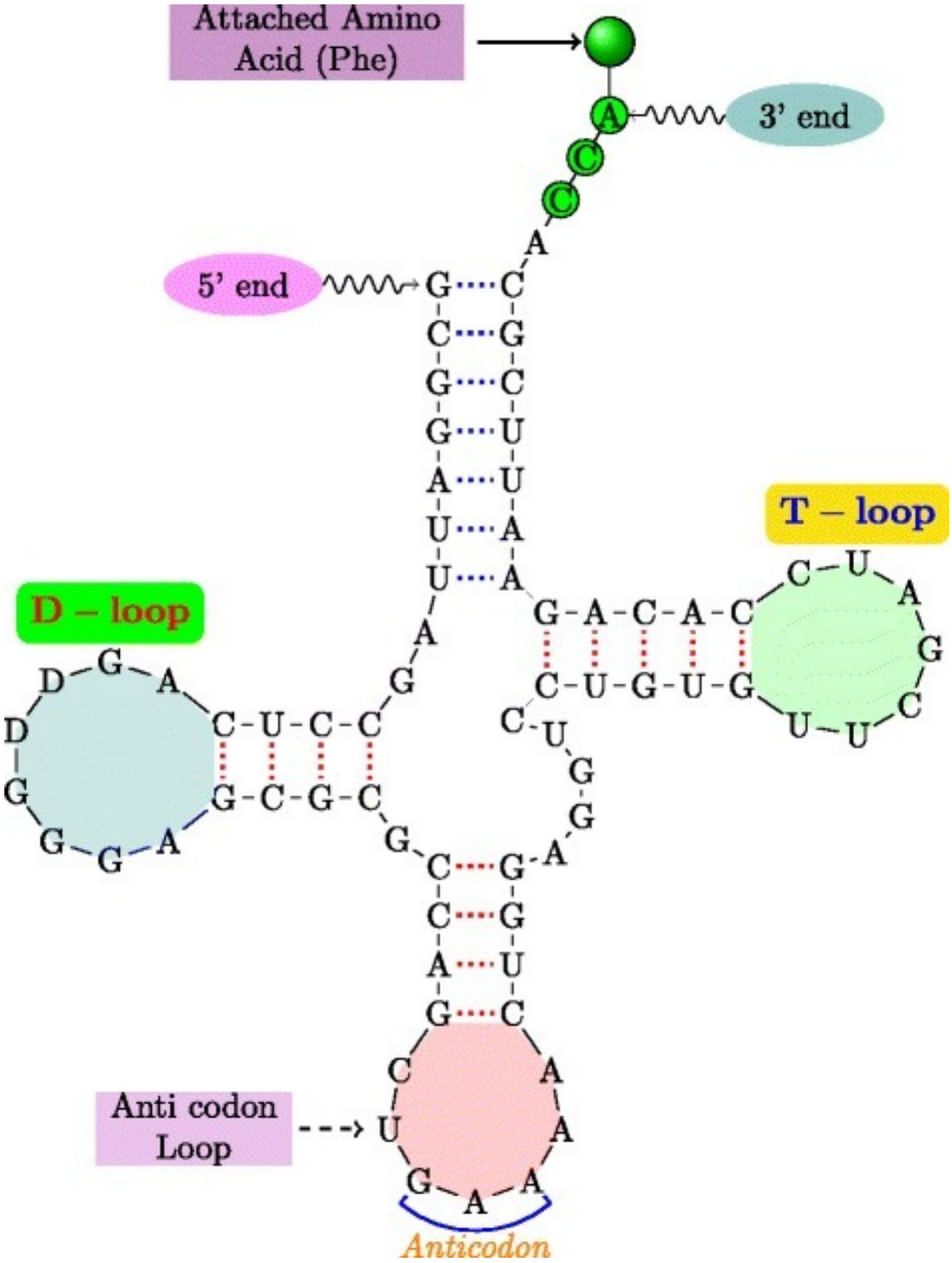


Figure 2.6: A yeast tRNA.

indicates crossing-over of the base-pair connection lines of two different stems, such that the RNA structure can no longer be presented as a planar (2-D) graph without edge crossings. An RNA structure can be differentiated based on the presence and absence of pseudo-knots. In general, handling RNA structures with pseudo-knots is computationally much harder than handling RNA 2-D structures without pseudo-knots. [3]

Chapter 3

Related Works

Noncoding RNAs (ncRNAs) are attracting recent research focus with their amazing and versatile cellular functions and many of them have significantly enriched our understanding of the molecular mechanisms. However, it remains challenging to automatically identify all known motif instances within a resolved RNA structure. There are some RNA structural motif search tools. These are discussed briefly in this chapter.

3.1 RNAMotifScan: Automatic Identification of RNA Structural Motifs Using Secondary Structural Alignment

Recent studies have shown that RNA structural motifs play essential roles in RNA folding and interaction with other molecules. Computational identification and analysis of RNA structural motifs remains a challenging task. Existing motif identification methods based on 3D structure may not properly compare motifs with high structural variations. Other structural motif identification methods consider only nested canonical base-pairing structures and cannot be used to identify complex RNA structural motifs that often consist of various non-canonical base pairs due to uncommon hydrogen bond interactions.

In this article, they have presented a novel RNA structural alignment method for RNA structural motif identification, RNAMotifScan, which takes into consideration the isosteric (both canonical and non-canonical) base pairs and multi-pairings in RNA structural motifs. The utility and accuracy of RNAMotifScan is demonstrated by searching for kink-turn, C-loop, sarcin-ricin, reverse kink-turn and E-loop motifs against a 23S rRNA, which is well characterized for the occurrences of these motifs. Finally, they have searched these motifs against the RNA structures in the entire Protein Data Bank and the abundances of them are estimated. [4]

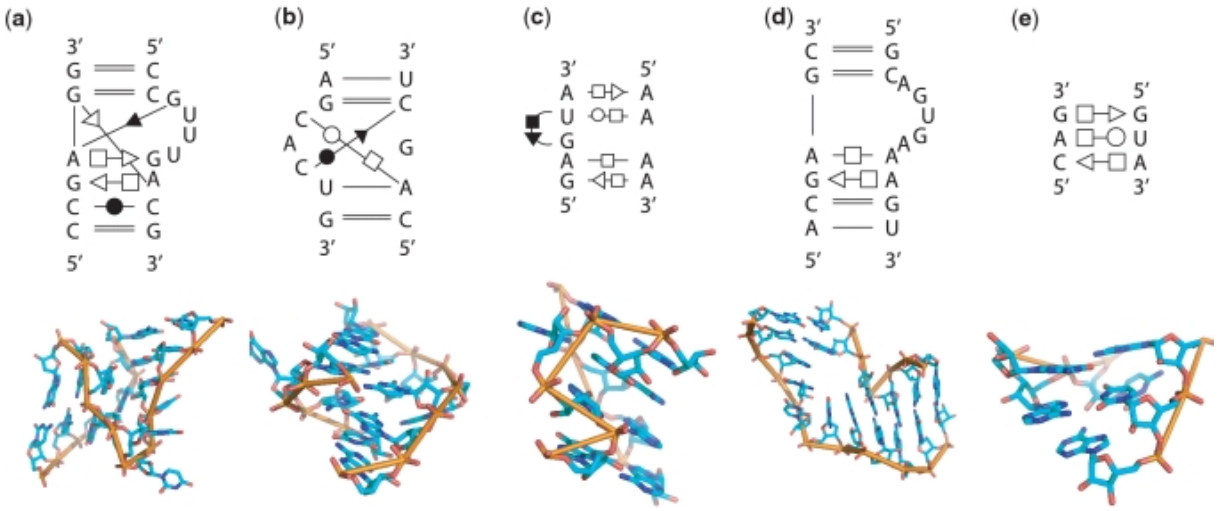


Figure 3.1: The 2D diagrams and 3D structures of newly identified motifs with sequence or base-pairing variations

3.2 RNAMotifScanX : A Graph Alignment Approach for RNA Structural Motif Identification

RNA structural motifs are recurrent three dimensional (3D) components found in the RNA architecture. These RNA structural motifs play important structural or functional roles and usually exhibit highly conserved 3D geometries and base-interaction patterns. Analysis of the RNA 3D structures and elucidation of their molecular functions heavily rely on efficient and accurate identification of these motifs. However, efficient RNA structural motif search tools are lacking due to the high complexity of these motifs.

In this work, they presented RNAMotifScanX, a motif search tool based on a base-intersection graph alignment algorithm. This novel algorithm enables automatic identification of both partially and fully matched motif instances. RNAMotifScanX considers noncanonical base-pairing interactions, base-stacking interactions, and sequence conservation of the motifs, which leads to significantly improved sensitivity and specificity as compared with other state-of-the-art search tools. RNAMotifScanX also adopts a carefully designed branch-and-bound technique, which enables ultra-fast search of large kink-turn motifs against a 23S rRNA. [5]

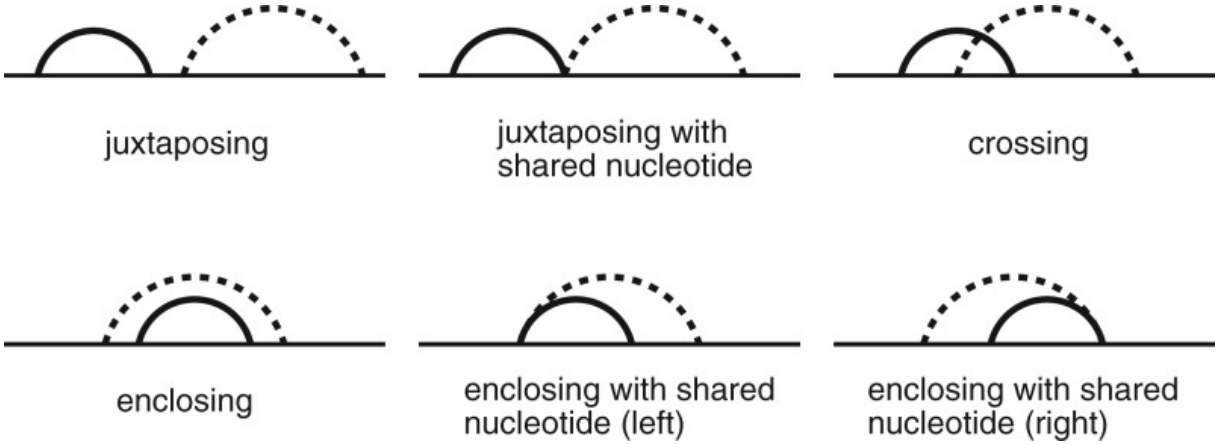


Figure 3.2: finding motif

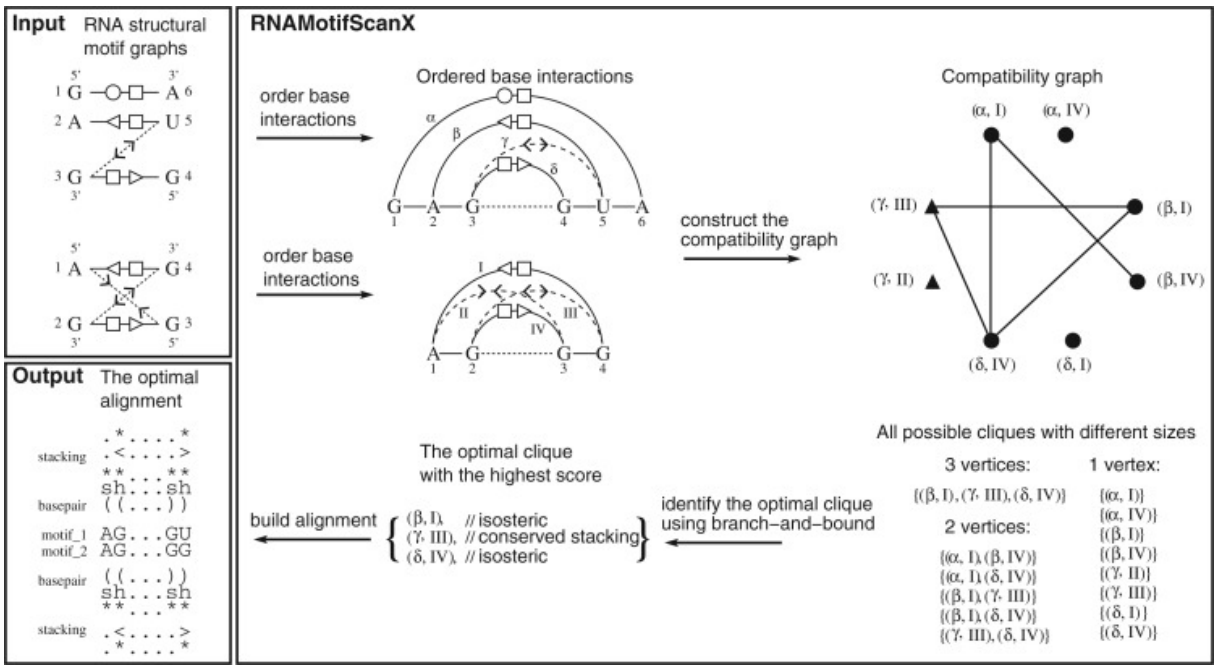


Figure 3.3: RNAMotifScanX

3.3 RegRNA

RegRNA is an integrated web server for identifying the homologs of Regulatory RNA motifs and elements against an input mRNA sequence. Both sequence homologs or structural homologs of regulatory RNA motifs can be identified. [6]

The regulatory RNA motifs supported in RegRNA are categorized into several classes:

- Motifs in mRNA 5'-UTR and 3'-UTR
- Motifs involved in mRNA splicing
- Motifs involved in transcriptional regulation

- Other motifs in mRNA, such as riboswitches
- Prediction of the splice sites, such as splicing donor/acceptor sites
- RNA structural features, such as inverted repeat
- miRNA target sites

3.4 RegRNA 2.0

RegRNA 2.0 is an integrated web server for identifying functional RNA motifs in an input RNA sequence. RegRNA 2.0 extends their previous work, RegRNA which is a widely used regulatory RNA motifs identification tool by incoporating more analytical methods and updated data sources. Through their integrated user-friendly interface, user can conveniently use these analytical approaches and observe results with good graphical visualization. Several kinds of functional RNA motifs and sites can be identified by RegRNA 2.0. [7]

3.5 CMfinder : A Covariance Model Based RNA Motif Finding Algorithm

CMfinder is a new tool to predict RNA motifs in unaligned sequences. It is an expectation maximization algorithm using covariance models for motif description, featuring novel integration of multiple techniques for effective search of motif space, and a Bayesian framework that blends mutual information-based and folding energy-based approaches to predict structure in a principled way. Extensive tests show that this method works well on datasets with either low or high sequence similarity, is robust to inclusion of lengthy extraneous flanking sequence and/or completely unrelated sequences, and is reasonably fast and scalable.

In testing on 19 known ncRNA families, including some difficult cases with poor sequence conservation and large indels, this method demonstrates excellent average per-base-pair accuracy–79 percent compared with at most 60 percent for alternative methods. More importantly, the resulting probabilistic model can be directly used for homology search, allowing iterative refinement of structural models based on additional homologs. They have used this approach to obtain highly accurate covariance models of known RNA motifs based on small numbers of related sequences, which identified homologs in deeply-diverged species. [8]

3.6 Modeling RNA Tertiary Structure Motifs by Graph Grammars

A new approach, graph-grammars, to encode RNA tertiary structure patterns is introduced and exemplified with the classical sarcin-ricin motif. The sarcin-ricin motif is found in the stem of the crucial ribosomal loop E (also referred to as the sarcin-ricin loop), which is sensitive to the a-sarcin and ricin toxins. Here, this approach generates a graph-grammar for the sarcin-ricin motif and apply it to derive putative sequences that would fold in this motif. The biological relevance of the derived sequences is confirmed by a comparison with those found in known sarcin-ricin sites in an alignment of over 800 bacterial 23S ribosomal RNAs. The comparison raised alternative alignments in few sarcin-ricin sites, which were assessed using tertiary structure predictions and 3D modeling.

The sarcin-ricin motif graph-grammar was built with indivisible nucleotide interaction cycles that were recently observed in structured RNAs. A comparison of the sequences and 3D structures of each cycle that constitute the sarcin-ricin motif gave us additional insights about RNA sequence-structure relationships. In particular, this analysis revealed the sequence space of an RNA motif depends on a structural context that goes beyond the single base pairing and base-stacking interactions. [9]

3.7 RNA Motif Search with Data-driven Element Ordering

In this paper, they study the problem of RNA motif search in long genomic sequences. This approach uses a combination of sequence and structure constraints to uncover new distant homologs of known functional RNAs. The problem is NP-hard and is traditionally solved by backtracking algorithms.

They have designed a new algorithm for RNA motif search and implemented a new motif search tool RNArobo. The tool enhances the RNAbob descriptor language, allowing insertions in helices, which enables better characterization of ribozymes and aptamers. A typical RNA motif consists of multiple elements and the running time of the algorithm is highly dependent on their ordering. By approaching the element ordering problem in a principled way, we demonstrate more than 100-fold speedup of the search for complex motifs compared to previously published tools.

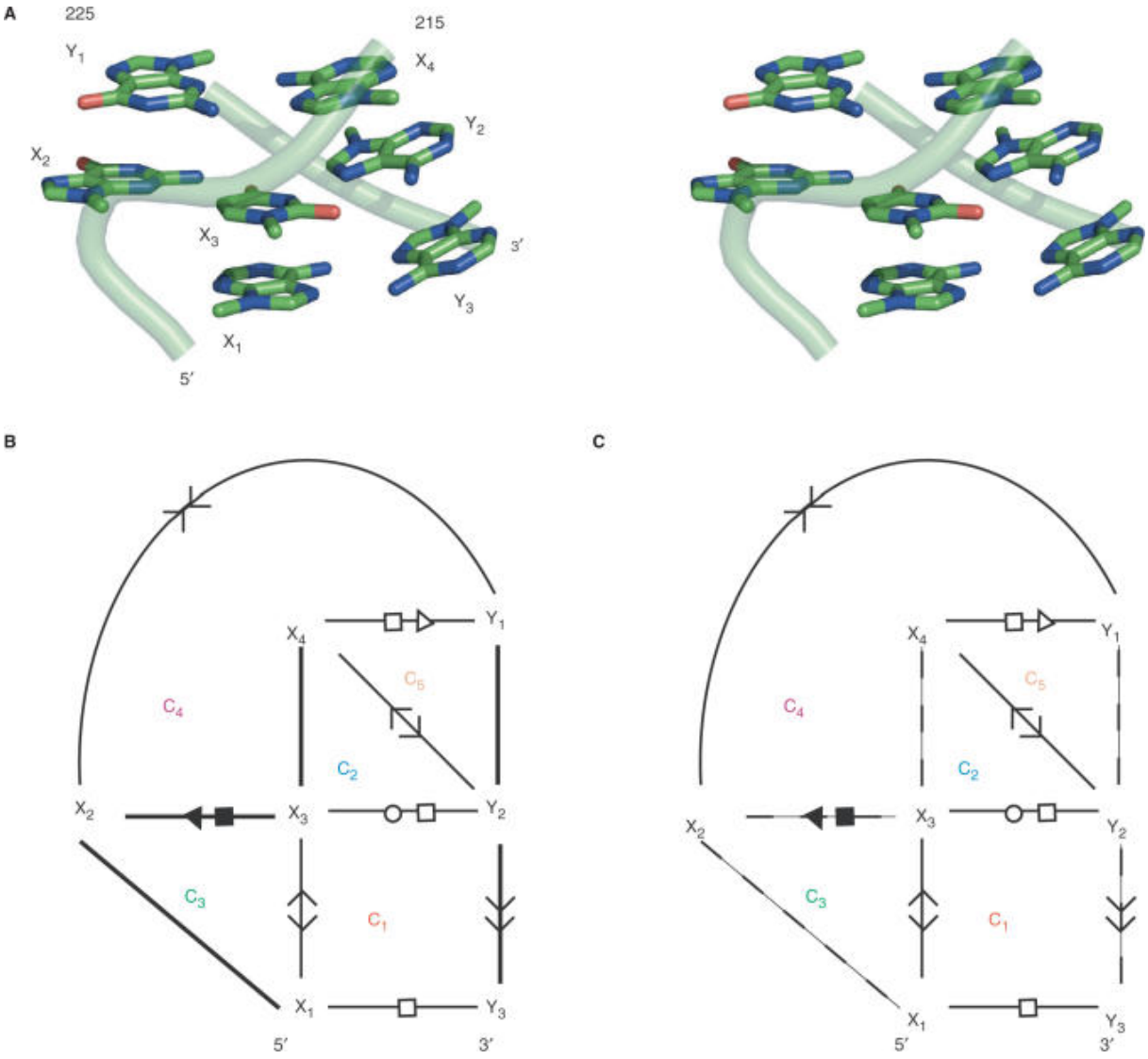


Figure 3.4: modeling RNA tertiary structure motifs by graph-grammars

They have developed a new method for RNA motif search that allows for a significant speedup of the search of complex motifs that include pseudoknots. Such speed improvements are crucial at a time when the rate of DNA sequencing outpaces growth in computing. [10]

3.8 A Novel Method for the Identification of Conserved Structural Patterns in RNA: From Small Scale to High-throughput Applications

Functional RNA regions are often related to recurrent secondary structure patterns (or motifs), which can exert their role in several different ways, particularly in dictating the interaction with RNA-binding proteins, and acting in the regulation of a large number of cellular processes. Among the available motif-finding tools, the majority focuses on sequence patterns, sometimes including secondary structure as additional constraints to improve their performance. Nonetheless, secondary structures motifs may be concurrent to their sequence counterparts or even encode a stronger functional signal.

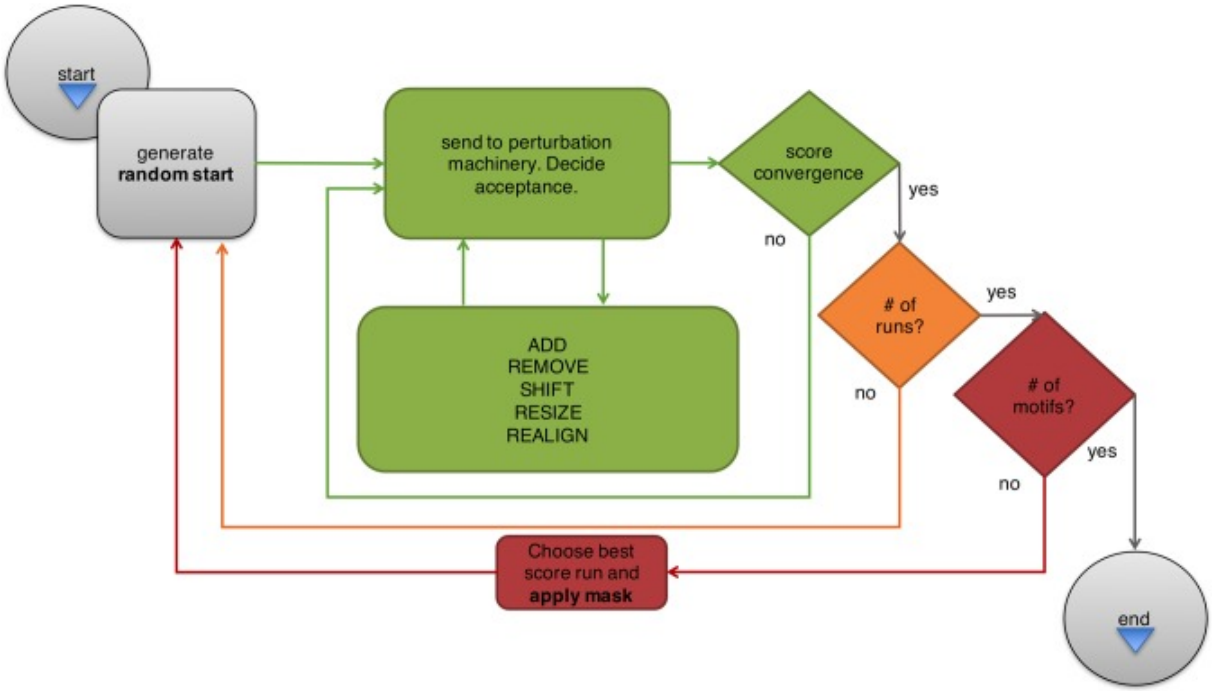


Figure 3.5: modeling RNA tertiary structure motifs by graph-grammars

Current methods for searching structural motifs generally require long pipelines and/or high computational efforts or previously aligned sequences. Here, we present BEAM (BEAr Motif finder), a novel method for structural motif discovery from a set of unaligned RNAs, taking advantage of a recently developed encoding for RNA secondary structure named BEAR (Brand nEw Alphabet for RNAs) and of evolutionary substitution rates of secondary structure elements. Tested in a varied set of scenarios, from small- to large-scale, BEAM is successful in retrieving structural motifs even in highly noisy data sets, such as those that can arise in CLIP-Seq or other

high-throughput experiments. [11]

Chapter 4

RNA Motif Identification : A Graph Traversal Approach

4.1 Edge Matrix

Edge matrix is a data structure to represent a graph similar to adjacency matrix, adjacency list, incidence matrix etc. But we are introducing this new data structure, because it saves the edges of a graph and thus its space complexity is $O(E)$. Although, for dense graph, this is not a good one to choose, for a sparse one, it might bring a drastic change in the space complexity, where adjacency matrix has a space complexity of $O(V^2)$, adjacency list has $O(V + E)$ and incidence matrix has $O(V * E)$. So space complexity of edge matrix is less than the others when $E \ll V$.

Fig. 4.2 shows an example of an edge matrix of a given graph, G in 4.1. If there is a vertex, v that has no edges with other vertices, then the edge matrix entry for this vertex will be (v,v) where both node1 and node2 of that edge is v.

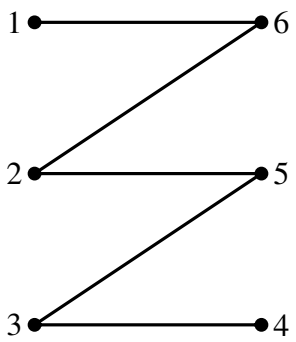


Figure 4.1: Graph G

Edge	Node1	Node2
a	1	6
b	2	5
c	2	6
d	3	4
e	3	5

Figure 4.2: Edge matrix of graph G

4.2 Graphical Representation of RNA

In the graphical representation of RNA, we replace the bases A(Adenine), G(Guanine), C(Cytosine), U(Uracil) by unique vertices and take the base-pair interactions and base-stack interactions as edges of type 1 and 2.

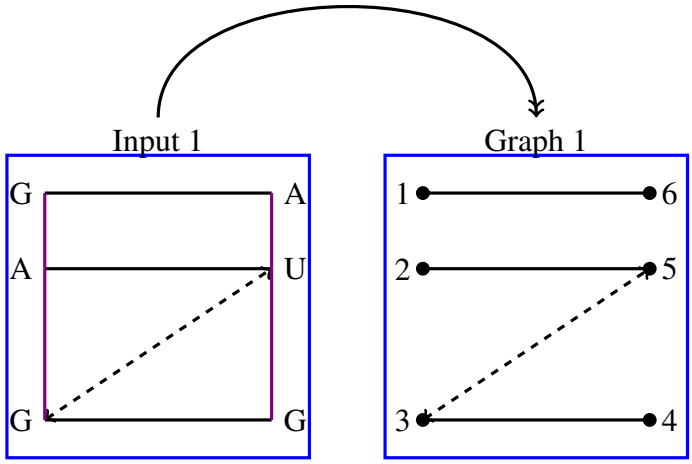


Figure 4.3: Graph1 of Input1

4.3 Degree of Edge

The degree of an edge of a graph is the number of edges incident to the end vertices of that edge. For example, from fig 4.9 we can see that, the degree of the edge (2,6) is 2, since it is connected with the edges (1,6) and (2,5).

Fig. 4.10 and fig. 4.11 shows the edge matrices of Graph1 and Graph2 with edge degrees respectively.

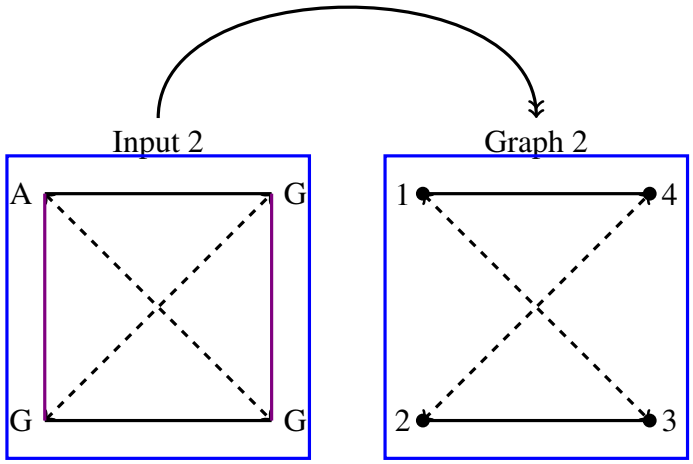


Figure 4.4: Graph2 of Input2

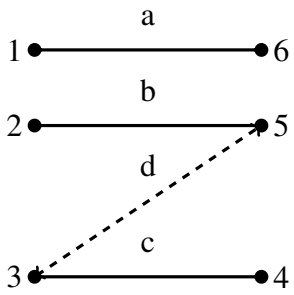


Figure 4.5: Graph1

4.4 Decreasing Order of Edges

We run edge based DFS simultaneously on both Graph1 and Graph2 every time selecting the edge having highest degree to explore. So, the edge having the maximum degree must be selected first from both of the graphs. The decreasing orders of edges of Graph1 and Graph2 are shown in fig. **??**.

Input1	Edge	d(3,5)	b(2,5)	c(3,4)	a(1,6)
	Degree	2	1	1	0
Input2	Edge	α(1,3)	β(1,4)	γ(2,3)	δ(2,4)
	Degree	2	2	2	2

Edge	Node1	Node2	Type
a	1	6	1
b	2	5	1
c	3	4	1
d	3	5	2

Figure 4.6: Edge matrix of Graph2

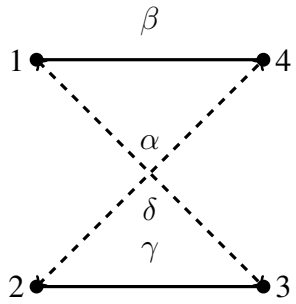


Figure 4.7: Graph2

4.5 Algorithm of RNA Motif Identification

input : array A of edge matrix of RNA
output : RNA Motif
Graph1 $\leftarrow A[1]$
i $\leftarrow 2$
while *i* $\leq A.length$ **do**
  Graph2 $\leftarrow A[i]$
  run max degree based edge DFS on both graphs Graph1 and Graph2 that results the MotifGraph
  Graph1 $\leftarrow MotifGraph$
end
return MotifGraph

Algorithm 1: RNA Motif Identification

4.6 Description of the Methodology

Let we have n RNA structures from which we have to identify the motif. These n structures are given as input of n-array of graphs represented by edge matrices. For n inputs, (n-1) iterations are done where at each iteration the simulation is run on two of them. Let, Graph1 and Graph2 are the current graphs that are dealed with. At the very beginning, the first graph element of the array is assigned to Graph1 and the loop runs for (n-1) times starting from 2. Let, the loop variable is i. At the beginning of each loop, the ith graph element of the array is assigned to

Edge	Node1	Node2	Type
α	1	3	2
β	1	4	1
γ	2	3	1
δ	2	4	2

Figure 4.8: Edge matrix of Graph2

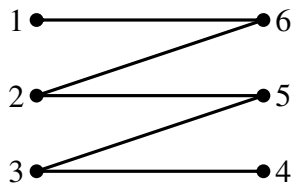


Figure 4.9: A graph

Graph1. When each loop ends, the common subgraph of Graph1 and Graph2 are assigned to Graph1 and the next graph from the array is assigned to Graph2.

Now, let us discuss the simplest case where n=2. So there is only one loop and the two inputs are assigned directly to Graph1 and Graph2. First of all, we have to compute the degree of the edges of Graph1 and Graph2. Let, there are e1 edges in Graph1 and and e2 edges in Graph2. Let us discuss elaborately the computation of finding degrees of edges of Graph1.

As we have told before, there are e1 edges in Graph1. So, to find the degree of edge a, we need to iterate from the second edge upto the last one to find the edges which are adjacent to the end points of edge a. This will take (e1-1) iterations. During these iterations, we also update the degree of the edges adjacent to edge a. For the next edge b, we do not need to look for edge a, since we have already calculated degree for the adjacency of edges a and b. So, for edge b, we need to consider the edges from the third until last. This will take (e1-2) iterations. In this way, the number of total iterations is $(n-1)(n-2)/2$ which is of $O(n^2)$.

In this way, we can compute the degree of the edges in both Graph1 and Graph2. After this, we need to concentrate the type of the edge that is given in the edge matrix. The type of an edge indicates the base-pairing interaction or base-stacking interaction. The edge indicating base-pairing interaction is drawn by solid line where the edge indicating base-stacking interaction is drawn by dashed line. These edges are to be handled seperately, that means when we run DFS simultaneously on both Graph1 and Graph2, the edges chosen to be explored must be of the same type.

Now, let us discuss about the DFS that should be run simultaneously on both graphs. The traversal is basically the Depth First Search, but this is an edge-based traversal. Now, the first

Edge	Node1	Node2	Type	Degree
a	1	6	1	0
b	2	5	1	1
c	3	4	1	1
d	3	5	2	2

Figure 4.10: Edge matrix of Graph1 with edge degrees

Edge	Node1	Node2	Type	Degree
α	1	3	2	2
β	1	4	1	2
γ	2	3	1	2
δ	2	4	2	2

Figure 4.11: Edge matrix of Graph2 with edge degrees

edge to be selected must be the edge with the highest degree. This is true for both the graphs and again their types must be the same one. After exploring these edges, the next edge to explore is decided depending on the degree of next edges, that can be explored, and the type of edges. The edges that are to be selected should be of highest degree as well as of the same type. If any of the edges can be no more explored and need to return to its previous edge, the same steps should be taken in the other graph though it may have opportunities for the edge to be explored. From 4.12, we can see that, firstly edge d in Graph1 and edge α in Graph2 is selected. Then, edge c in Graph1 and edge β in Graph2 is selected. Now, although edge β can explore more in depth, edge c does not have that opportunity in Graph1. So, we must return back to edge d in Graph1 and edge α in Graph2. Similarly, we next explore edge b in Graph1 and edge γ in Graph2. Finally, there are no more edges to be explored. So, this is the common subgraph, or more definitely the motif present in these RNA structures.

4.7 Analysis of Our Approach

Through RNA motifs are really important in biological processess , there is not many efficient searching tools for them. This is due to the complex structure of motifs. In our thesis work we tried to find a solution to reduce the structural complexity and to find both partial and fully observable motifs. Our propsed grpah traversal technique methodoly was explained in the previous chapter. In this section we analyze different aspects of our proposed approach.

4.7.1 Input Strategy

We took the RNA sequences as graphs to reduce input complexity. We repsesented each nucleotide base as vertex and each base pair or base stacking as edges. We differentiated between two types of edges as type 1 and type 2 respectively. This helps to ensure an easy representation of the complex RNA sequence. We proposed an edge matrix as input. This edge matrix helps to take input in space complexity O(E) where E is the number of edges. This structure reduces space complexity better than adjacency matrxi, vertex matrix etc.

Data structure	**Space complexity**
Adjacency matrix	$O(V^2)$
Adjacency list	$O(V + E)$
Incidence matrix	$O(V * E)$
Edge matrix	$O(E)$

4.7.2 Analysis of Our Algorithm

After taking the RNA sequences input as graphs we calculated the Degree of each edge. As the graph is a sparse one so this takes time complexity of O(E). As we want to run depth first search process baes on descending order of degree, for this we need rearrange the edges with respect their degree's descending order. To do so we run quicksort on degree of egdes. We get an edge matrix organized with respect to their degree after sorting. The sorting process takes O(ElogE) time. Then we do depth first search on the two graphs simultaneously where the starting node is the edge with maximum degree in two graphs. The algorithm chooses a path if the two edges are of same type. This depth first search takes O(V+E) time complexity.

Our main approach watch to find a subgraph isomorphism between two graphs representing RNS sequences. But as it is a NP hard problem we tried to solve the search space using degree of edges as a constraint.

4.7.3 Pitfalls of Our Approach

The graph traversal approach that we proposed progresses using descending order of degree of edges. But it doesn't work efficiently when degrees of edges are same. It may take a wrong path in that case. thus traversing wrong edges. So out approch aligns with pattern recognition in some way where degree is like a feature. To make the algorithm more accurate and efficient

we need to find more constraints to work with.

4.7.4 Comparison with Related Works

n chapter 3 we gave description of many other RNA motif identification techniques. Here we compare our proposed solution with RNAMotifScanX algorithm which follows a graph alignment approach.

In RNAMotifScanX also takes the RNA sequence inuput as grpahs. Then it finds 6 types of relationship in the graphs separately : juxtaposing, juxtaposing with shared nulceotide, crossing, enclosing, left align, right align. Using these 6 relationships as nodes for both graphs this tools draws an matvhing graph. Then it runs Bron and Kerbosch algorithm to find maximum clique from the matching graph. This maximum clique is the desired motif or similar sequece.

It takes O(N*N) time to find the 6 types of relationship from both graphs where N is the number of nodes in the graphs. Constructing the matching graph takes O(A,B) time where A=no of relationships in one graph and B= number of relationships in second graph. Finally the clique finding algorithm takes $O(dn3^{(d/3)})$, where d is the degeneracy of the graph, a measure of its sparseness and n is the number of vertices.

Our proposed algorithm takes O(E^2) time at degree finding step. Then the sorting process takes $O(ElogE)$ time. FInally the DFS takes $O(E + V)$ time. Though the both processes are different in nature but RNAMotifScanX includes more computational processes than ours. Also RNAMotifScanX is totally computaional. It prunes the graph with respect to the 6 types of relationships. Our proposed method is computational model. Our proposed process prunes the grpah with respect to degree of edges.

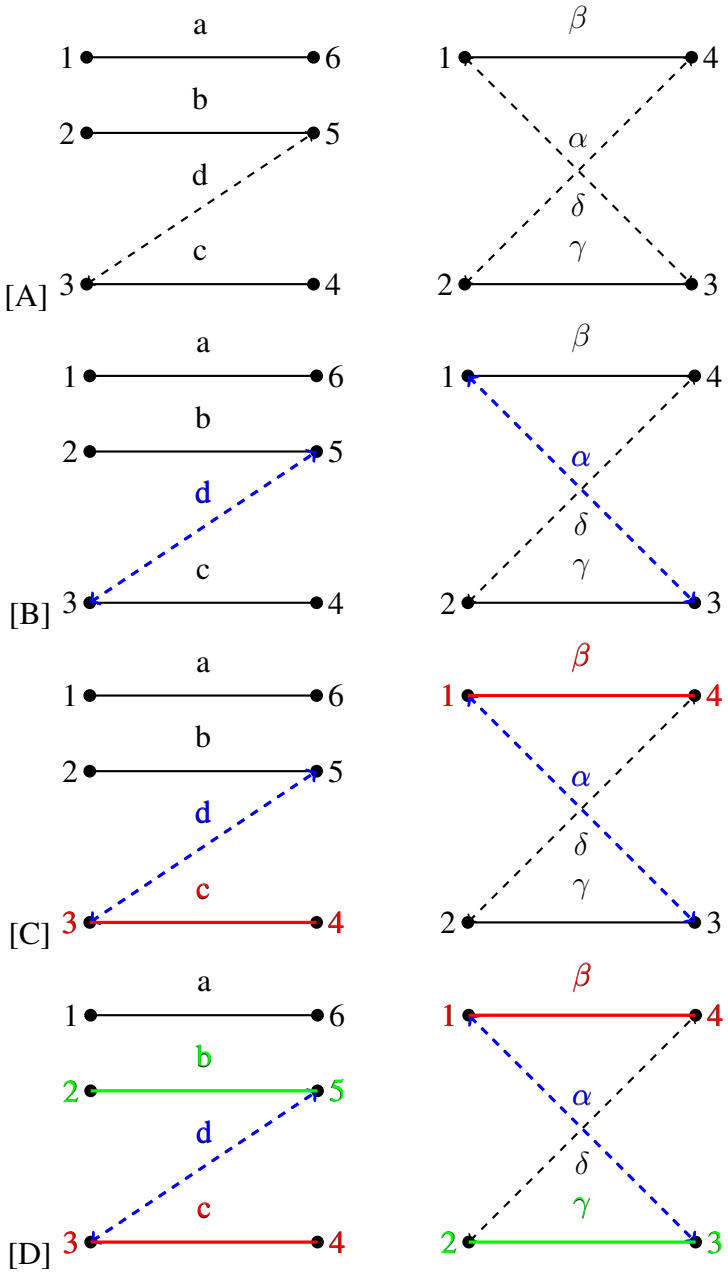


Figure 4.12: Simulation of RNA Motif Identification; [A] Initialization; [B] edge d in Graph1 and edge α in Graph2 is selected; [C] edge c in Graph1 and edge β in Graph2 is selected; [D] edge b in Graph1 and edge γ in Graph2 is selected;

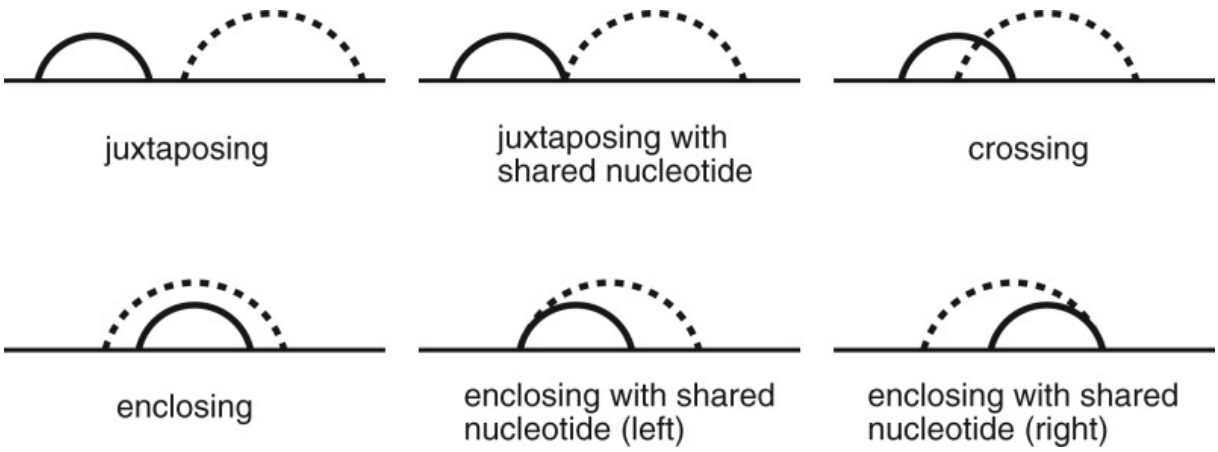


Figure 4.13: finding motif

Chapter 5

Conclusion

5.1 Findings

In this book an effective scheme to recognise RNA secondary structural motif is proposed. This proposal is a graph traversal approach that enables automatic identification of both partially and fully matched motif instances. To construct this algorithm we first used graph representation of the RNA sequences using our proposed edge marix. Here each nucleotide base was defined as a vertex and any base pair bond or base stacking bond was defined as an edge. The two types of bonds were labeled as type 1 edge and type 2 edge respectfully. Then we calculated the degree of each edge and ran a sorting process on the edges based on the degree. Then we ran depth first search algorithm on the both graphs simultaneously by traversing the edges with maximum degree. In our proposed process we choose a edge that is both graph and is of same type. The searching prcess ends when we return to the starting edge after bachtracking. The process often fails to work well when the degree of edges of a graph is same and it ends up choosing a wrong edge to traverse. Due to this the accuracy is not high enough. So in future we consider the implementation of more constraints other than degree to amke better accuracy of this method.

5.2 Future Study

Number of researchers worked with RNA motif identification problem but there is only a few searching tools that finds RNA motif effectively and optimally. Efficient RNA motif tools are lacking due to the high complexity of the structure of these motifs. In our research work we wanted to adopt a subgraph isomorphisom technique for finding similar parts of the graphs. Though it is a NP hard problem it works for graphs with less edges. This technique is infeasible for graphs with greater number of nodes as we have to consider all psible combination of nodes. For ths reason we tried to put a constraint on the search space by ruuning a depth first search on

the basis of degree of edges. As this algo traverses the graphs in the basis of degree it doesnt work well same degrees of edges. This is why we need to introduce more constraint on the searching process. Further research work in this field can help to make the algorithm more efficient. Also The current data structure reduces space complxity but in depth first search process it increases search complexity. So future research works can help to find a more accessable data structure. So the below mentioned points can be improved by future study:

- Finding more constraints that reduces ambiguity and search space.
- Finding a more efficient data structure for depth first search process.

Our future study will be on finding satisfactory ccnstraint approach, data structure and if possible any other grapg traversal technique that will eventually maximize our RNA motif identification rate to a satisfactory level.

References

[1] V. B. Zhurkin, M. Y. Tolstorukov, F. Xu, A. V. Colasanti, and W. K. Olson, "Sequence-dependent variability of b-dna," *DNA conformation and transcription*, pp. 18–34, 2005.

[2] D. K. Hendrix, S. E. Brenner, and S. R. Holbrook, "Rna structural motifs: building blocks of a modular biomolecule," *Quarterly reviews of biophysics*, vol. 38, no. 3, pp. 221–243, 2005.

[3] A. Achar and P. Sætrom, "Rna motif discovery: a computational overview," *Biology direct*, vol. 10, no. 1, p. 61, 2015.

[4] C. Zhong, H. Tang, and S. Zhang, "Rnamotifscan: automatic identification of rna structural motifs using secondary structural alignment," *Nucleic acids research*, vol. 38, no. 18, pp. e176–e176, 2010.

[5] C. Zhong and S. Zhang, "Rnamotifscanx: a graph alignment approach for rna structural motif identification," *RNA*, vol. 21, no. 3, pp. 333–346, 2015.

[6] H.-Y. Huang, C.-H. Chien, K.-H. Jen, and H.-D. Huang, "Regrna: an integrated web server for identifying regulatory rna motifs and elements," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W429–W434, 2006.

[7] T.-H. Chang, H.-Y. Huang, J. B.-K. Hsu, S.-L. Weng, J.-T. Horng, and H.-D. Huang, "An enhanced computational platform for investigating the roles of regulatory rna and for identifying functional rna motifs," *BMC bioinformatics*, vol. 14, no. 2, p. S4, 2013.

[8] Z. Yao, Z. Weinberg, and W. L. Ruzzo, "Cmfindera covariance model based rna motif finding algorithm," *Bioinformatics*, vol. 22, no. 4, pp. 445–452, 2005.

[9] K. St-Onge, P. Thibault, S. Hamel, and F. Major, "Modeling rna tertiary structure motifs by graph-grammars," *Nucleic acids research*, vol. 35, no. 5, pp. 1726–1736, 2007.

[10] L. Rampášek, R. M. Jimenez, A. Lupták, T. Vinař, and B. Brejová, "Rna motif search with data-driven element ordering," *BMC bioinformatics*, vol. 17, no. 1, p. 216, 2016.

[11] M. Pietrosanto, E. Mattei, M. Helmer-Citterich, and F. Ferrè, "A novel method for the identification of conserved structural patterns in rna: From small scale to high-throughput applications," *Nucleic acids research*, vol. 44, no. 18, pp. 8600–8609, 2016.

Generated using Undegraduate Thesis LaTeX Template, Version 1.3. Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.

This thesis was generated on Thursday 21st November, 2019 at 8:22pm.