MASTERARBEIT | MASTER'S THESIS

Titel | Title

Computational Studies of Exoribonuclease-Resistant RNA Structures in Flaviviruses

verfasst von | submitted by

Denis Skibinski B.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

This thesis is dedicated to my beloved fiancée and family members for their limitless support and encouragement in this turbulent time

ACKNOWLEDGMENTS

ABSTRACT

We focused on the computational analysis of exoribonuclease resistant RNA (xr-RNA) structures in flaviviruses. We aimed to understand the particularities of the xrRNA structure across the different flaviviral groups and establish a foundation for producing artificial xrRNA. Consensus secondary and 3D structures of xrRNAs were predicted and analysed for each flavivirus subgroup. We reveal distinct similarities of the xrRNA in flaviviruses. We show that a multi-loop with a three-way junction and two pseudoknots is a common motif in the secondary structure of flaviviruses. In the 3D structure of xrRNA we show that the ring-like motif is conserved between the flaviviruses, which is assumed to be the major mechanism in resisting the degradation of the host enzyme XRN1. With our findings we lay out a guideline that can be followed to create artificial xrRNA sequences. These could be used in developing xrRNA based therapeutic applications like integration into a riboswitch, enabling a robust on and off switch for gene expression or overall increasing the stability of RNA therapeutics.

Keywords: RNA, Flaviviruses, xrRNA, XRN1, ViennaRNA

ABSTRAKT

Diese Masterarbeit beschäftigt sich mit der Analyse von Exoribonuklease resistenten RNA (xrRNA) Strukturen in Flaviviren. Hierbei versuchten wir die Sekundär- und Tertiärstruktur der xrRNA in den verschiedenen Flavivirengruppen zu analysieren und eine Grundlage für die Herstellung künstlicher xrRNA zu schaffen. Im Rahmen der Masterarbeit wurden Konsenssekundär-Strukturen und 3D-Strukturen von xrRNAs vorhergesagt und für jede Flavivirusgruppe analysiert. Hier konnten deutliche Ähnlichkeiten der xrRNA Strukturen in den verschiedenen Flavivirusgruppen aufgezeigt werden. Des Weiteren wurde demonstriert, dass eine Multiloop mit drei Stems und zwei Pseudoknoten ein zentrales Motiv in der Sekundärstruktur der xrRNA von Flaviviren darstellt. In der 3D-Struktur zeigen wir, dass eine eindeutige ringartige Struktur der xrRNA zwischen den Flaviviren konserviert ist. Es wird angenommen, dass das der Hauptmechanismus ist, der dem Abbau durch das Enzym XRN1 widersteht. Mit unseren Erkenntnissen legen wir einen Leitfaden vor, der zur Herstellung künstlicher xrRNA verwendet werden kann. Diese künstliche xrRNA könnte für die Entwicklung von xrRNA basierten therapeutischen Anwendungen verwendet werden, beispielsweise für die Integration einer xrRNA Struktur in einen Riboswitch, oder für die Erhöhung der Stabilität von RNA-Therapeutika.

Keywords: RNA, Flaviviruses, xrRNA, XRN1, ViennaRNA

Contents

Nomenclature

CM   Covariance Model

ISFV Insect-Specific Flavivirus

IUPAC International Union of Pure and Applied Chemistry

MBFV Mosquito-Borne Flavivirus

MFE   Minimum Free Energy

MIS   Most Informative Sequence

MSA   Multiple Sequence Alignment

NKV   No Known Vector Flavivirus

PCR   Polymerase Chain Reaction

RMSD Root Mean Square Deviation

RNA   Ribonucleic Acid

RpRP RNA-Dependent RNA Polymerase

sfRNA Subgenomic Flavivirus RNA

TBFV Tick-Borne Flavivirus

UTR   Untranslated Region

XRN1 Exoribonuclease 1

xrRNA Exoribonuclease Resistant RNA

Chapter 1

Overview

1.1 Introduction

The field of bioinformatics and virology have made substantial progress in understanding viral RNA over the years. This resulted in advancements in therapeutic applications and especially vaccine development. Among these RNA, the exoribonuclease-resistant RNA (xrRNA) in flaviviruses is an area of study due to its unique mechanisms and roles in disabling the host enzyme XRN1. This thesis focuses on computationally analysing xrRNA structures across different flaviviral groups. The objective is to understand a bit more about their complex architectures and explore the potential for developing artificial xrRNAs for therapeutic purposes.

1.2 Research Gap and Motivation

Our initial research identified a gap in the understanding of xrRNA structures across flaviviruses. Although the two-dimensional structures of xrRNAs for specific viruses within the mosquito-borne flaviviruses (MBFV), tick-borne flaviviruses (TBFV), no-known vector flaviviruses (NKV), and insect-specific flaviviruses (ISFV) have been predicted, a detailed 2D consensus structure for each flaviviral group has not yet been established. The xrRNA's three-dimensional structure has only been instrumentally analysed for MBFV and only broadly hypothesized for the other flaviviral group. Our research goal is to analyse all the structural elements of xr-

RNA. The final objective is to create a foundational guideline that could lead to the development of artificial xrRNA.

1.3 Objective of the thesis

The work done in this thesis can be divided into three main objectives:

Objective 1: Prediction and Analysis of 2D Consensus Structures. The first objective is to predict and analyze a consensus structure of xrRNA for each subgroup of Flaviviruses. This involves a methodological approach for structure prediction and a comparative analysis framework. First we search for a suitable database and predictive models used for xrRNA identification and defining the selection criteria for xrRNA sequences. Then we focus on the 2D structure, only then do we try to understand the 3D structure. By understanding the consensus structures, we aim to identify variations and conserved motifs of the xrRNA structure in and between the different flaviviral groups.

Objective 2: Statistical Analysis of Consensus Structures. The second objective is to perform a statistical analysis of the consensus structures. This includes defining the statistical methods used and explaining their relevance to understanding the structure of xrRNA. Through this analysis the thesis aims to provide a quantitative basis for interpreting the importance of different parts of the xrRNA structure.

Objective 3: Exploration of Artificial xrRNA in Therapeutics. The third and final objective is to explore the potential of artificial xrRNA in therapeutic applications. This includes discussing the criteria for the design of artificial xrRNA and outlining the steps for testing their efficacy. The research goal is to check if the utilization of xrRNA structures in therapeutical methods is feasible and explore theoretical fields of applications.

This research goal is to enhance our understanding of the structural diversity and similarities of xrRNA in Flaviviruses. This will contribute to the development of future xrRNA based therapeutics. For example an improvement of mRNA vaccine development or gene therapy through the development of xrRNA Riboswitches.

1.4   Outline of the Thesis

This thesis can be split into six chapters. Each Chapter focuses on specific aspects of the analysis and characterization of xrRNA in the Flavivirus family.

Chapter 2 Literature Review, creates a basic understanding of RNA and tries to give an overview of the current knowledge of xrRNA structures in flaviviruses. It also explores the role of XRN1 in RNA degradation. In the end it provides an overview on the computational tools used for RNA analysis in our study.

Chapter 3 Materials and Methods, shows the theoretical and computational methods used to analyze the xrRNA structure. This chapter provides an in-depth explanation of the tools and procedures used, with a primary focus on the utilization of the ViennaRNA Package for RNA secondary structure prediction and analysis. The usage of covariance model, the process of obtaining Flavivirus sequences from a database, the methodology for predicting a 2D consensus structure, creating a 3D model of a sequence and statistical techniques for structure analysis are explored.

Chapter 4 Results, presents the results of the analysis process. This section presents the features of the xrRNA consensus structure for different Flavivirus families in 2D and 3D. It also includes the outcomes of the statistical analysis and the identification of similarities.

Chapter 5 Discussion, discusses the results in the context of the scientific literature, exploring their implications for our understanding of RNA structure based on the 2D consensus structure, the 3D structure, and the statistical analysis. This section discusses the implications of artificial xrRNA design and the limitations of the study.

Chapter 6 Conclusion, summarizes the key findings of this thesis and showcasing future research goals in this area.

Chapter 2

Literature Review

2.1 Introduction to RNA

2.1.1 An Overview

Ribonucleic acid (RNA) is a biopolymer that plays a vital role in all biological life. Its monomers are made up of nucleotides. These nucleotides contain a ribose sugar, a phosphate group, and either a purine (adenine or guanine) or a pyrimidine (cytosine or uracil). The RNA backbone is negatively charged and consists of phosphodiester bonds between one carbon atom of a nucleotide's ribose and one phosphate group of another adjacent nucleotide, giving the RNA a direction. RNA is normally single-stranded, allowing it to fold into structures with itself. Its secondary and tertiary structure are determined by the sequence of bases and the rules of complementary base pairing (adenine with uracil and guanine with cytosine). The formation of base pairs is guided by hydrogen bonds. These hydrogen bonds between nucleotides are part of the overall stability and specificity of the RNA secondary structure [1].

The secondary structure of the RNA is the local spatial arrangement of a single RNA strand. The primary stabilizing energies involve base stacking between adjacent nucleotides and base pairing between complementary nucleotides. The stem-loop structure comes into existence when an RNA strand folds back on itself, causing complementary bases to pair and form double-stranded helices [2]. For a better understanding of the energy in the secondary structure, the RNA structure can be decomposed into different types of loops [3], as shown in Figure 2.1.
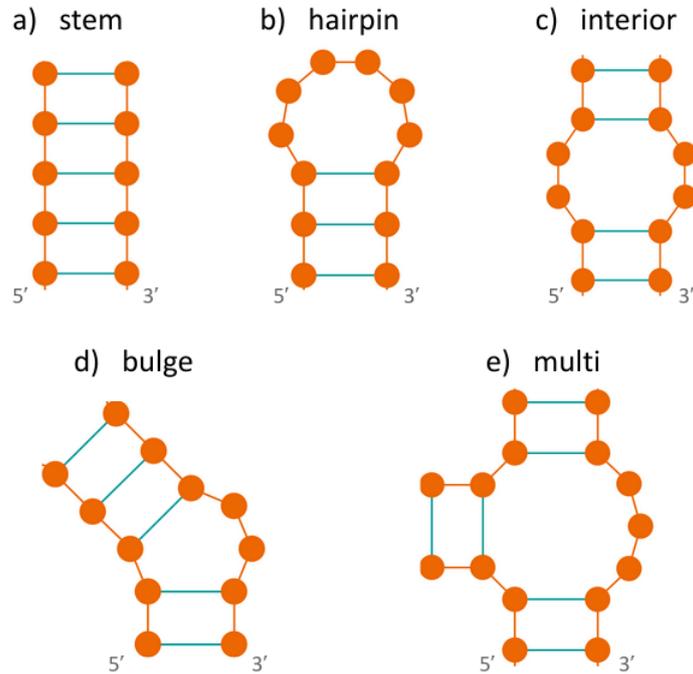
Figure 2.1: **A loop decomposition diagram.** The orange nodes are nucleotides which are connected through a phosphodiester backbone with the neighboring nucleotides. The green connections are the hydrogen basepairs the nucleotides build with their canonical partner.

- **Stem loop** structures are a double-stranded stem with complementary base pairing. The stability comes from the base stacking energy of the nucleotides in the stem. The energy is lower for more stable, longer and GC-rich stems.

- **Hairpin loop** structures are similar to stem loops. They consist of double-stranded stem but with an unpaired loop at the end of, giving them a hairpin shape. This structure is formed when the RNA molecule folds back on itself, forming a stable stem through base pairing and leaving a loop of unpaired nucleotides at the end. The energy of the hairpin loops depends on the size and sequence of the loop. Smaller loops tend to have higher energy due to the strain caused by tight turns in the molecule, while larger loops tend to have lower energy. This trend is reversed when the loop size exceeds a certain threshold.

- **Interior loop** structures are formed when there are unpaired nucleotides on either side of a double-stranded region, disrupting the regular helical structure.

5

The energy of the inner loop can be changed by the size of the loop, the identity of the unpaired nucleotides and the sequence context.

- **Bulge** structures happen when one or more nucleotides on one side of a double-stranded region are unpaired, resulting in a distortion or 'bulge' in the regular stem structure. The energy of the bulges varies depending on the number and type of unpaired nucleotides, the sequence context and the overall conformational flexibility of the bulge. Bulges can potentially disrupt the stability of the RNA structure and are typically related with higher energy states.

- **Multi-loop** structures consist of multiple stems connected by stems. They can be thought of as multiple internal loops connected by the same junction. These structures are highly flexible and their energy varies depending on the number of stems and loops, the sequence and the flexibility of the individual stems and loops. Multi-loops can form complex 3D RNA structures and can act as key determinants of RNA function.

Loop decomposition will be important for comparing different flaviviral xrRNAs. When creating the different loop types, only canonical base pairs that do not cross each other are considered. RNA structures also contain non-canonical base pairs, such as base triples and pseudoknots. These can have important stabilizing function for the tertiary structure of RNA [3].

Base triples are formed when a third nucleotide joins an existing base pair. These base triples contribute to the formation of complex 3D RNA structures. The energetics of these triples could be affected by the type of nucleotides involved and the overall conformation of the RNA molecule [4].

Pseudoknots are formed when nucleotides form a base pair that crosses with other base pairs in the secondary structure. This creates interwoven loops in the RNA. Pseudoknots in RNA can regulate the gene expression or viral RNA replication. The stability and formation of a pseudoknot depends on the nucleotide composition and length of the base pairing. As GC-rich pseudoknots are more stable than others [5]. In our study of xrRNA, we will encounter some pseudoknots that are essential for the structure of xrRNA and its mechanism of switching off the host immune response.

2.1.2 RNA Secondary Structure Visualization

Visualising the RNA secondary structure is important in understanding the function of an RNA. The secondary structure of RNA can be represented in a number of ways, each offering different advantages. One of the most common ways of representing RNA structure is in dot-bracket notation, also known as Vienna notation. This is a linear text format that provides a simple way of describing the pairing of nucleotides. In this format, paired nucleotides are represented by matching brackets "(" ")" and unpaired nucleotides are represented by dots "." [6] as seen in Figure 2.2. To visualise pseudoknots one could use another type of matching brackets "[" "]".

[7] introduced the Abstract Shape as a novel way to view the RNA folding space, that is based on the dot-bracket notation. This view focuses on the structural layout of the structure, abstracting from sequence details. It retains structural motifs such as hairpins and multi-loops, which can be important for understanding RNA function and differentiating RNA molecules. This visualization aids in efficient computational analysis and comparison of RNA structure. Different levels of coarse-graining can be chosen based on the required level of detail for research purposes.

For more visually intuitive representations, RNA secondary structures can be displayed as graphs. This graphical method uses nodes for nucleotides and lines for covalent bonds between adjacent nucleotides for the backbone (in black) and non-adjacent nucleotides for base pairs (in blue Figure 2.2). Various layout algorithms can be used to generate such a graph, the most common being the planar graph layout. This layout minimises base pair crossings and overlaps, making the structure easier to analyse [8].

Dot-plots and mountain plots are both visualisation techniques for RNA secondary structure. Dot-plots display possible base pairings in a two-dimensional matrix, showcasing the diversity of the structure ensemble and pairing probabilities. Mountain plots show the structure as a line graph, representing the RNA structure based on base pairings, where peaks and valleys correspond to structural features [9]. Although not used in this work, I wanted to mention them for completeness.

```
graphic planar            sequence
                          GCCGGCAAAAGCGAAAACGCUGCGCAAUAGCGCAAAGCCAGGC

                          dot-bracket
                          ((((((....(((....))).(((....)))...))).)))

                          Abstract shape
                          lvl 1 [[_[]_[]_]_]
                          lvl 2 [[[][]]_]
                          lvl 3 [[[][]]]
                          lvl 4 [[][]]
                          lvl 5 [[][]]
```
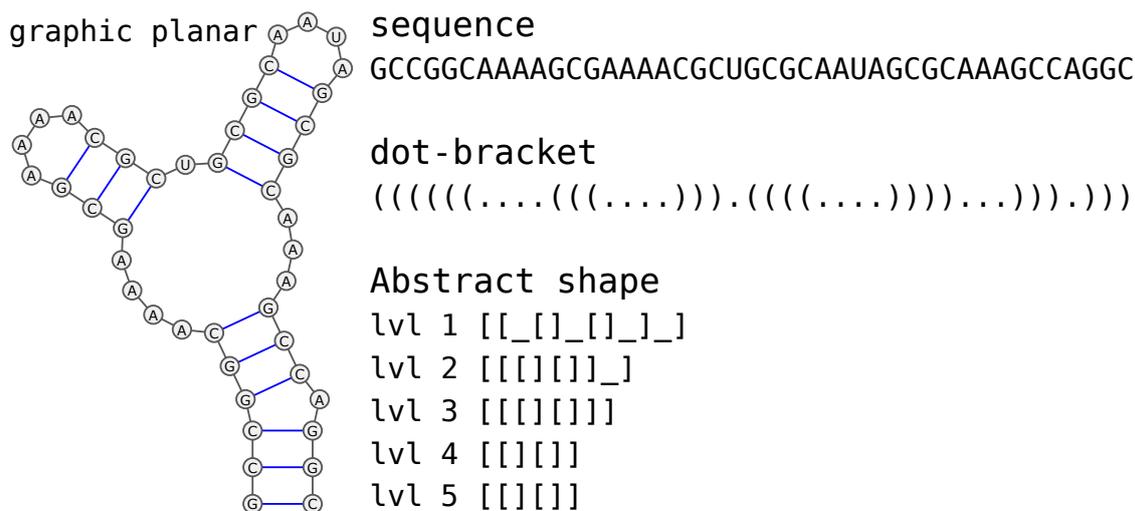
Figure 2.2: **Three styles of visualisation of RNA secondary structure**. On the left is the graphical planar layout, on the middle right is the dot-bracket notation, and on the bottom right is the abstract shape as calculated with the abstract_shape function of the ViennaRNA package.

2.1.3   RNA Secondary Structure Prediction

Understanding the structure of RNA is fundamental to understanding its function in a cell and its potential interactions. RNA structure prediction is a computational method used to find the most likely biologically functional secondary and tertiary structure of RNA based on the nucleotide sequence. Since RNA folding is a hierarchical process in which the secondary structure is built before the tertiary structure, it is important to accurately predict the secondary structure first. Various approaches can be used to predict the secondary structure of an RNA sequence, such as stochastic, machine learning and minimum free energy methods. This thesis focuses on the minimum free energy (MFE) method as used in RNAfold of the ViennaRNA package. The MFE method assumes that the biological functional structure is the one with the lowest energy [10]. Computational tools such as RNAfold use a dynamic programming algorithm to predict the secondary structure of RNA [11]. RNAfold divides the RNA sequence into smaller segments and the structure of each segment is predicted. These predictions are then combined to find the most stable overall structure. To calculate the energy of the smaller segments, these tools rely on experimentally analysed thermodynamic parameters, the most prominent of which

8

is the Turner parameter. The energy assigned to each possible RNA structure is based on the sum of the loop element energies. These consist mainly of the stacking energy of adjacent nucleotides and in buldges for example on the length of the unpaired region. The functional structure is the one with the lowest total energy [12].

Although MFE methods became the state-of-the-art when predicting RNA secondary structures, they have limitations. They often predict only the most stable structure, which is not necessarily the biological functional structure due to its higher energy and small differences. The accuracy of the MFE method is highly dependent on the correct and complete thermodynamic parameters used [10].

2.1.4 Energy Parameter Turner and Andronescu

Accurate prediction of RNA secondary structure using the MFE method is dependent on the energy parameters used in the computational models. These models use different types of energy, including base pair energies, base stacking energies, and loop energies, which include the energies of different types of loops in RNA structures, such as hairpin loops, internal loops, bulges and multiple loops. These include mismatch and terminal mismatch energies, which account for the energy contribution when non-Watson-Crick pairs are adjacent to helices or at the end of helices, pseudoknot energies and, most importantly, stacking energies, which represent the energy contributions of stacked base pairs. Two of the most prominent energy parameters used are those developed by Turner and later based on Turner by Andronescu. The Turner parameters, developed in 1980, are based on thermodynamic studies [12]. These parameters have been continuously updated, the last update being in 2004 [13]. Currently, the Turner parameters are widely used as standard parameters for predicting secondary structure with computational tools such as RNAfold. However, these tools allow the user to change the parameter model used, such as the Andronescu parameters.

The Andronescu parameters do not replace the Turner parameters, but build on and refine them. They have been developed to address some of the limitations of the Turner model, notably in the context of predicting more complex RNA structures. While the Turner parameters provide a good basic model for predicting RNA

structures, Andronescu's work focuses on more complex aspects of RNA structures, such as multi-loops and pseudo-knots. This is achieved by using more experimental data and refining the energy model using machine learning algorithms [14]. When predicting the secondary structure of RNA, it is advised to recognise the limitations and advantages of the energy parameters and to assess whether another model would give more fitting results.

2.1.5 RNA Consensus Structure Prediction

Consensus structure prediction can increase the likelihood of identifying the biologically functional secondary structure of an RNA sequence. This method predicts a consensus structure from an alignment of similar RNA sequences, rather than predicting the secondary structure from a single sequence. Consensus predictions are more reliable as they reduce sequence-specific errors and they find conserved structural features more reliable. Predicting a single RNA sequence can often lead to errors due to its variability and complexity. However by analyzing a set of similar sequences, consensus prediction can smooth out these anomalies. This can result in a more reliable and representative structure. A consensus structure can help in deducing the function of unknown RNA sequences by comparing their sequence to one with a known structure. The identification of conserved RNA structures across species can reveal evolutionarily structural motifs [15]. To improve consensus structure prediction, a computational tool such as RNAalifold from the ViennaRNA package can combine sequence from multiple sequence alignment (MSA) and structural data from covariance analysis [11].

2.1.6 RNA Multiple Sequence Alignment (MSA)

Having discussed the importance of RNA secondary structure prediction and consensus structure prediction, it is essential to explore RNA MSA. MSA is used for example in comparative genomics and RNA structure prediction. As mentioned in the section on consensus structure prediction, alignment of multiple RNA sequences improves secondary structure prediction. MSA is a useful tool for identifying conserved or covarying regions across multiple sequences, which often reveal important functional and structural features. By aligning these sequences, MSA can reveal

10

structural motifs that may be missed by single sequence analysis. The process of MSA involves aligning two or more sequences to identify small regions. This alignment is achieved by a computational method that optimises the similarity between the different sequences. Computational tools often keep track of the alignment score and penalise gaps and mismatches that happen in MSA. A common approach is progressive alignment, where we start with the most similar pair of sequences and add others based on similarity [16]. Tools such as ClustalW uses heuristics to efficiently align large numbers of sequences. These heuristics, such as progressive alignment, pairwise sequence comparison, or various gap penalty schemes, are computationally efficient but do not guarantee to find the optimal alignment due to their approximate nature [17]. To improve the accuracy of MSA tools such as locARNA, secondary structure information is incorporated into the alignment process. This is helpful in regions with conserved structure but varying sequences [18]. The MSA provides several insights that cannot be obtained from a single sequence analysis. It can highlight conserved sequence segments or structural clues of functional importance. On the other hand, it could reveal variable regions suggesting functional diversification, which could be important for species-specific adaptation, such as the creation of the diverse range of hosts and vectors that flaviviruses have [19]. MSA could also provide a means of analysing the phylogenetic relationship between RNA sequences. Conserved sequences and structures between different species could suggest a common evolutionary origin or ancestor [16].

2.1.7 Sequence Logo

The sequence logo is a graphical representation of a MSA and was presented as early as 1990. The sequence logo is a frequency table showing the count of each nucleotide at each position in the alignment. The nucleotides are then sorted by frequency, with the most frequent nucleotide at each position appearing at the top of the sequence logo graph. If only the top nucleotides are read, the sequence consensus is obtained. The visualisation allows clear identification of sequence patterns and provides insight into the variability and conservation of RNA sequences [20].

2.1.8 Covariance Analysis

In bioinformatics, covariance is the correlated changes of base pairs of nucleotides in an RNA MSA between sequences. So when two paired nucleotides in an RNA sequence show covariance, it infers that changes at one position are compensated by changes at another position to maintain their base pairs [21]. Applying covariance analysis to MSA can identify pairs of nucleotides that covary between different species. This analysis can help find conserved sequences and improve our understanding of how variation in one part of the sequence correlates with variation in another part. Covariance analysis improves RNA consensus structure prediction by identifying covariant positions across an MSA. This allows the prediction of conserved and functionally important base pair interactions and secondary structures. Additionally, base pair covariation can provide insight into the evolution of RNA structures over time. Understanding covariation patterns can reveal the evolutionary pressures that have shaped RNA structure and function over time. Many computational tools, such as RNAalifold uses covariance analysis to improve their secondary structure predictions [15]. Probabilistic models that flexibly describe the secondary structure and primary sequence consensus of an MSA could also be used to build a covariance model that predicts a motif in unaligned sequence [22].

2.1.9 Most Informative Sequence

The concept of the Most Informative Sequence (MIS) in RNA MSA goes beyond the consensus sequence approach. The consensus sequence represents the most frequent nucleotide at the position of the MSA, whereas the MIS highlights specific nucleotide positions that are most informative about the RNA sequence and function. Focusing on the MIS provides greater precision and specificity. The consensus sequence could mask subtle but biological important sequence variations. The MIS on the other hand can highlight specific nucleotides that may be important to the function or structure of the RNA. The International Union of Pure and Applied Chemistry (IUPAC) nucleotide code represents MIS. This code allows for the representation of ambiguities where multiple nucleotides may occupy a position in an MSA. The abbreviations are given in Table 2.1 according to [23].

Symbol	Meaning	Symbol	Meaning
R	A or G	B	C or G or T
Y	C or T	D	A or G or T
S	G or C	H	A or C or T
W	A or T	V	A or C or G
K	G or T	N	any base
M	A or C	. or -	gap

Table 2.1: Symbol Representations of nucleotides in MIS after IUPAC

2.1.10 Shannon Entropy

Shannon Entropy is a concept from information theory that can be used in MSA analysis to provide additional information. This concept was introduced by Claude Shannon and in this context entropy quantifies the degree of uncertainty or variability at a single position [24]. When applied to RNA MSA, Shannon entropy can be used to assess the variability at each nucleotide position. High entropy values show high variability, suggesting a less conserved region in the aligned sequences. Conversely, a low entropy shows a highly conserved region in the MSA. This variability in conservation could help in understanding the structural function of sequence regions. Conserved regions often correlate with structurally important domains, while variable regions may suggest adaptive evolutionary processes.

The Shannon Entropy for a specific position in an MSA is calculated using the formula:

$$H(i) = - \sum_{x \in \{A,C,G,U\}} p(x) \cdot \log_2 p(x) \tag{2.1}$$

where H(i) is the entropy at position i, p(x) is the frequency of nucleotide x at that position, and the summation is over all number of sequences. The maximum Shannon Entropy in an RNA MSA without gaps and four nucleotides is $H_{\max} = 2$ bits, and it is achieved when all four nucleotides are equally distributed at a given MSA position.

For example, in an RNA MSA, highly conserved regions (with low Shannon Entropy) may point to functionally important structural motifs that are necessary for the RNA's structural function. On the other hand, regions with high entropy could suggest areas that have undergone evolutionary changes, possibly to adapt to

different hosts [25].

2.1.11   RNA Tertiary Structure Prediction

Due to the inclusion of noncanonical base pairs that build base triples and pseudoknots, which are typically not predicted in the secondary structure, the prediction of the tertiary structure is computationally more extensive. Monte Carlo simulation is one of the computational method that can be used to predict the tertiary structure of RNA. The simulation tries to find the most stable structure that the RNA could adopt in a biological environment. Theoretically an RNA could adopt a large number of structures. The Monte Carlo simulation is used to explore the conformational space of the RNA molecule by applying random structural changes defined by the user. The simulations targets to identify conformations with the lowest free energy, which are typically the most stable and biologically relevant structures. The energy calculation in these models is based on several factors, including base pairs, stacking energies and the interaction of the RNA with its environment. The simulation attempts to transform the RNA structure into a lower energy at each step. If a structural change results in lower energy, the conformational change is more likely to be accepted than if it results in higher energy. Additionally the probability of accepting an unfavourable conformation change is higher at the beginning of the simulation than at the end. This ensures that the simulation does not become trapped in an energy minimum and helps in the exploration of the entire energy landscape of the RNA [26]. To reduce computational costs, especially for large RNA molecules, the RNA is not modelled at an all-atom level but its representation is coarse-grained. For example could one nucleotide of the RNA be represented by five points. Another advantage is the ability to run the Monte Carlo simulation in parallel. This parallelization increases computational efficiency, allowing us to sample the conformational space of an RNA molecule more effectively and converge to the optimal structure more quickly [27].

2.2 Flaviviruses

2.2.1 Overview of Flaviviruses

Flaviviruses belong to the Flaviviridae family. They can cause high-impact diseases such as yellow fever and encephalitis. The genus is divided into subgroups based on vector and host specificity. These vectors include mosquito-borne, tick-borne, insect-specific and non-known vector groups. Insect-specific flaviviruses can be further divided into classical and dual-host flaviviruses [28, 29]. Flaviviruses can infect a wide range of hosts, like humans, birds, and insects. Controlling and predicting the virus's evolution can be especially challenging due to this wide host range [30, 31]. For instance, West Nile virus infects more than 300 bird species, which act as a reservoir before spreading to humans [32]. Deforestation and urbanisation of areas can increase human exposure to wildlife reservoirs. This problem was demonstrated by the Nipah virus outbreak in Malaysia in 1999. Bat habitats overlapped with pig farms, resulting in the virus jumping from insects, to bats, to pigs and then to humans [33]. Several factors affect the global distribution of flaviviruses, including climate change. Climate change, for instance, could expand the habitat of mosquito vectors. This expansion is evident in the increasing range of Aedes mosquitoes capable of transmitting Zika or dengue viruses [34]. Flavivirus outbreaks can have severe health risks. The Zika virus epidemic 2015-2016 in Brazil for example, has caused birth defects in newborns. Yellow fever outbreaks that lead to large-scale vaccination efforts create immense economic costs for the affected countries [35, 36]. While Singapore effectively monitors and educates its public to prevent the spread of dengue virus, other places with limited resources and infrastructure may struggle to do so [37].Chronic common diseases such as obesity, hypertension and diabetes have been shown to increase the severity of flavivirus infections. This finding highlights the need for an integrated approach to flavivirus research and management, considering both viral and host factors, to mitigate the global health impact [38].

The most medically significant flaviviral groups are those transmitted by mosquitoes and ticks, as identified by [39, 40]. The burden of MBFV is clearly demonstrated by the epidemiology of diseases such as Dengue, which affects millions annually. Urbanization and global travel have catapulted the spread of these diseases across

continents [41]. Tick-borne flaviviruses like Powassan or tick-borne encephalitis virus have a specific epidemiological niche like forested areas. The expansion of human activity into these areas increases the risk of exposure [42].

The flavivirus is an enveloped virus with an icosahedral hull of a 40-65 nm diameter. The genome of the virus is linear, single-stranded and positive-sense RNA. It has a length of 10-12 kb and encodes a single open reading frame. The genome functions as a mRNA and consists of three functional regions. The two untranslated regions (UTRs) located at both the 5' and 3' ends and the coding sequence that is located in the middle of the UTRs. The 5' UTR contains a 5' terminal cap 1 structure that mimics the host's mRNA. The coding sequence encodes a large polyprotein, which is later cleaved by host and viral proteases at specific sites. The use of a single ORF allows for compact genetic storage and coordinated regulation of translation, which is necessary due to the small size of the flaviviral genome. The 3' UTR usually does not have a poly(A) tail, but it does contain variable and conserved sequences that form secondary structures that are essential for viral replication and genome cyclization [43]. Additionally, the 3' UTR contains secondary structures, such as the exoribonuclease-resistant structure (xrRNA), which can interact and disable the host immune response, thereby changing the viruses virulence [44].

2.2.2 Lifecycle Flavivirus

The infection of the host by flaviviruses is initiated via binding of the virus to specific receptors on the host cell surface. Once docked to the surface, the virus is internalised by the cell through clathrin-mediated endocytosis. This involves the virus being enveloped in a vesicle that bursts through the cell membrane into the cell. Once inside the cell, the low pH medium of the endosome triggers a fusion between the viral envelope and the endosomal membrane. The host enzyme then uncoats the nucleocapsid, resulting in the release of the viral RNA genome into the cytoplasm. Here begins the viral replication cycle. The flavivirus genome is now translated by host proteins [45]. Flavivirus RNA replication takes place in the cytoplasm of the host cell. The viral RNA is designed to mimic cellular mRNA. It does this in order to help evade the host defence mechanisms. A major part of flaviviral strategy to do this is the generation of subgenomic flavivirus RNA (sfRNA) in the host.

The sfRNA is actually the 3' untranslated region (UTR) of the viral genome that ends up of the viral genome after being digested by the host's cell. XRN1 is an exoribonuclease enzyme and part of the host RNA decay mechanism, which is inhibited by xrRNA in the 3' UTR. This leads to the accumulation of sfRNA. The sfRNA interferes with the host's interferon signaling pathway, increasing the virulence of flaviviruses [46]. Newly assembled viruses are build in the host's endoplasmic reticulum. This involves the assembly of viral components such as the capsid, envelope and newly polymerized RNA genome. After the assembly, the virus goes into the lumen of the endoplasmatic reticulum, resulting in the formation of mature virions that are released from the host cell by cell lysis [45].

2.2.3 Exoribonuclease-resistant Structure (xrRNA) in Flaviviruses

The 3' untranslated region (3'UTR) of flaviviruses impacts the viruses replication, stability and interaction with the host [47]. its primary function is to interact with viral and host proteins to enhance viral replication and escape the hosts immune response. The 3'UTR contains RNA motifs such as the dumbbell structure, the xrRNA and the stem-loop IV structure [48, 49]. These elements are conserved across species and are of importance for replication, virulence and adaptation. However, there is also variation between species in the elements present in the 3'UTR, which has implications for virulence and host specificity. This variation allows flaviviruses to adapt to different hosts and environmental conditions, contributing to their widespread and diverse nature. The xrRNA structure is important in the pathogenic cycle of flaviviruses. It protects the viral RNA from complete degradation by the host XRN1. The structural features of xrRNA, such as a ring-like knot around the RNA strand, create a mechanical blockage that prevents XRN1 from proceeding with degradation. Therefore, this resistance is not just a passive barrier, but actively helps to evade the host immune response. xrRNA prevents the degradation of viral RNA, thereby avoiding the triggering of immune sensors that normally recognise foreign RNA. Incomplete degradation of viral RNA leads to the formation of subgenomic flaviviral RNA (sfRNA), as symbolized in Figure 2.3.

So by protecting specific regions of viral RNA from complete degradation, xrRNA
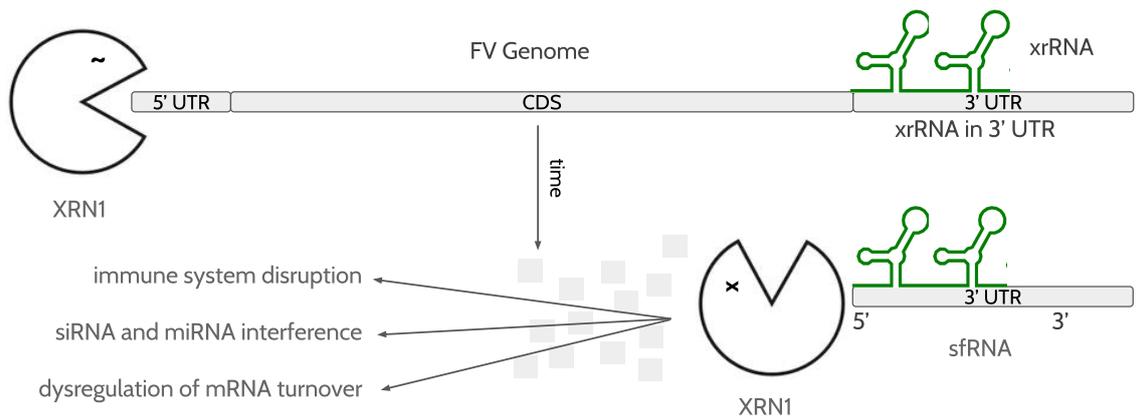
Figure 2.3: **Symbolic representation of subgenomic flaviviral RNA production by XRN1 and xrRNA** XRN1 attempts to degrade the flaviviral genome but is halted at the start of the 3' UTR by the xrRNA structure.. The rest of the flaviviral RNA is called subgenomic flaviviral RNA (sfRNA) and can lead to various negative side effects in the host, such as disruption of the immune system, siRNA and miRNA interference, and dysregulation of mRNA turnover.

promotes or better enables the formation of sfRNA. The sfRNA is not only a by-product, but also enhances viral pathogenicity and immune evasion by modulating the host immune response and potentially altering virus-host interactions [47, 50].

This shows us that the 3'UTR and xrRNA of flaviviruses are of research interest, as they have the potential to become targets for antiviral strategies. Understanding their function and structure could lead to the development of novel therapeutic approaches aimed at disrupting their vital role in the viral life cycle.

2.3 Exoribonuclease-resistant RNA (xrRNA) Structures

2.3.1 Structure of xrRNA

The xrRNA has a multi-loop with a three-way junction and some unpaired nucleotides in the multi-loop. The three-way junction is surrounded by conserved nucleotides and is useful for the overall stability and function of the xrRNA. In addition, some helices in xrRNA are stacked coaxially, which contributes to the sta-

bility of the structure. The exoribonuclease-resistant structure of the RNA molecule adopts a ring-like conformation, with the 5'-end of xrRNA passing through the ring. The ring consists of 15-16 nucleotides and this unique structure plays a major role for its resistance to degradation by XRN1. The structure relies on conserved nucleotides and non-canonical bindings, such as base triples, which help to maintain the unique topology. The xrRNA structure involves the coordination of a magnesium ion, a common feature of RNA structures, which stabilises the phosphate backbone [51]. There are two types of pseudoknot interactions that appear to be important to the structure of xrRNA. The first is a short pseudoknot that connects the unpaired region in the multiloop to the start of the first stem. This helps to form the ring-like structure at the 5' end. The second is a long pseudoknot interaction that connects the hairpin region of the third stem to the end of the xrRNA. This locks the structure in place. The structural configurations of xrRNAs, such as the double pseudoknot and base triplets, show a conserved topology in one form or another across all flaviviral groups [52]. In many flaviviruses, xrRNA structures are often found in tandem in the 3' UTR. Typically there are two structures, but there can be up to five [48]. The interdependent relationship between xrRNAs suggests that the structural integrity of one may affect the function of another. A study examining the structure of a wide range of flaviviruses has demonstrated the universality of the ring-like structure in xrRNAs as a fundamental mechanism for resisting degradation by XRN1. This ring-like structure in xrRNA, despite sequence variations, represents a conserved evolutionary strategy to block enzymatic digestion. The presence of the ring-like feature in different flavivirus groups highlights the importance of its structure in maintaining viral RNA stability and function. This underlines its importance in the viral life cycle and host interactions [53].

2.3.2 Structural Diversities and Uniform Function of xrRNA in Flaviviruses

A comparison of genomics study showed structural diversity and evolutionary conservation of xrRNAs across different flavivirus groups [48].

1. MBFVs xrRNA exhibit the structural features as described above.

2. In 1997, the secondary structure of the 3'UTR of TBFVs was predicted [9]. It was later discovered that it also contains the typical xrRNA structure of TBFV. This structure displays distinct attributes from MBFVs, although it is less characterized in terms of pseudoknots and base triples due to the lack of an instrumentally analysed 3D structure. The first stem towards the multi-loop appears to be longer, and only one pseudoknot extend to the end of the structure, with a longer multi-loop [52].

3. NKVs xrRNA share the common characteristics of a three-way junction with both MBFVs and TBFVs. Non-canonical basepairs like base triples and pseudoknots were not analysed. The structure seems more familiar to TBFVs than to MBFVs xrRNA [51].

4. ISFVs xrRNA needs to be observed in two categories. The dualhost ISFVs have a nearly identical xrRNA structure as the MBFVs. And the classic ISFVs xrRNA also look similar to the MBFV but they appear to have a longer pseudoknot1 from the beginning of the sequence to the multi-loop.

Despite structural variation, xrRNAs play a consistent role across all flavivirus groups. This includes primarily resistance to XRN1, which leads to sfRNA production and contributes to viral evolution and host adaptation. A recent study revealed a novel class of xrRNA elements in Rana tamanavirus (RaTV) identified in European common frogs, expanding our understanding of flavivirus diversity and the evolutionary adaptability of xrRNA [54].

2.3.3 Mechanism of xrRNA

The xrRNA has a unique mechanism of resistance to degradation by XRN1, which is attributed to the topology of its ring-like tertiary structure. The ring-like structure forms a knot around the 5' end of the RNA molecule, creating a mechanical barrier that prevents the enzyme XRN1 from degrading the RNA. XRN1 faces a mechanical unwinding problem as it tries to pull the 5' end of the RNA through the ring, which its normal thermal unwinding abilities cannot overcome. The xrRNA appears to exhibit mechanical anisotropy [55]. The xrRNA is resistant to degradation by XRN1 when pulled from the 5' end, but can be replicated by viral

RNA-dependent RNA polymerase (RpRP), which works from the 3' to the 5' end of the RNA molecule. This mechanical anisotropy of xrRNA is an evolutionary adaptation of RNA viruses that balances the need to protect their RNA from degradation with the requirement for efficient replication. Recent studies have shown that the stability of xrRNA against XRN1 is positively impacted by the prolonged lifetime of its ground state conformation, which is orders of magnitude longer than that of canonical RNA structures, making xrRNA more resistant to degradation. Mutations in the pk2 region of MBFV that reduce this conformational lifetime have been shown to impair xrRNA resistance and the ability of the virus to replicate. [56]. Further studies have shown that a longer PK2 region in MBFV xrRNA increases XRN1 resistance and reduces the dependence on magnesium ions for structural stability. It also enables the xrRNA to withstand mutations in these regions that would otherwise destabilise the xrRNA structure. [57]. The resistance of xrRNA to XRN1 depends on a structural motif common to all flaviviruses. This motif involves a ring-like architecture that surrounds the 5' end of the RNA, creating a strong barrier against XRN1 [53].

2.4   Exoribonuclease XRN1

2.4.1   Role of XRN1

Exoribonuclease XRN1 is a member of the $5' \to 3'$ exoribonuclease family. XRN1 plays a major role in the post-transcriptional modification and regulation of RNA molecules within a cell. The enzyme degrades one nucleotide at a time from the 5' end to the 3' end. XRN1 also regulates the mRNA in a cell. It degrades unstable mRNA molecules that are no longer needed, improperly made or damaged. XRN1 plays a vital role in the cells RNA metabolism and in defence against RNA viruses [58]. XRN1 degrades viral RNA. This limits the viral replication and spread in the host's cell. This antiviral effect has been observed for RNA viruses like the negative-sense RNA viruses such as influenza [59]. XRN1 degrades parts of the viral mRNAs reducing the ability of the virus to synthesise proteins necessary for their life cycle. Or as in positive-sense RNA viruses such as picornaviruses, XRN1 targets the whole viral genome directly preventing its replication [60]. However, it can also have a

more complex interaction, such as with flaviviruses, which generate sfRNA and help evade the host immune response [61].

2.4.2   Mechanism of XRN1

XRN1 binds to the monophosphate of the first nucleotide at the 5' end, which is mediated by aromatic side chains in the enzyme that stack with the bases of the first three RNA nucleotides. The catalytic site of XRN1 is activated by a metal ion like magnesium (Mg^{2+}). The catalytic mechanism of XRN1 involves the activation of a water molecule by the Mg^{2+}. The activated water molecule performs a nucleophilic attack on the phosphodiester bond between the nucleotide bound to the enzyme and the next nucleotide in the RNA. This attack cleaves the 5'-nucleotide from the rest of the RNA. The XRN1 releases the nucleotide and leaves the RNA molecule with a new nucleotide with a 5'-monophosphate end. This monophosphate is important for the XRN1 to bind the next nucleotide. After cleavage XRN1 moves the RNA one step further to its activated site, which positions the next nucleotide for cleavage. This is important to note, because it shows us that XRN1 requires a minimum of three unpaired nucleotides to initiate RNA degradation. XRN1 can unwind RNA structures, such as stem loops, to pull the RNA molecule into position. This unwinding happens through the RNA translocation mechanism and steric barriers within the enzyme that exclude double-stranded regions and allow only single-stranded substrate processing. XRN1 is called a processive enzyme. This means that the XRN1 normally degrades an RNA molecule completely without releasing partially degraded RNA fragments [58].

2.5   Computational Tools for RNA Structure Analysis

2.5.1   ViennaRNA Package

The Vienna package [11] is a suite of tools and algorithms designed for the analysis, prediction, and comparison of RNA secondary structures. Key tools of the ViennaRNA Package include:

1. RNA secondary structure prediction can be achieved using RNAfold and RNAalifold. These tools can predict the secondary structure of single or multiple RNA sequences based on minimum free energy.

2. RNA-RNA Interaction Prediction involves the use of RNAcofold to predict the structures of RNA dimers, which is useful for understanding RNA-RNA interactions. Additionally, RNAduplex can be used to find the optimal interaction between two RNA molecules, making it ideal for studying RNA hybridization.

3. RNA design can be achieved using RNAinverse, which solves the inverse folding problem by designing RNA sequences that fold into a given structure.

4. RNAplot and RNAdistance are visualization tools used for graphical representation of RNA secondary structures and measuring structural similarities between RNA molecules, aiding in comparative analysis

2.5.2 RNAfold

RNAfold is a tool from the ViennaRNA package [11]. It predicts the secondary structure of a single RNA molecule by calculating the minimum free energy structure. The minimum free energy is calculated using a set of thermodynamic parameters including energies for different loop types and base pair stacking. A dynamic programming algorithm is used to explore the space of secondary structure conformations that the RNA sequence can adopt. In addition to predicting the minimum free energy (MFE) structure, RNAfold can also calculate the partition function for the ensemble of all possible secondary structures. This allows base pairing probabilities to be estimated, providing insight into the flexibility and dynamics of RNA structures.

2.5.3 RNAalifold

RNAalifold is a tool used to predict the secondary structure of RNA. It extends the typical RNA secondary structure prediction method used by RNAfold to MSAs. RNAalifold combines thermodynamic energy minimization with evolutionary conservation. RNAalifold uses evolutionary information into the prediction process, by

analysing these data points from a MSA. Like the score for the covariation of base pairs in the MSA [15].

2.5.4 VARNA

VARNA [62], the Visualization Applet for RNA, is a interactive tool for drawing and editing of RNA secondary structure. With Python API integration, it is a versatile tool designed for the automation of visualisation and annotation of RNA secondary structures. Using VARNA made it simple to streamline the analysis of RNA structures. This was achieved by incorporating a pipeline that predicts secondary structure, calculates positional entropy, and performs covariance analysis on multiple sequences in series. VARNA's support for visualizing non-canonical base pairs and pseudoknots aided in understanding complex xrRNA structures.

2.5.5 MAFFT

Multiple Alignment using Fast Fourier Transform (MAFFT) is a tool for performing MSAs. Mafft uses Fast Fourier Transformation to identify homologous regions in sequences, which reduces computational time while maintaining high accuracy in alignments. The core of MAFFT's approach lies in its scoring system and a sophisticated gap penalty mechanism. Starting with a progressive alignment based on a guide tree, MAFFT refines this initial alignment through iterative techniques, optimizing the overall structure and accuracy of the MSA [63].

2.5.6 simRNA

SimRNA is a computational method for simulating RNA folding and predicting its 3D structure [27]. SimRNA uses a coarse-grained model to simplify the representation of RNA molecules, making it easier to simulate complex interactions. The model represents the RNA backbone with two pseudo-atoms and the bases with three pseudo-atoms, reducing the complexity of the RNA structure from 20-34 atoms per residue to just five key atomic positions. This preserves the essential properties of the RNA strand. SimRNA uses the Metropolis algorithm to scan the conformational space of RNA molecules. This algorithm is a Monte Carlo method

in which defined random changes are made to the RNA structure. These changes are accepted or rejected based on their effect on the energy of the RNA molecule. SimRNA can also use additional information in the form of constraints. These constraints can come from computational predictions and take the form of secondary structure information. They help to guide the folding process and can be useful in accurately modelling 3D structures.

2.5.7 QRNAS

QRNAS is a tool for refining nucleic acid structures. It is effective in energy minimisation following 3D simulations such as Monte Carlo. The minimum free energy structure obtained from Monte Carlo simulations can sometimes be physically impossible or broken. It is therefore important to perform energy minimisation to refine the structure towards a biologically plausible and physically coherent state. QRNAS improves the local geometry of the structure, increasing its quality and accuracy without affecting global measures such as RMSD. This is important when performing a comparison of structures after the resulting Monte Carlo simulation [64].

2.5.8 PyMOL

PyMOL is a tool that allows users to visualise 3D molecular structures. We used PyMOL to visualize 3D RNA structure and to calculate the Root Mean Square Deviation (RMSD) between two RNA molecules. We could also align two RNA molecules, which was helpful for a comparison and visualisation of structural similarities or differences [65].

2.5.9 Infernal

Infernal (INFERence of RNA ALignment) is a tool that is used to analyze RNA sequence data [66]. Infernal is a tool that identifies homologous RNA sequences in RNA databases. It does this by building consensus RNA secondary structure profiles known as covariance models (CMs) and searching these RNA databases with them. Infernal uses stochastic context-free grammars (SCFGs) to integrate both sequence

and secondary structure information into a statistical framework. With these CMs, it can quantitatively score and rank potential homologous RNAs in a database. In addition, Infernal is useful for annotating RNA structures in a given genome.

2.5.10   CMCompare

CMcompare is a tool designed to measure the discriminatory power of RNA covariance models. It allows the comparison of different covariance models and the analysis of their similarities. The Link Score is a key metric in CMCompare as it quantifies the overlap between covariance models. This allows comparison of how well different covariance models capture the variability and conservation of the RNA molecules they are built from [67].

2.5.11   Bedtools

BEDtools is a software package for the analysis of genomic features. The suite is designed to efficiently compare and explore genomic datasets. Each tool performs a simple operation, ensuring ease of use and accuracy [68]. These tools include intersection, windowing, finding the closest non-overlapping intervals and applying functions to columns for overlapping intervals.

2.5.12   NCBI

The National Center for Biotechnology Information (NCBI) provides online resources for searching and analysing biological data. These resources include GenBank (less curated) for nucleotide sequences, PubMed for biomedical literature, and RefSeq for curated nucleotide sequences.

Chapter 3

Materials and Methods

3.1 Methodological Overview

The Figure 3.1 shows the workflow used to predict the secondary and tertiary structure of xrRNA in flaviviruses. The process begins with the acquisition of flaviviral sequences, which are then filtered according to specific criteria to ensure usable data. A homology search is then performed to identify potential xrRNA sequences, followed by a further screening step to ensure data fit. Finally, an MSA is generated from these sequences. The first step is to predict the structure of a randomly picked sequence from the different viruses. This sequence can then be used as a constraint when predicting the consensus structure. After alignment, a secondary prediction is made on the consensus structure. At this point, the workflow branches out and a covariance model is built from this consensus structure. This model can be used to search for other xrRNA in the original data set. However, the consensus structure is represented visually and includes data such as covariance and positional entropy. A 3D Monte Carlo simulation is then performed to predict a tertiary structure. This structure is then compared with a target structure. This approach allows the computational analysis of numerous flavivirus sequence inter- and intra- viruses for different xrRNA structures.
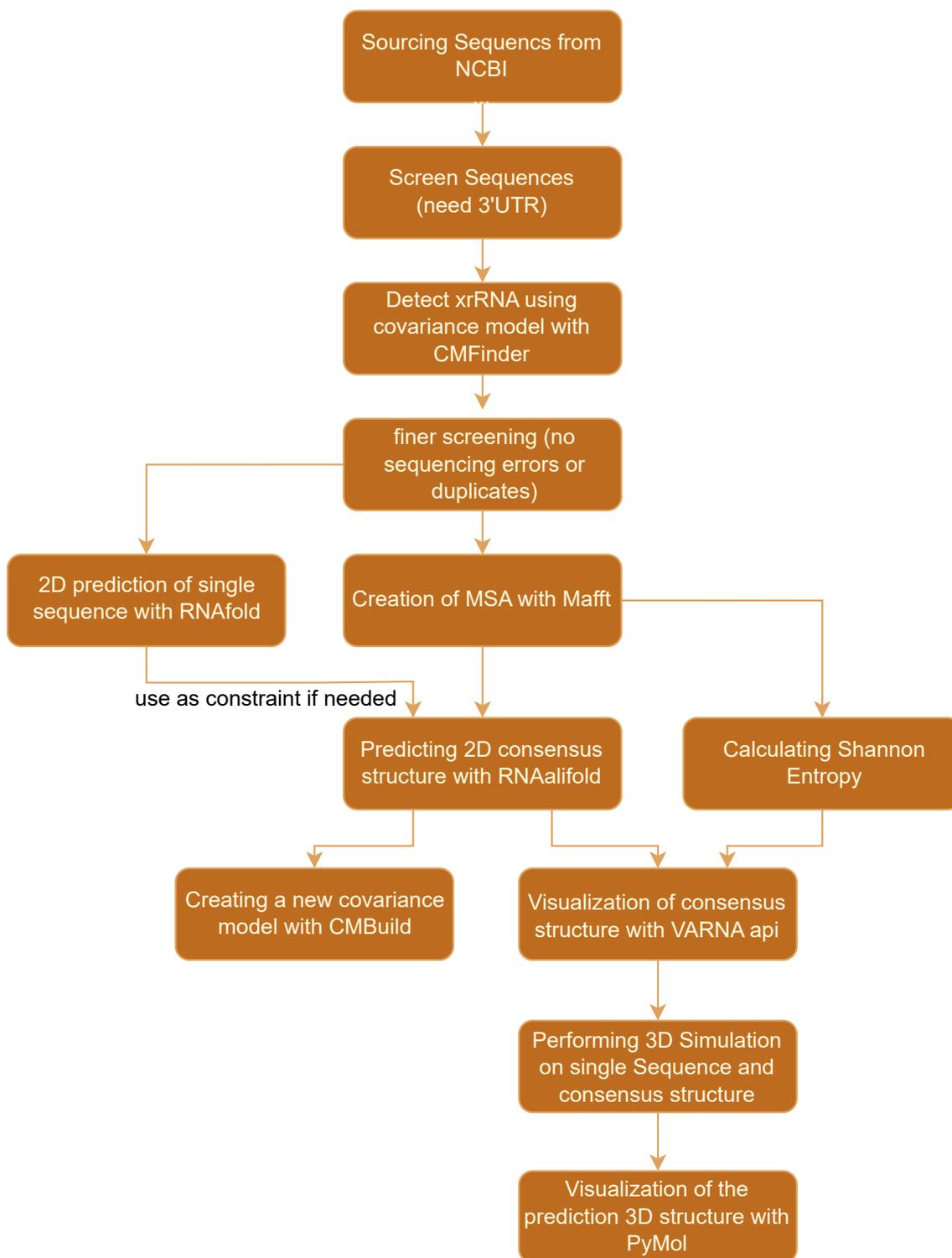
Figure 3.1: **Workflow chart** illustrating the methodological approach on the computational analysis of finding the 2D and 3D structure of xrRNA in different falviviruses.

3.2 Sequences from NCBI

The flavivirus sequences used in this study were obtained from the NCBI RefSeq and GenBank databases as of 11 November 2023. We downloaded viral genomes classified under taxonomic ID 11051, which corresponds to the genus Flavivirus. Our selection was further refined to include genomes belonging to MBFV, TBFV, NKV and ISFV. The downloaded data set consisted of 13606 MBFV, 872 TBFV, 203 ISFV and 16 NKV isolate 3'UTR sequences. The 3' UTR sequences of the viruses listed in Table 3.1 were selected for further analysis based on their medical significance or the number of isolates available for each virus. The 3' UTR sequences of the viruses listed in Table 3.1 were selected for further analysis based on their medical significance or the number of unique sequences available for each virus.

3.3 Structural Homology Search with Covariance Models

To identify structural homologs of xrRNAs in flaviviruses, we used the xrRNA covariance models defined in [49]. The search for homologous sequences was performed using INFERNAL 1.14, a package specifically designed for RNA sequence analysis [66]. This method effectively identifies conserved RNA structures by using covariance models that use both sequence and secondary structure information of the RNA. CMFinder [22], a subtool of the INFERNAL package, used the xrRNA covariance model to scan relevant 3'UTR sequences obtained from the Genbank search. The hits obtained were screened to ensure high specificity in the selection of potential homologues. The screening process involved setting an E-value cut-off of 0.00001. It is important to note that some viruses have several xrRNAs in series. Therefore, the resulting hits were further sorted based on the number of xrRNA in a virus and grouped into xrRNA1, xrRNA2 and xrRNA3, 4 and 5 based on their order in the sequence. After filtration, BEDTools 2.31.2 [68] was used to expand the sequences. The pseudoknot2 region was added to the end of the sequence (approximately 15 nucleotides) and the pseudoknot1 region was included for the MBFV, NKV and ISFV subgroups (approximately 5 bases). A screening step was then per-

Table 3.1: Alphabetically sorted list of flaviviruses selected for analysis, categorized by groups. List with full names can be found in appendix

MBFV	TBFV	ISFV	NKV
AROAV	ALKV	AEFV	MMLV
BAGV	DTV	BinJV	MODV
BANV	GGV	BJV	RBV
DENV1	KAMV	CFAV	
DENV2	KFDV	DONV	
DENV3	KSIV	EPEV	
DENV4	LGTV	GUAV	
JEV	LIV	KRV	
KOKV	MPFV		
KOUV	NEGV		
KUNV	OHFV		
KVEV	POWV		
NMV	SGEV		
SLEV	SREV		
TMUV	TBEV		
UGSV	TYUV		
USUV	XiFV		
WESSV			
WNV			
YFV			
ZIKV			

formed to eliminate sequences with sequencing errors (sequences containing 'N') and duplicate entries.

3.4   Creating Multiple Sequence Alignment (MSA)

Following the structural homology search, I generated a MSA using the sequences obtained from CMFinder and the subsequent screening steps. The chosen sequences, which show potential xrRNA homologs across various viruses, were aligned using MAFFT version 7.520 [63]. The purpose of this alignment process was to compare the sequences, highlighting both conserved regions and differences that could

suggest structural or functional similarities and variations. The alignment was also checked for sequence outliers to ensure the specificity of the sequences to the xrRNA group of interest. This was a significant step in focusing on xrRNA1 sequences and excluding inadvertent inclusions of xrRNA2, xrRNA3, and so on, due to their structural similarity. This step is important in the workflow of my study because a false MSA could compromise the accuracy of my findings and their interpretations. To optimize the process of MSA, we used the –auto flag with MAFFT. This allowed for the automatic selection of the most suitable alignment strategy based on the data size.

3.5   xrRNA Shannon Entropy

To evaluate the variability of xrRNA sequences in various Flavivirus groups, we calculated the Shannon entropy for each position in the MSA [24]. The process involved using the Biopython AlignIO module to read the MSA file. Then, each column of the alignment was iterated to calculate the frequency of each nucleotide. The Shannon entropy was computed for each column to quantify the sequence conservation. Higher values show greater variability, while lower values show lower variability. This approach yielded a quantitative measure of conservation and variability across the aligned xrRNA sequences. This aided in identifying conserved regions that may be essential for xrRNA's function.

3.6   xrRNA Secondary Structure Prediction

To analyse the secondary structure of xrRNA in Flaviviruses, we folded the sequences identified through homology searches into their secondary structures using RNAfold and RNAalifold, tools from the ViennaRNA package. However, these tools have limitations, and we were unable to predict a secondary structure with pseudoknots, which we believe to be a critical aspect of the xrRNA structure. To identify potential pseudoknot regions, the folding constraints were modified to allow the formation of duplexes only within unpaired regions of the initially predicted secondary structure, where pseudoknots might build. If a compatible duplex resem-

bling a pseudoknot was identified, its secondary structure was integrated with the primary secondary structure prediction to effectively incorporate the pseudoknots. The VARNA python API was used to visualize the secondary structures, which provides a graphical representation of the xrRNA folds.

3.7 xrRNA 2D Consensus Structure

To predict the secondary structure of xrRNA across the four flaviviral families, we analysed the consensus structure of various viruses (as described in 3.1) within each family. For instance, we predicted the secondary consensus structure of xrRNA in TBFV by first generating a consensus structure prediction for each virus in TBFV. To achieve this, we used RNAalifold and Andronescu parameters with each MSA. However, the initial consensus predictions did not always align with the expected xrRNA structure. To address the discrepancies in the initial predictions, a deeper analysis was conducted on a single sequence from each virus using RNAfold and its subopt() function. This allowed for the identification of suboptimal structures, which were overlooked by RNAalifold, and led to the discovery of candidate structures. After identifying a suitable xrRNA structure within individual sequences in the viruses, parts of this structure were used as constraints in a subsequent RNAalifold prediction to guide the process of finding the correct secondary structure. The final step in refining the 2D consensus structure prediction involved identifying potential pseudoknot formations. The consensus structure is now established, and the unpaired regions are identified. By modifying the folding constraint to allow duplex formation only in the unpaired region, we predicted the possible location of pseudoknots. These were then integrated with the previously predicted consensus structure to form the final structure. The consensus structure of all viruses within a family ensures that they consist of the same structural elements, allowing for the prediction of the consensus structure over all TBFV. Outliers detected between the viruses within the flaviviral groups are excluded from the consensus prediction. For each virus of the TBFV, one sequence is selected, and an MSA is created using MAFFT. A consensus structure prediction is performed using RNAalifold with constraints outlining the end or beginning of structural elements. For instance, a three-way junction is present in all xrRNA flaviviral groups.

3.8 xrRNA Structure Statistics

In this study I conducted a detailed analysis of the structural elements of the xrRNA structure for each subgroup after predicting the consensus structure for each selected virus within the four distinct flavivirus groups. The study counted the minimal and maximal lengths of each structural element in each consensus structure, providing insight into the structural diversity and constraints characteristic of xrRNA across different flaviviruses. The study visualized the four distinct flavivirus xrRNA using VARNA as a coarse-grained symbolic representation.

A correlation analysis was conducted to evaluate the dependence of the length of various elemental structures of the xrRNA on each other. The analysis was performed using the pandas 1.5.3 library in a Python 3.11.6 environment.

The Pearson correlation coefficient was calculated using the following equation based on the length of pairwise xrRNA1 and xrRNA2 structures in each subgroup of flaviviruses, with two from each virus:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

where r_{xy} is the Pearson correlation coefficient between the number of nucleotides in the structural elements x and y. x_i and y_i are the individual values of variables x and y. $\bar{x}$ and $\bar{y}$ are the mean values of x and y. n is the number of structures analyzed in this statistic. The correlation coefficients will be interpreted to understand the strength and direction of relationships between structural elements.

3.9 xrRNA 3D Structure Analysis

Monte Carlo simulations were performed using SimRNA 3.20 to predict and analyse the 3D structure of xrRNA [27]. The simulation was set to run 20 million iterations, saving the state of the system every 20 thousand iterations. The initial and final temperatures were set to 1.65 and 0.60, respectively, to create a gradual cooling process. This allows the system to more easily escape local minima and explore the entire conformational space. To increase the sampling efficiency, we used the Replica Exchange Monte Carlo (REMC) method by activating it with the

'E-16' switch. This allowed 16 simulations to be run at different temperatures. The simulations target is to exchange configurations, allowing for more efficient bypassing of energy barriers and better sampling of conformational space. The sequence information and secondary structure, including pseudoknot1 and pseudoknot2 in parenthesis notation, were entered as separate input files. The input included structural constraints to guide the folding process and increase the likelihood of accepting motions consistent with the predefined secondary structure. This approach ensures that the simulation not only explores conformational space, but also stays close to the biologically relevant structure provided in the input files. The simulations were run on a high performance computing system using SLURM, a workload manager, to efficiently manage resources. The following SLURM parameters were used:

- Job Array: –array=1-25 established a job array with 25 tasks, enabling the parallel execution of simRNA. The job array feature of SLURM is useful for studies like Monte Carlo Simulations, where multiple configurations are evaluated in parallel to explore the conformation space of xrRNA.

- Memory Allocation: –mem=15G was specified to allocate 15 gigabytes of RAM to each job.

- CPUs per Task: With –cpus-per-task=16, each task within the job was assigned 16 CPUs.

After conducting 25 parallel simulations, we clustered the trajectory data. The clustering process was performed on the top 1% of frames based on their energy, with an RMSD threshold of 5 angstroms. We then conducted a numerical analysis on the number of trajectories in each cluster to find the distribution and ratio of the different conformations. Prior to visualising the 3D structure with Pymol 2.5.0 (Delano, 2002), an energy minimisation step was performed using the QRNA software to optimise the conformations to the lowest possible energy state. This helps to eliminate any steric clashes or unfavourable interactions, leading to increased structural accuracy and biological relevance.

Chapter 4

Results

In this Thesis I tried to analyse the xrRNA structures across different groups of flaviviruses. Bioinformatical tools such as RNAalifold, simRNA and Infernal were used to gain a better understanding of the complexities and variabilities in the xrRNA structures of MBFV, TBFV, ISFV and NKV flaviviruses. This chapter presents the results of the computational analysis, starting with the consensus structure predictions for xrRNA1 and xrRNA2 in different viruses within each flavivirus group. I tried to identify conserved structural elements in xrRNA that may be important in xrRNAs resistant to the host XRN1. This was achieved by analysing base pairing patterns, covariation, and Shannon entropy. Statistical analysis, such as Welch's t-test and Pearson correlation coefficients, were also used to provide quantitative insight into the structural differences between xrRNA1 and xrRNA2 elements within MBFV and TBFV.

I present the results in four main sections, each dedicated to a distinct group of flaviviruses. In each section, the 2D consensus structures resulting from the analysis are described, with findings related to structural conservation and variability highlighted. The comparison of xrRNA1 and xrRNA2 in the MBFV and TBFV groups is presented, with an explanation of how the length of the structural elements correlates with each other. Due to limited data, ISFV and NKV have been excluded. Finally, the predicted 3D structure of the xrRNA in each group is shown.

4.1 MBFV xrRNA Structure Analysis

The study of xrRNA structures in MBFV provides a fresh perspective on the mechanisms of viral RNA that aid in resilience against host cellular defenses. MBFV, which have an impact on global public health, include viruses such as Dengue, Zika, and West Nile viruses. The computational analysis presented here focuses on the structural elements and sequence conservation/variability across xrRNA1 and xrRNA2 within MBFV. The objective is to understand the mechanism that leads to XRN1 resistance. Before going into details of xrRNA1 and xrRNA2 and their comparison, it is essential to ensure a clear understanding of the components of the xrRNA structure. In Figure 4.1 the structure of MBFV's xrRNA is characterised as a three-way junction consisting of the α, β and γ strands. In addition, the β and γ stems form a hairpin loop. There are also two pseudoknots in xrRNA, pk1 and pk2. Pk1 is short and consists of two base pairs extending from the start of the structure to the multiloop section. On the other hand, pk2 extends from the hairpin γ to the end of the xrRNA and contains more base pairs compared to pk1. The median consensus structure for xrRNA1 was calculated using 21 viruses, while for xrRNA2 it was based on 16 viruses.

4.1.1 MBFV xrRNA1

The analysis of xrRNA1 across multiple MBFV viruses shows a landscape of structural conservation interspersed with regions of notable variability. Consensus structure predictions were made using RNAalifold and are exemplary shown in Figure 4.2 for four viruses, including Aroa virus (AROAV), Bagaza virus (BAGV), Dengue virus (DENV3), and Zika virus (ZIKV).

These four structures demonstrate a common structural motif, namely a three-way junction. The consensus structures comprise a highly conserved stem α, which is likely significant for the structural integrity and function of the xrRNA. This is followed by a multiloop of grade 3, which serves as a connecting region between the other structural elements, namely stem β with an adjacent hairpin loop β, and again a highly conserved stem γ leading to hairpin γ. A conserved cytosine is present within the multiloop, between stems β and γ. The conservation of this cytosine
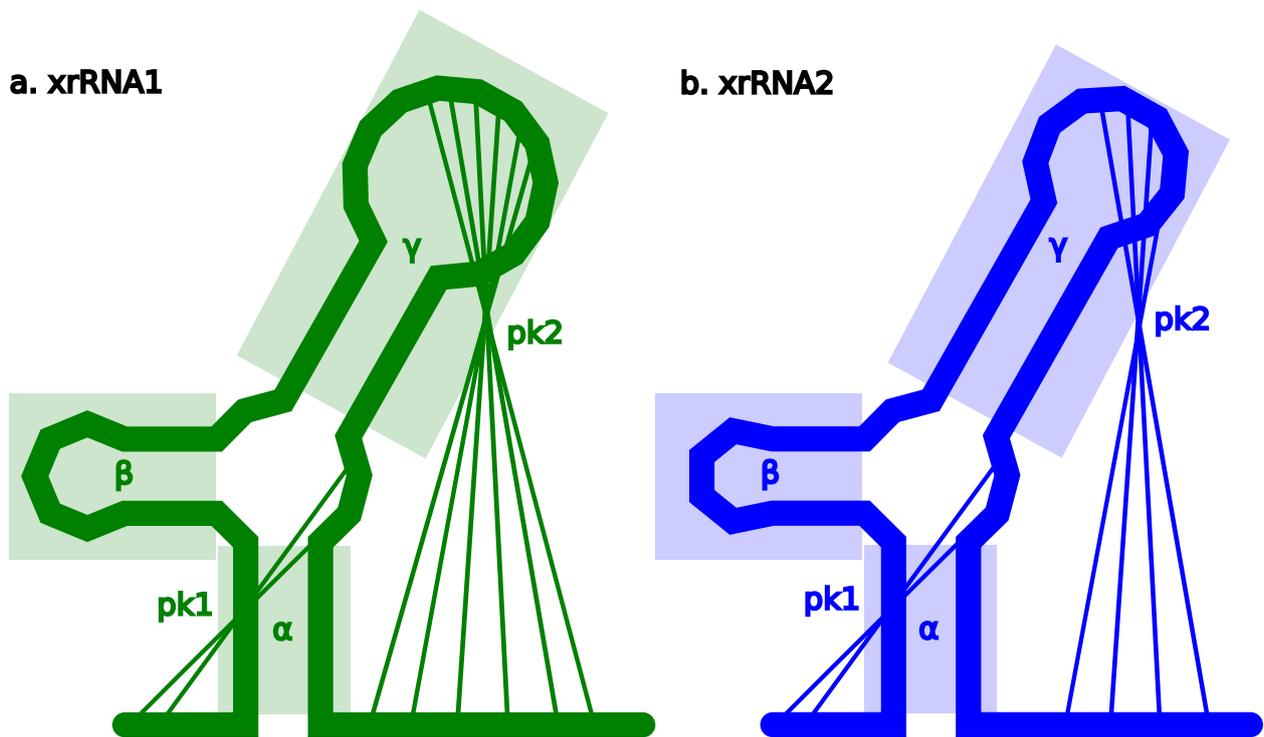
36

Figure 4.1: **Median Consensus Structure of MBFV a. xrRNA1 and b. xrRNA2**. A highlevel overview of the architecture of MBFV xrRNA1 and xrRNA2, highlighting a three-way junction formed by stems α, β, and γ, complemented by two pseudoknots, pk1 and pk2. The length of the structural elements was calculated by taking the median length from each MBFV xrRNA consensus structure.

suggests a potentially vital role, possibly in the folding and stability of the xrRNA structure or in the interaction with host proteins. Two pseudoknots are identified within these structures. Pseudoknot 1 is short, comprising two base pairs with evidence of covariation, extending from the beginning of the sequence to the unpaired nucleotides in the multiloop. Pseudoknot 2 is longer, spanning from the hairpin γ across a stretch of unpaired nucleotides at the unpaired end of the structure, which also exhibits covariation. The figure's color-coding shows the level of conservation and covariation across the viruses. Red represents conserved base pairs, orange represents covariation with two possible base pairs, and green represents covariation with three possible base pairs. This colour-coding system visually represents the sequence variability and structural conservation within the xrRNA1 across flavivirus viruses. The nucleotides are coloured based on the positional Shannon entropy
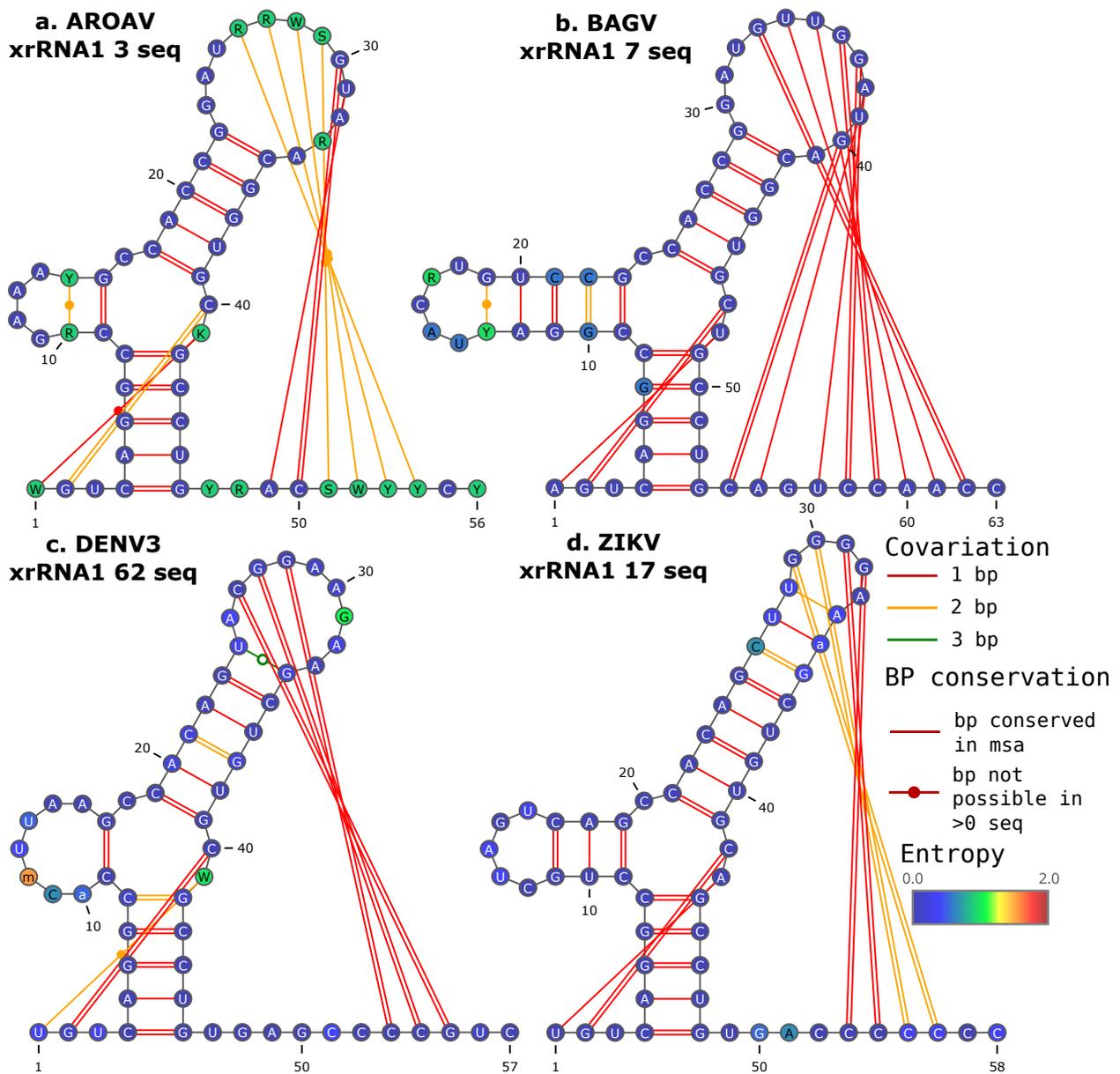
Figure 4.2: **Consensus Structure Predictions of xrRNA1 across four MBFV viruses.** The figure displays the consensus structure predictions for xrRNA1 using RNAalifold on four selected viruses in MBFV: Aroa virus (AROAV) with 3 sequences, Bagaza virus (BAGV) with 7 sequences, Dengue virus (DENV3) with 62 sequences, and Zika virus (ZIKV) with 17 sequences. The base pairing color shows the covariation in the structure: red for conserved base pairs, orange for two possible bps, and green for three possible bp. The colours of the nucleotides are picked after the positional Shannon entropy within the MSA of each virus, suggesting nucleotide variability. Dark blue represents no variability (entropy = 0), green represents moderate variability (entropy = 1), and red represents high variability (entropy = 2). The sequences displayed are the most informative sequence, from each MSA.

38

within the MSA of each virus, suggesting variability and sequence conservation. Lower entropy areas (closer to the blue end of the spectrum) suggest conserved regions, while higher entropy areas (closer to the red end of the spectrum) suggest variability.

In Figure 4.3 panel A. we see the consensus structure of the 21 viruses. This structure consists again of the typical a three-way junction with stems—α, β, and γ and two hairpin loops (from β and γ) and a pair of pseudoknots (short and long). The analysis benefits from the inclusion of diverse virus sequences, providing a stronger signal for covariation and positional entropy, key indicators of structural conservation and variability than the consensus structures of each virus as seen in Figure 4.2.

The consensus structure shows low entropy in the first pseudoknot and along stem α, along with some covariation. This suggests a highly conserved sequence and structure across different viruses. The nucleotides and basepairs within stem β exhibit high entropy and some covariation, suggesting a less conserved structure. This is supported by the base pairs that do not consistently form across all sequences, denoted by the central dot within the base pairs. The β hairpin contains many gaps, which were introduced by a single virus, YFV. Stem γ has moderate entropy but strong covariation, suggesting an intermediate level of structural conservation. Pseudoknot 2 does not reliably form across the consensus structure, suggesting that its formation may not be a stable feature in these RNA sequences. The high entropy observed in the region corresponding to pk2 suggests a high degree of variability. The inconsistency in the formation of pk2 could be attributed to either evolutionary divergence among the viruses. Further investigation into the covariation patterns may reveal the evolutionary pressures and functional demands placed upon this region of the xrRNA1.

In Figure 4.3 panel B. we see the MSA of the 21 virus sequences that were used to create the consensu structure. The alignment displays the consensus structure in dot-bracket notation, which simplifies the complex secondary structure. It highlights regions of high conservation, within stem α and pk1 (region nucleotide number 1-8 and 55-63), coloured in red and orange, suggesting conserved base pairs or covariation with two possible base pairs. On the other hand, the alignment shows regions with gaps, especially within hairpin β (nucleotide number 12-28) and pk2
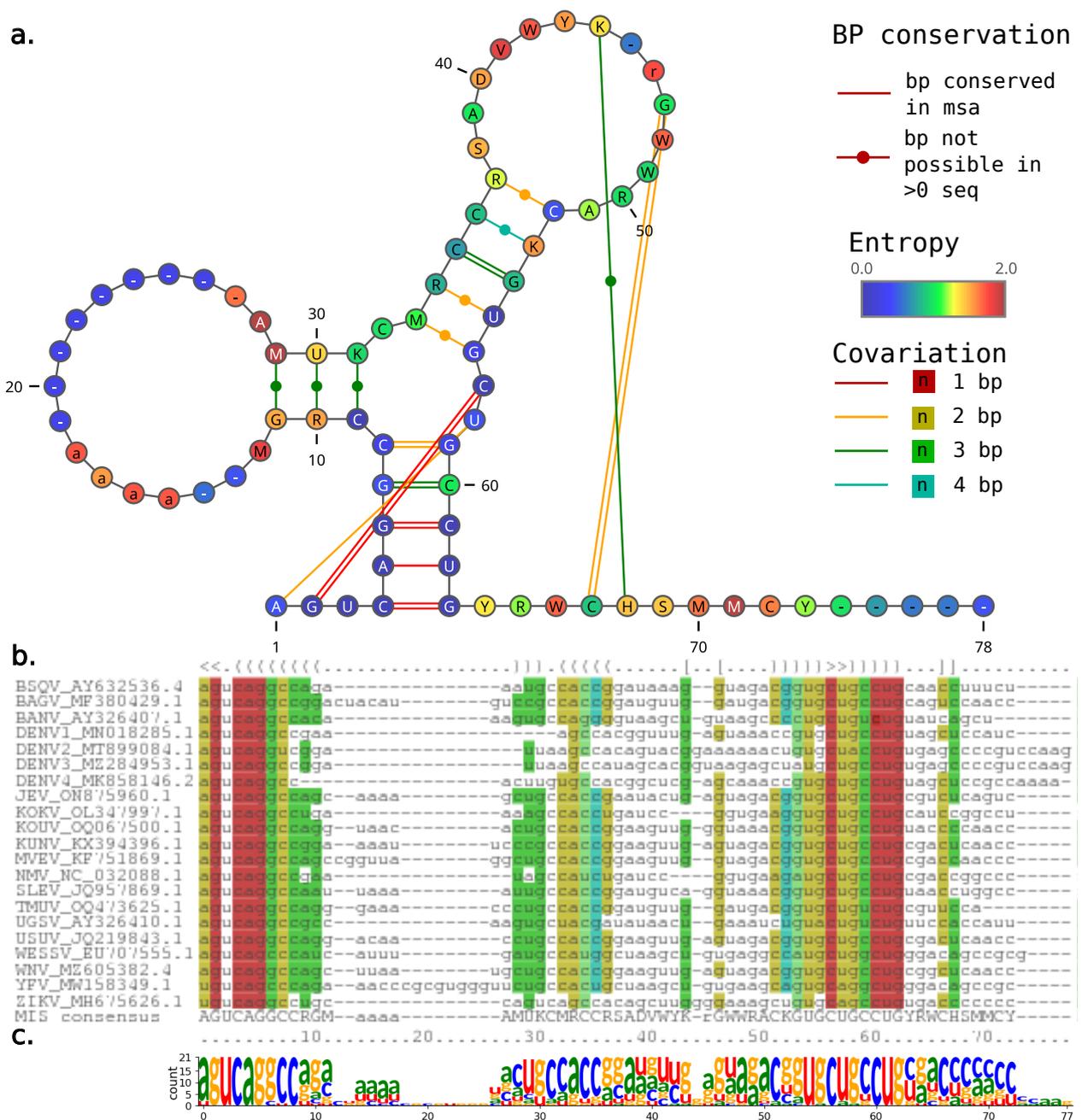
Figure 4.3: **Analysis of MBFV xrRNA1 Secondary Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA1 across 21 MBFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA1 sequence. The stack heights are proportional to the nucleotide frequency at each position.

region (nucleotide number 43-48 and 67-73). The final line of the MSA displays the most informative sequence, conforming to the International Union of Pure and Applied Chemistry (IUPAC) nomenclature. This facilitates the identification of conserved nucleotides and potential functional motifs within the alignment.

The sequence logo for the xrRNA1 sequences of the MBFV viruses MSA provides a quantitative illustration of nucleotide conservation Figure 4.3 panel C.

By observing the positional, enumerated nucleotide stacks, it is confirmed that there is high conservation in region 1-10 and 55-63, corresponding to stem α and pk1. Additionally, these positions show a predominance of cytosine and guanine. The Uracil at position 3 is completely conserved, potentially suggesting a vital role in the structural or functional aspects of the xrRNA1. Regions with variable nucleotide stacks, such as those found at positions 30-40 and 64-73, suggest a higher degree of sequence variability. This could show a region that tolerates mutations which lead to covariation as seen in the msa and structural integrity. The presence of gaps or insertions/deletions in the alignment across the viruses is suggested by the reduction in the total height of stacks at positions 12-28 and 74-77. This may hint at structural flexibility or regions that are less important to the conserved function of xrRNA1.

Figure 4.4 displays a similarity score matrix of xrRNA1 covariance models based on 20 MBFV viruses calculated with CMCompare. The viruses have been sorted according to the similarity of their xrRNA1 covariance models using hierarchical clustering with average linkage. The intensity of the color shading corresponds to the degree of similarity, with darker shades suggesting higher similarities. The matrix is symmetrical, so the comparison between any two models remains unaffected by their order. The highest score is achieved when comparing a covariance model (CM) with itself. Lower scores show greater divergence in the structure and sequence of xrRNA1 between the compared viruses.

Dengue virus 1 (DENV1), Dengue virus 2 (DENV2), Dengue virus 3 (DENV3), and Dengue virus 4 (DENV4), form a cluster with high similarity within this matrix. There is another cluster with high similarity that is central in the matrix and consists of the following viruses: Tembusu virus (TMUV), St. Louis encephalitis virus (SLEV), Kunjin virus (KUNV), Japanese encephalitis virus (JEV), Usutu virus (USUV), West Nile virus (WNV), Bagaza virus (BAGV), and Murray Valley
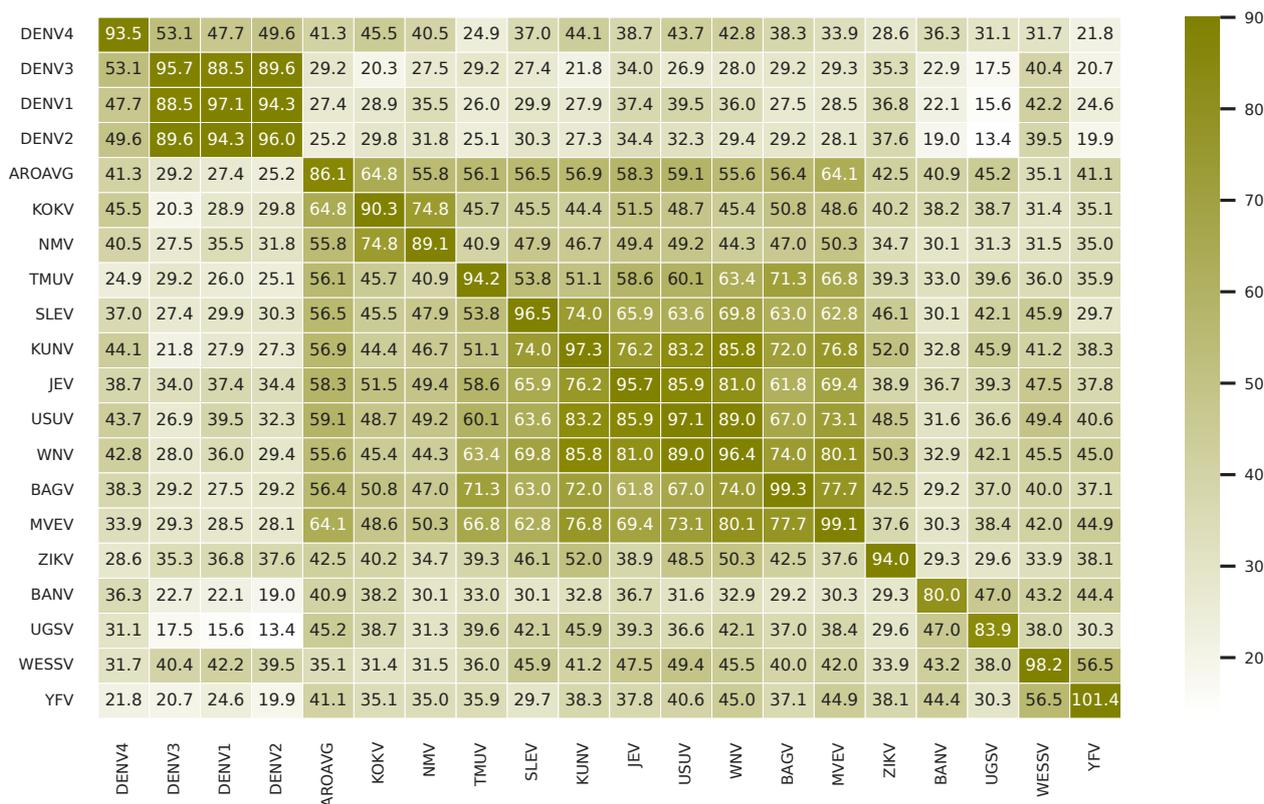
Figure 4.4: **Pairwise Similarity Between Representative xrRNA1s CM in MBFV.** xrRNA1 similarity scores calculated from pairwise covariance model comparison. Diagonal values show the maximum value that can be achieved. A hierarchical clustering with average linkage was performed, grouping xrRNAs based on the similarity of their covariance model scores.

encephalitis virus (MVEV). The following viruses on the on the other hand: Zika virus (ZIKV), Banna virus (BANV), Uganda S virus (UGSV), Wesselsbron virus (WESSV), and Yellow Fever virus (YFV) were found to be the most dissimilar in their xrRNA1 CM from the other viruses. WESSV and YFV show some clustering, although not such much than other virus groups.

4.1.2 MBFV xrRNA2

Based on the analysis of MBFV xrRNA1, the computational analysis of MBFV xrRNA2 2D consensus structure shows similar structural motifs, as demonstrated in Figure 4.5 for the viruses AROAV, BAGV, DENV3, and ZIKV. These consensus

structures of xrRNA2 exhibit the same characteristic three-way junction structure as xrRNA1, but are unique to each virus. The consensus structure retains the triple-stemmed elements accompanied by hairpin loops (β, γ) and a cytosine between the β and γ stems in the multiloop, which is conserved in all viruses. The short pk1 from the beginning to the multiloop and the pk2 from the hairpin γ to the unpaired end are present in every consensus structure. Although detailed intricacies are not restated, the overall structural framework remains consistent with previous findings from the xrRNA1 analysis. In DENV3, we observe strong covariation in the stem γ through the green and orange coloring of three base pairs, suggesting structural significance.

In Figure 4.6 panel A. we see the consensus structure of the 17 viruses. The structure comprises the standard three-way junction with stems (α, β, and γ) and two hairpin loops (β and γ). Additionally, there is a short pseudoknot from the beginning of the structure to the unpaired nucleotides of the multiloop. This analysis benefits from the inclusion of diverse virus sequences, which provides a stronger signal for covariation and positional entropy. These are key indicators of structural conservation and variability, unlike the consensus structures of each virus, as seen in Figure 4.5.

The consensus structure shows relatively low entropy at the first pseudoknot and along stem α, along with some covariation. This suggests a highly conserved sequence and structure across the different viruses. The first base pair of stem β appears to be conserved across all viruses, as shown by the red base pair (no covariation) and low sequence variability (low entropy). However, the last two base pairs of stem β exhibit high entropy and covariation. The central dot within the base pairs shows that not all sequences consistently form the base pairs, suggesting a less conserved structure. The β hairpin contains many gaps, which were introduced by a single virus, YFV. Stem γ's first two base pairs are conserved over all viruses and show no entropy. The base pairs of stem γ exhibit high entropy but strong covariation, suggesting an intermediate level of structural conservation. The teal coloring shows covariation with three different base pairs, while violet/pink shows covariation with four different base pairs. The central dot within the base pairs denotes inconsistency in their formation across all sequences. Notably, pk2 does not form in the consensus structure. The high entropy observed in the region corresponding to

43
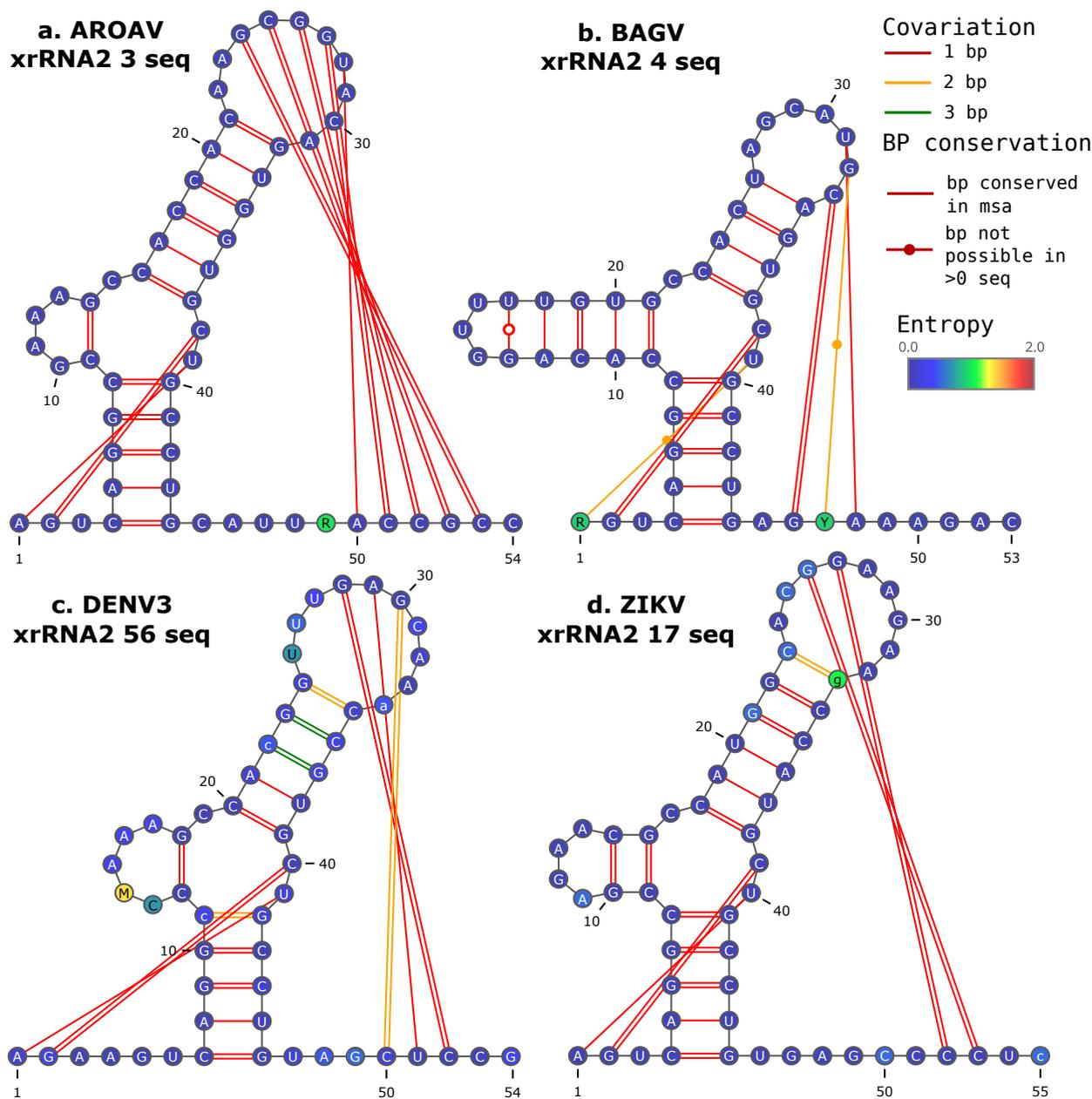
Figure 4.5: **Consensus Structure Predictions of xrRNA2 across MBFV viruses.** This figure illustrates the consensus structure predictions for xrRNA2 performed using RNAalifold on four exemplary picked viruses in MBFV: a. Aroa virus (AROAV) with 3 sequences, b. Bagaza virus (BAGV) with 4 sequences, c. Dengue virus (DENV3) with 56 sequences and d. Zika virus (ZIKV) with 17 sequences. The base pairing color represents the covariation in the structure: red for conserved base pairs, orange for two possible bps, and green for three possible bps. Nucleotide colors are based on positional Shannon entropy within the MSA of each virus, showing nucleotide variability: dark blue (entropy = 0, no variability), green (entropy = 1, moderate variability), and red (entropy = 2, high variability).

Figure 4.6: **Analysis of MBFV xrRNA2 Secondary Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA2 across 17 MBFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA2 sequence. The stack heights are proportional to the nucleotide frequency at each position.

pk2 suggesting a high degree of variability and gaps introduced by some sequences.

In Figure 4.6 panel B. we see the MSA of the 17 virus sequences that were used to create the consensus structure.The alignment displays the consensus structure in dot-bracket notation, simplifying the complex secondary structure. It highlights regions of high conservation, within stem α and pk1, as well as the beginning of stem β and stem γ. These columns exhibit uniform coloring without any disruptions. However, the alignment exposes gaps in certain regions, specifically within the hairpin β (nucleotide number 11-25) and pk2 region (nucleotide number 37-48 and 63-73). The final line of the MSA displays the most informative sequence. The mutual information score (MIS) can be obtained from the sequence logo of the MSA, as shown in Figure 4.6 panel C.

Observing the positional, enumerated nucleotide stacks, we can observe the following and confirm the previous observation of high conservation in region 1-8 and 52-60, corresponding to stem α and pk1. Also these positions show a predominance of cytosine and guanine. The Uracil at position 3 is completely conserved as in xrRNA1 of MBFV as seen in Figure 4.6 panel C., potentially suggesting an important role in the structural or functional aspects of the xrRNA. The first base pair of pk1 (nucleotide number 0 and 55) exhibits mixed stacks, suggesting compensatory mutation and recovery of the base pair. Variable regions with lower stacks and a mix of nucleotides, such as those observed in positions 35-44 and 61-73, suggest a higher degree of sequence variability in these regions. This could show a region that compensates mutations which lead to covariation as seen in the MSA and structural integrity. The presence of gaps or insertions/deletions in the alignment across the viruses is suggested by the reduction in the total height of stacks at positions 11-25 and 74-77. This may hint at structural flexibility or regions that are less important to the conserved function of xrRNA1.

Figure 4.7 displays a matrix of similarity scores for xrRNA2 covariance models based on 16 MBFV viruses. DENV4, WESSV, BANV and UGSV were excluded from this matrix as they do not have xrRNA2. The highest score is obtained when a CM is compared with itself. Lower scores suggest greater divergence in the structure and sequence of xrRNA2 between the compared viruses.

DENV1, DENV2 and DENV3 show similarity to their xrRNA1 counterparts. However, unlike their xrRNA1 counterparts, the other viruses do not cluster to-
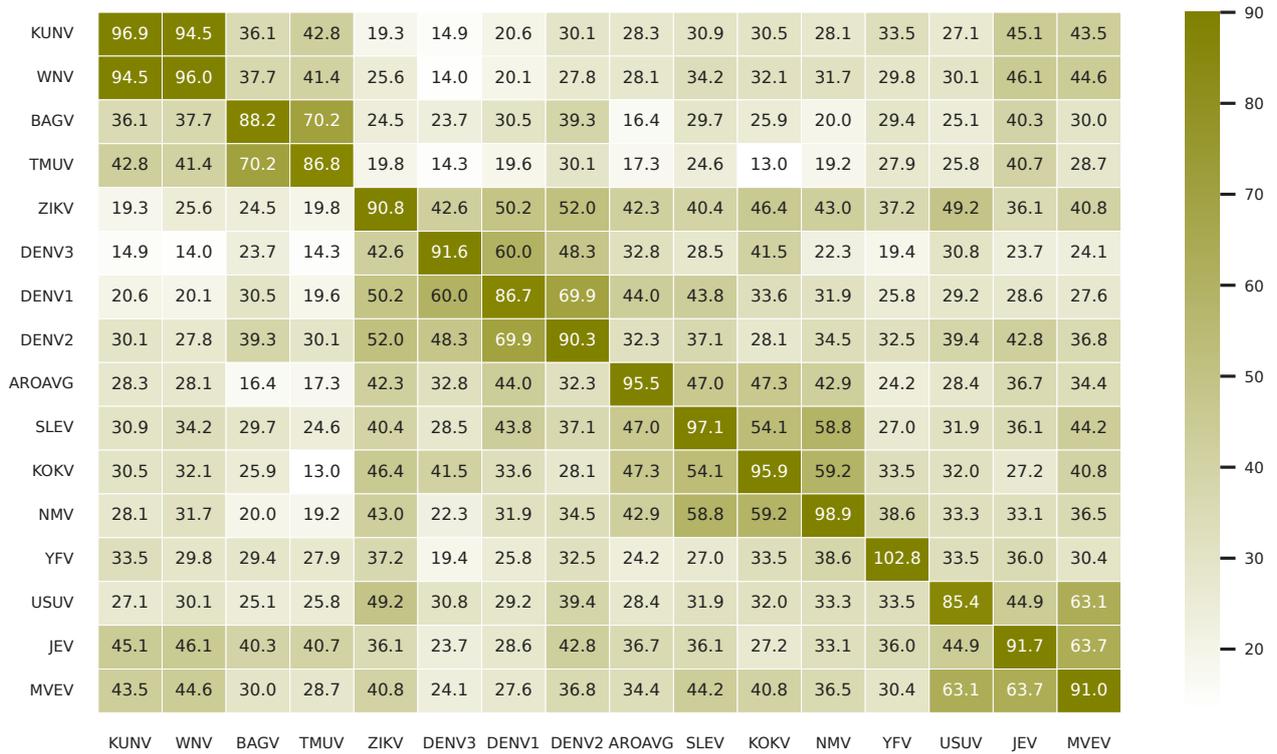
Figure 4.7: **Pairwise Similarity Between Representative xrRNA2s in MBFV.** xrRNA2 similarity scores calculated from pairwise covariance model comparison. Diagonal values show the maximum value that can be achieved. A hierarchical clustering with average linkage was performed, grouping xrRNAs based on the similarity of their covariance model scores.

gether. The xrRNA2 of KUNV and WNV, BAGV and TMUV, and JEV, MVEV and USUV show high similarity scores. YFV and ZIKV appear to be the most dissimilar to other xrRNA2.

4.1.3 MBFV 3D Structure Analysis

Figure 4.8 panel A shows the 3D structure of the AROAV xrRNA1 molecule predicted by simRNA and energy-minimized using QRNAS. The core of the molecule forms a ring-like structure, anchored around the 5' region of the sequence and stabilized by the presence of pk1. The ring is composed of nucleotides spanning positions 34 to 46, aligned with the second base-pairing sequences of the α and γ stems. The arrangement and positioning of these nucleotides suggest that the stems play an

important role in the xrRNA stalling mechanism against the XRN1 enzyme. PK2 seems to attach to the terminal segment of the molecule, possibly providing a mechanical counterbalance or tension relief as the XRN1 enzyme exerts force on the structure during its degradation process. The β stem and hairpin loop appear to be positioned peripherally and isolated from the central interactive dynamics of the xrRNA1 structure. This suggests that they may have a limited or non-contributory role in the direct resistance mechanism against XRN1 activity, possibly hanging adjacent to the main structure without much interaction with the core resistance features.

4.1.4 MBFV xrRNA1 and xrRNA2 Comparison

This subsection presents a comparison of xrRNA1 and xrRNA2 within MBFV across 17 viruses that have at least two xrRNA. The main objective is to find out whether the observed structural features of these RNA elements are uniform or display much variability. Visual assessment is employed through boxplot representations that illustrate the length distribution of each structural component. Additionally, we support our visual observations with Welch's t-test to calculate if the differences in lengths are statistically significant or due to random variation. Finally, we will compare the CM of each xrRNA to observe their clustering patterns, enabling us to evaluate the similarities and differences between xrRNA1 and xrRNA2.

The boxplots in Figure 4.9 provide a visual summary of the structural element lengths for MBFV xrRNA1 and xrRNA2 molecules. The data are segmented into seven categories representing different structural elements: α stem, β stem, β loop, β sum, γ loop, γ stem, and pseudoknot2. In the comparison, xrRNA1 is colored in green, and xrRNA2 in blue. For the α stem, the lengths for both xrRNA1 and xrRNA2 are the same, with a variance of 0 and a median of 5 base pairs, suggesting high structural conservation. The β stem shows a more varied range, but it appears to be very similar for xrRNA1 and 2. The β loop presents as the smallest structural element for both xrRNA types with a median of five nucleotides for xrRNA1 and four nucleotides for xrRNA2. A noticeable contrast is seen in the γ elements; xrRNA1 exhibits a shorter γ stem but compensates with an extended
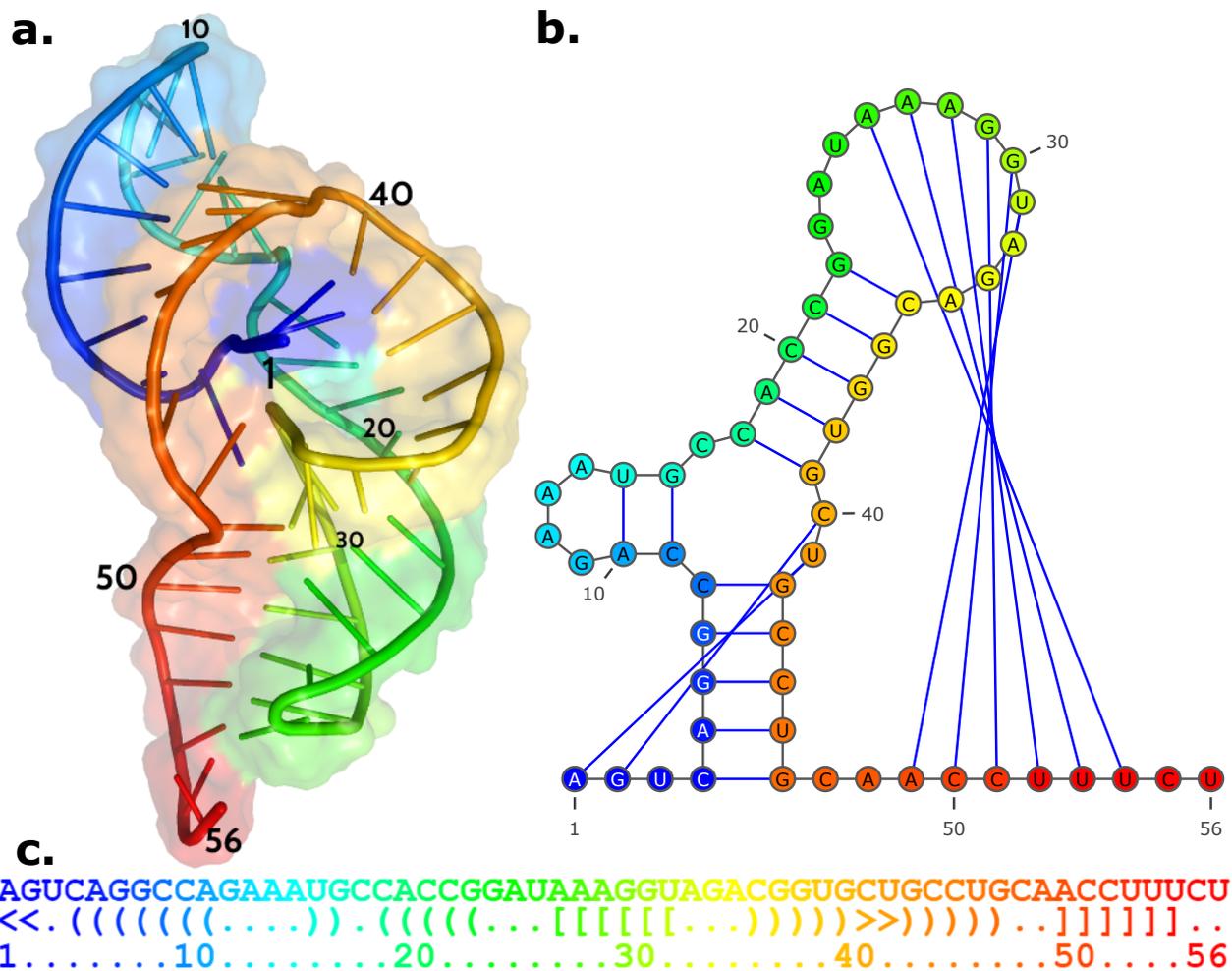
Figure 4.8: **3D structural analysis of AROAV xrRNA1.** Panel A shows the three-dimensional structure of AROAV xrRNA1 in cartoon representation, as predicted by sim-RNA and subsequently energy-minimised using QRNAS. The visualisation uses a gradient rainbow colour scheme, starting with dark blue at the beginning of the sequence and progressing through the spectrum to red at the end of the sequence, to correlate to the nucleotide positions. A semi-transparent surface rendering, created using PyMOL, is superimposed on the cartoon model to provide a contour perspective of the molecular surface. Panel B displays the corresponding secondary structure, using the same colour scheme as the 3D structure, to facilitate correlation between the two-dimensional layout and the three-dimensional conformation. Panel C shows the dot-bracket notation sequence of AROAV xrRNA1 with highlighted secondary structure elements. The nucleotide sequence follows the same colour coding, providing a consistent visual reference across all panels and emphasising the continuity from sequence to structure.
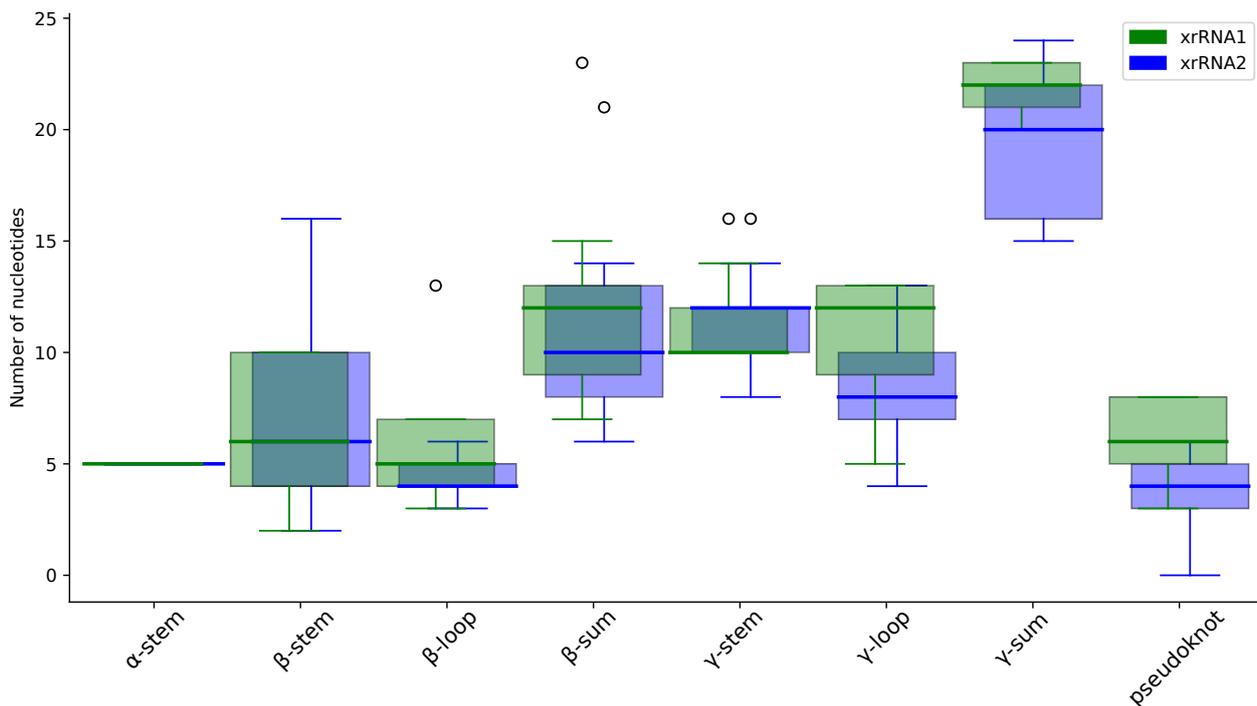
49

Figure 4.9: **Structural Element Length Distribution for MBFV xrRNA1 and xr-RNA2.** Boxplots illustrating the length distribution of different structural elements in MBFV xrRNA1 (in green) and xrRNA2 (in blue). The analyzed structural elements include the α stem, β stem, β loop, γ stem, γ loop, and the pseudoknot. The median length is represented by the thick line in each box, the interquartile range is represented by the edges of the box. Outliers are shown as individual points. These distributions offer a visual representation of the range and median sizes of each structural element in the two xrRNA species.

γ loop relative to xrRNA2. This inverse relationship in the length of γ structural elements may suggest a correlation between them that warrants further examination. Finally, the pseudoknot2 element length appears to be greater in xrRNA1 than in xrRNA2, which conform with the previous assessment that the xrRNA1 γ loop is longer than xrRNA2, since the pk2 paires in this region.

The statistical comparison of the lengths of various structural elements between xrRNA1 and xrRNA2 of MBFV is provided by the Welch's t-test in Figure 4.10.

The analysis did not include the α stem, which always measured 5 base pairs without variance. As for the β elements, the β stem did not show a huge difference in length between xrRNA1 and xrRNA2, as shown by its high p-value of 0.8540. However, the β loop showed a p-value of 0.0465, which is just below the α signif-
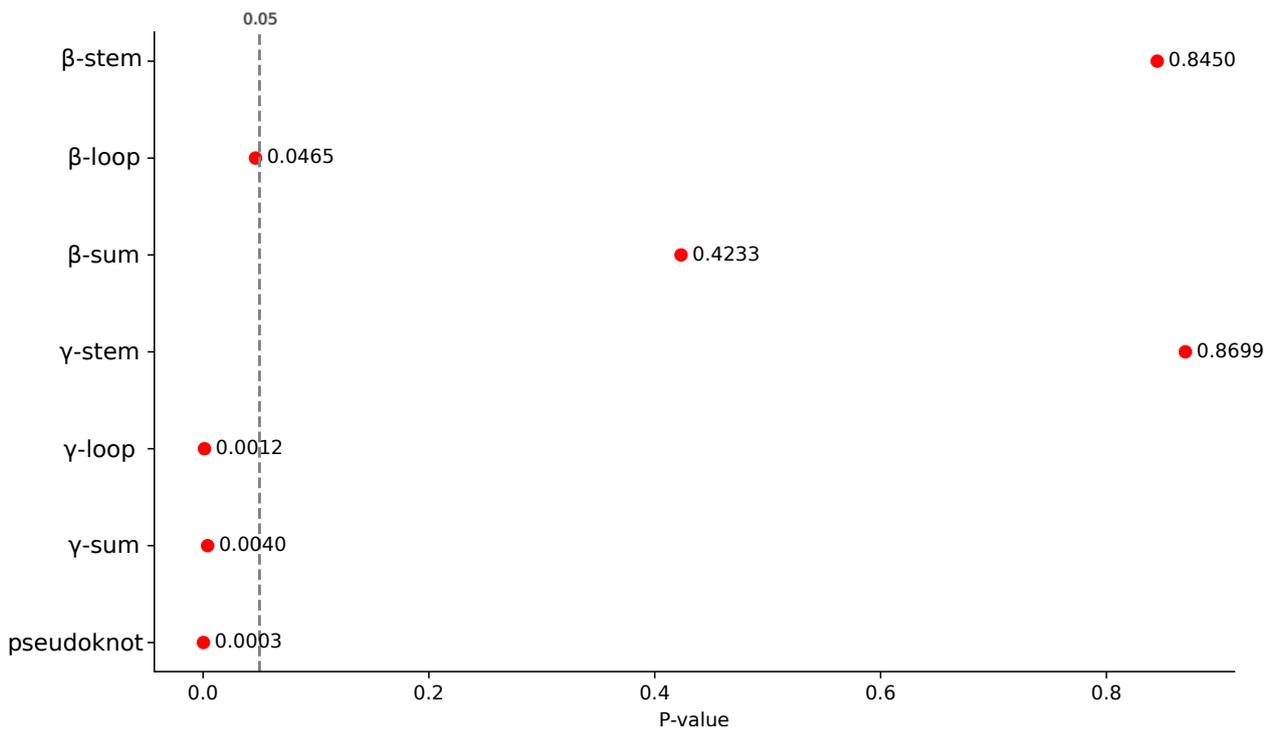
Figure 4.10: **Welchs t-test between length of xrRNA1 and 2 elements of MBFV.** Statistical Analysis of Nucleotide Length Variation in Structural Elements of xrRNA1 and xrRNA2. This figure presents the results of Welch's t-test comparing the number of nucleotides (length) in various structural elements between xrRNA1 and xrRNA2. The p-value measures the probability that an the mean of the two elements are the same. The lower the p-value, the less likely the observed data would be if there were no actual difference.
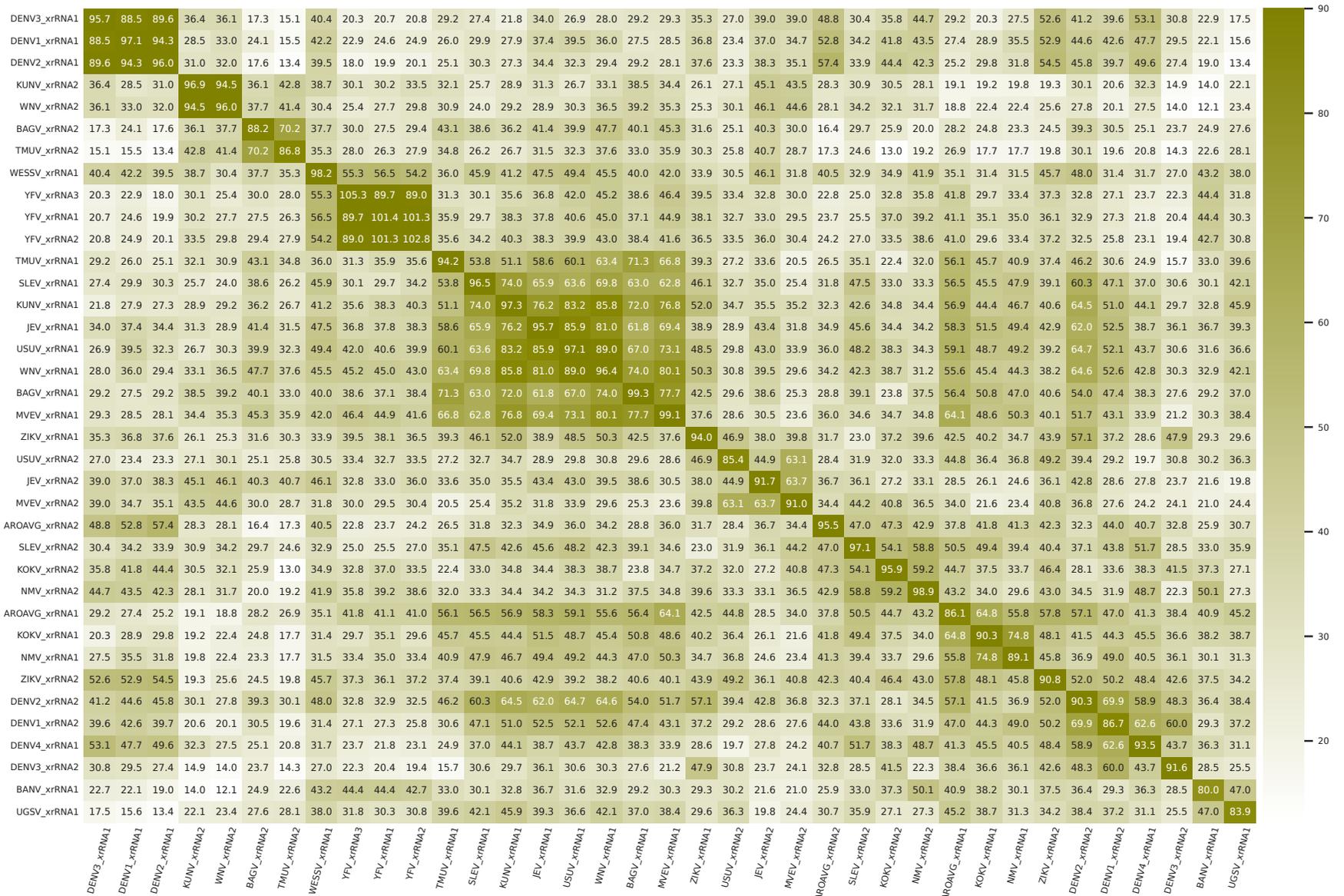
icance level of 0.05, suggesting a significant difference in length between xrRNA1 and xrRNA2 for the β loop. The combined β elements show a p-value of 0.4233, suggesting that the length variation does not reach statistical significance. The γ stem also shows no significant difference in length with a p-value of 0.8699, similar to the β stem. However, the γ loop and γ sum have p-values of 0.0012 and 0.0040 respectively, both well below the 0.05 threshold. The analysis shows a significant difference in length between xrRNA1 and xrRNA2 for these elements, as shown by the pseudo-knot structure with a p-value of 0.0003. This suggests that the two RNA variants may have different structural properties. The results of Welch's t-test show significant differences in length between xrRNA1 and xrRNA2 for certain structural elements, including the β loop, γ loop, γ sum and pseudoknot. However, no sig-

nificant differences were observed for other elements such as the β stem and the γ stem.

In our study of MBFV xrRNA1 and xrRNA2, we conducted a Link Score analysis across all covariance models (CMs) of xrRNA between each MBFV virus. The results, as shown in Figure 4.11, primarily grouped xrRNA1 and xrRNA2 from the same virus together, with some exceptions.

Unusual clustering was observed with the xrRNA1 from DENV4, which appears to cluster more closely with xrRNA2 from viruses DENV1 through DENV3 rather than with other xrRNA1. This clustering pattern might suggest a unique evolutionary path for DENV4, which could involve the loss of a previous xrRNA1 and the now xrRNA1 is the supposed xrRNA2.

The Yellow Fever virus (YFV) xrRNAs exhibit a unique clustering pattern. xrRNA1-3 of YFV primarily group with each other and differ from the rest, except for a similarity with WESSV xrRNA1. xrRNA1 shows a distinct cluster, suggesting a higher degree of similarity within this group across various viruses. In contrast, the xrRNA2 cluster appears less defined, suggesting a greater degree of divergence among xrRNA2 across different viruses. This observation suggests that there may be greater variability in the structural and functional aspects of xrRNA2 compared to xrRNA1 among the MBFV viruses.

Figure 4.11: **Pairwise Similarity Score between all MBFV xrRNAs CM.** xrRNA similarity scores calculated from pairwise covariance model with CMCompare. Diagonal values show the maximum value that can be achieved. A hierarchical clustering with average linkage was performed, grouping xrRNAs based on the similarity of their covariance model scores.

4.1.5 MBFV xrRNA Design

In this section we focus on the foundational work for future xrRNA design. It builds upon the analysis of the xrRNA structures within the MBFV. Our objective is to pioneer design strategies for xrRNA, using our understanding of the structural nuances of xrRNA1 and xrRNA2, as well as their lengths across 17 viruses.

Figure 4.12 shows detailed boxplot analysis that quantitatively map the lengths of various structural elements within xrRNA1 and xrRNA2. This data serves as a scaffold for the design blueprint.
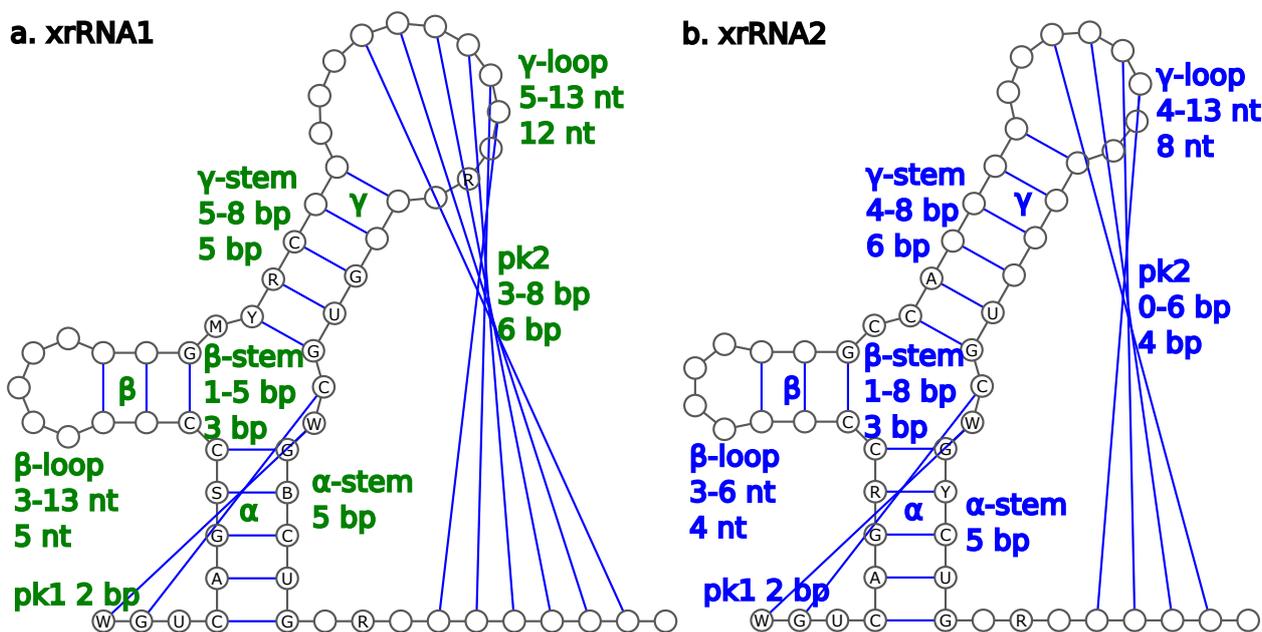

Figure 4.12: **Comparison of Secondary Structures of MBFV xrRNA1 and xrRNA2.** Panel A displays the structure of xrRNA1 and panel B shows the structure of xrRNA2. Both structures were constructed using median structural element lengths calculated in boxplot analysis for each consensus structure in MBFV. The shown nucleotides represent highly conserved regions with at least 90% sequence identity in the MSA for each respective xrRNA. Each structural element, identified by its functional annotation (α, β, γ, and pseudoknot pk1 and pk2), is quantified by the range of lengths (minimum to maximum) followed by the median length in nucleotides (nt) or base pairs (bp), allowing for a direct comparison of the structural variance between xrRNA1 and xrRNA2.

The non-variable nucleotides in a design should be the ones present at each position in the structure. This is because they correlate to a sequence identity

of 90% or more in the corresponding MSA, providing insights into the conserved and variable regions across the viruses. It can be observed that pk1, stem α, the multiloop, and the first 2-3 bp of stem γ are sequence-conserved in and between xrRNA1 and xrRNA2, suggesting their importance to the xrRNA mechanism.

Calculating the correlations between the lengths of structural elements in xrRNA is important for understanding their integrity and functionality. Figure Figure 4.13 shows a near-zero correlation between the lengths of the β stem and β loop, suggesting that their lengths may not be important for their function. This observation is consistent across both xrRNA1 and xrRNA2, highlighting potential flexibility in the structural design of these regions.
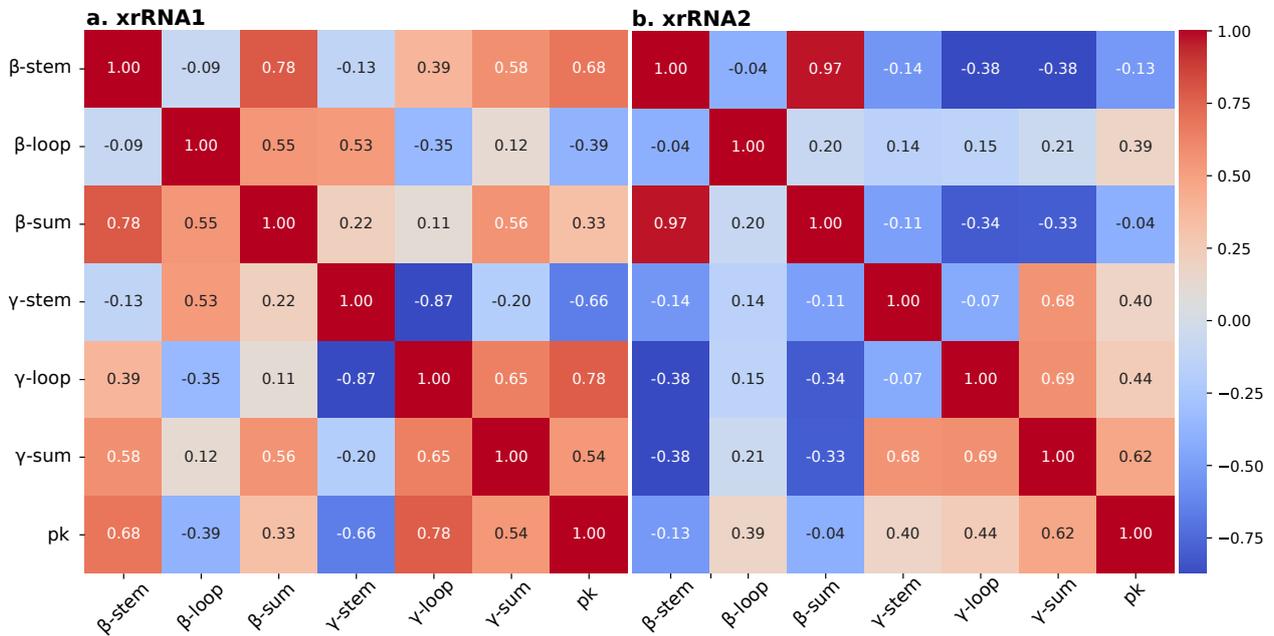

Figure 4.13: **Pearson Correlation Coefficients for Lengths of different xrRNA1 and xrRNA2 elements in MBFV.** This table presents the calculated Pearson correlation coefficients for the lengths of structural elements of 17 xrRNA1 adn 14 xrRNA2, with one representatives from each virus. The coefficient ranges from -1 (dark blue) to 1 (dark red) and provides insights into the linear relationship between the structural elements of the xrRNA. A positive number in the correlation matrix shows us an increase in the length of one structural element with an increase in the length of another, while a negative number shows us an inverse relationship, where the lengthening of one element gives us the shortening of another.

On the other hand the xrRNA1 shows a negative correlation between the length

of the γ stem and the length of the γ loop, which is not observed in xrRNA2. In addition, a positive correlation could be observed between the length of the γ loop and the pseudoknot region. This pattern makes sense, since the pseudoknot2 forms base pairs in the γ region.

4.2   TBFV xrRNA Structure Analysis

Tick-borne flaviviruses (TBFV) pose unique challenges and opportunities for studying xrRNA structure and function. Similar to MBFV, TBFVs, which include the Tick-borne encephalitis virus (TBEV) and Powassan virus (POWV), as well as 14 other viruses, have been computationally analyzed to show the complexities of their xrRNA configurations. This section presents the computational findings of xrRNA1 and xrRNA2 within TBFV.

Beginning with an analysis of consensus structure predictions, this investigation utilises RNAalifold to capture the fundamental base pairing patterns present in TBFV xrRNAs. SimRNA and Infernal are employed as auxiliary tools to provide three-dimensional insights and alignment-based comparisons, respectively. The study identifies structurally conserved regions that play an important role in the survival of viral RNA within the host. This is done by assessing the covariation and Shannon entropy within MSA. Statistical methods, such as Welch's t-test and Pearson correlation coefficients, were used to quantify the variability and inter-element relationships within the xrRNA structures. This study examines the structural motifs of TBFV xrRNA. The analysis reveals a common three-way junction at the core of the architecture, composed of the α, β, and γ stems. The β and γ form adjacent hairpin loops, as seen in Figure 4.14. These common motifs are evident in the median consensus structures for TBFV xrRNA1 and xrRNA2, which are built from the median length of each element in the consensus structure of each virus. The central link of the structure is a pseudoknot, known as pk1, which spans from the multiloop to the midpoint of the unpaired sequence at the end of the structure. This pseudoknot is made up of a compact three-base pair formation. The xrRNA structure has a long α stem that extends into a multiloop, which serves as the core for the three-way junction. The β and γ stems, each with their corresponding hairpin loops, extend from this central multiloop. At the end of the xrRNA, there is a

second pseudoknot (pk2) that connects back to the γ hairpin loop and extends to the end of the unpaired sequence. Additionally, we measured the segment labelled as External M, which extends from the end of the α stem to the beginning of the first pseudoknot, providing a measurement for this external region of the xrRNA.
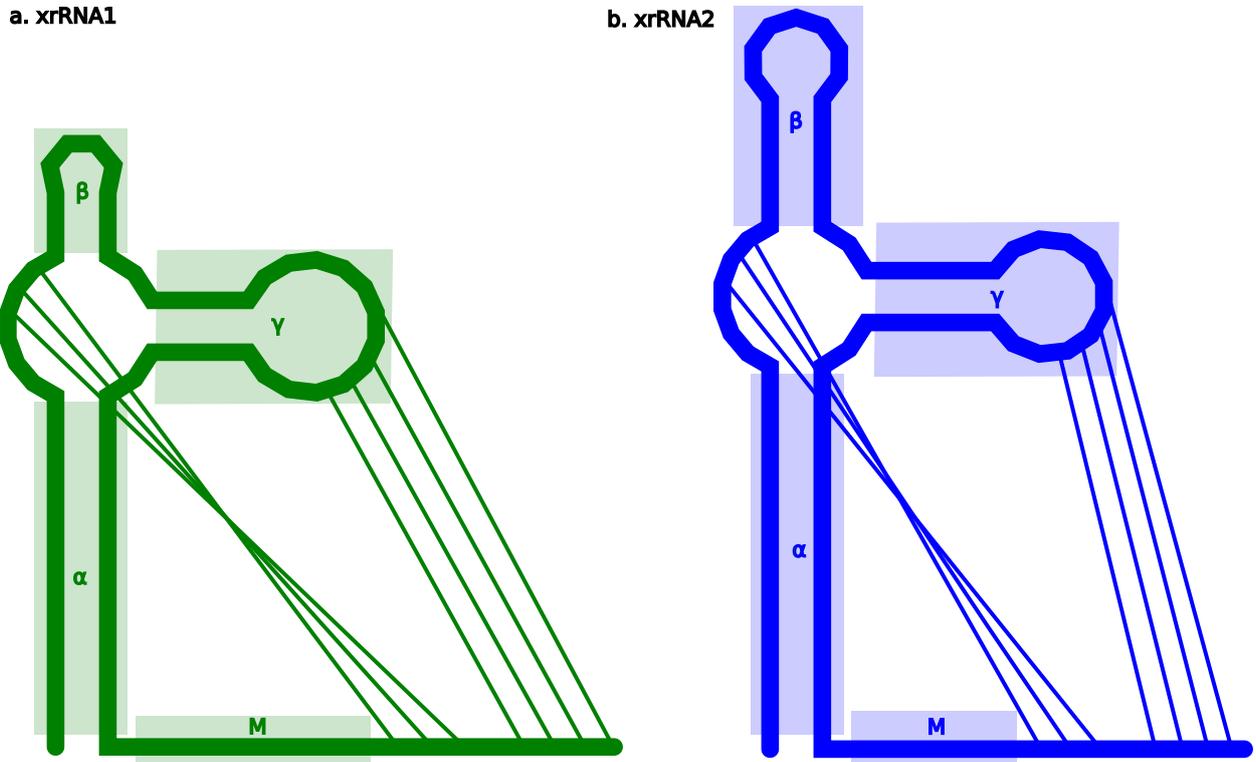

Figure 4.14: **Median Consensus Structure of TBFV xrRNA1 and xrRNA2**. This figure offers a side-by-side comparison of TBFV xrRNA1 and xrRNA2 median structures, coming from 14 (for xrRNA1) and 10 (for xrRNA2) virus structures. The length of the structural elements was calculated by taking the median length from each TBFV xrRNA consensus structure.

4.2.1 TBFV xrRNA1 and xrRNA2

The analysis conducted using RNAalifold reveals both conserved and variable regions in the structural analysis of TBFV xrRNA1 and xrRNA2, which may hold significance in their functional mechanisms. The data from a variety of TBFV viruses is used to build a detailed representation of the xrRNA architecture. As

an example, we examine the xrRNA1 and xrRNA2 of tick-borne encephalitis virus (TBEV) and Powassan virus (POWV) in Figure 4.15. We calculated the consensus structure for each virus. For TBFV xrRNA1 and xrRNA2, the three-way junction remains a central motif, with the multiloop at its core exhibiting a high degree of conservation. This suggests an important role for the multi-loop in the structural stability and potential interaction with viral or host proteins. The α stem, especially as it nears the multiloop, shows a growing conservation of G-C base pairs, suggesting a structurally important area.

Pseudoknot 1 is identified by covariation and strong conservation. It extends from the multiloop and may be an important element for the structural integrity of the xrRNA. The β stem also exhibits conserved base pairing, notably near the multiloop, showing an evolutionarily preserved function. Conserved adenosines are present across viruses in the β and γ stems, as well as between the γ stem and α stem, highlighting their structural importance. The γ stem is characterised by the robust conservation of guanine-cytosine base pairs. This leads into the γ hairpin loop, which also exhibits conserved base pairing within pk2. There is little variability noted immediately before and after this region. The conservation of the γ stem and loop, along with their relationship to pk2, highlights their importance in maintaining the structural framework of the xrRNA. The section referred to as External M, which spans from the end of the α stem to the beginning of pk1, exhibits the greatest degree of variability.

Figure 4.16 panel A. shows the consensus structure of xrRNA1 from 14 different viruses of TBFV. This analysis confirms previous findings on the structural conservation and variability within TBFV. Specifically, the consensus structure's multiloop core and γ stem exhibit low Shannon entropy, with a prevalence of guanine-cytosine (G-C) base pairs. Additionally, the pseudoknots pk1 and pk2 in the TBFV xrRNA1 structure exhibit a high degree of conservation, as demonstrated by covariation. The covariation observed within pseudoknot regions suggests that there is an evolutionary pressure to maintain base pairing, which is essential for the stability and function of the xrRNA. Conversely, the external M region and the β loop, as well as segments preceding and following the pk2 region within the γ loop, display a higher degree of variability. The xrRNA1 regions exhibit increased Shannon entropy, showing sequence diversity across the analyzed viruses. The external M region and the
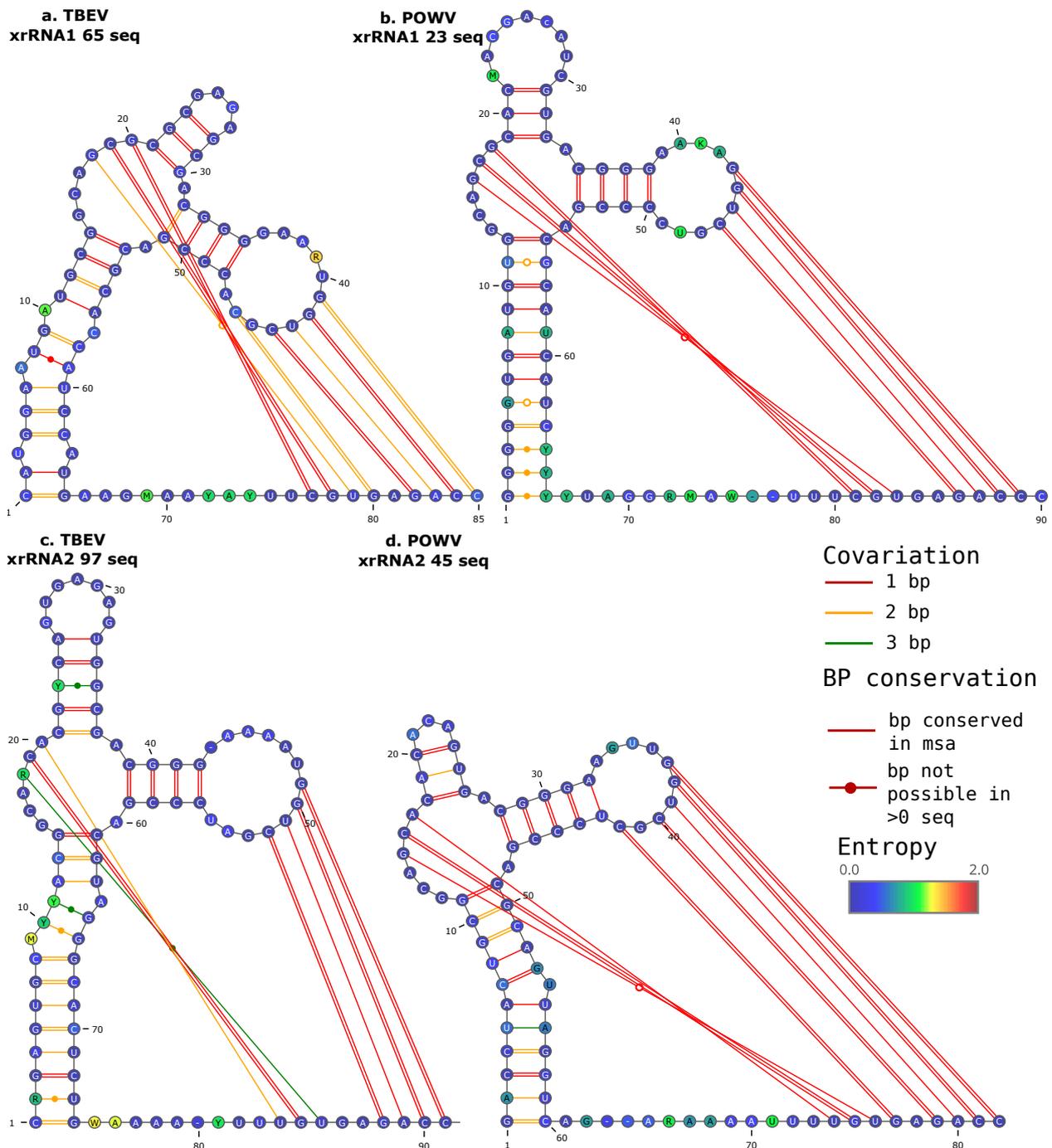
Figure 4.15: **Consensus Structure Predictions of xrRNA1 and xrRNA2 across two TBFV viruses.** The figure shows the consensus structure predictions for xrRNA1 and xrRNA2 in two viruses of TBFV, namely Tick-borne encephalitis virus (TBEV) with 65 sequences and Powassan virus (POWV) with 23 sequences for xrRNA1, and TBEV with 97 sequences and POWV with 45 sequences for xrRNA2. The predictions were performed using RNAalifold. The colours used in base pairing represent covariation in the structure: red for conserved base pairs, orange for two possible bps, and green for three possible bps. Nucleotide colours are based on positional Shannon entropy within the MSA of each virus, suggesting nucleotide variability: dark blue (entropy = 0, no variability), green (entropy = 1, moderate variability), and red (entropy = 2, high variability). The sequences displayed are the most informative sequence coming from each MSA.
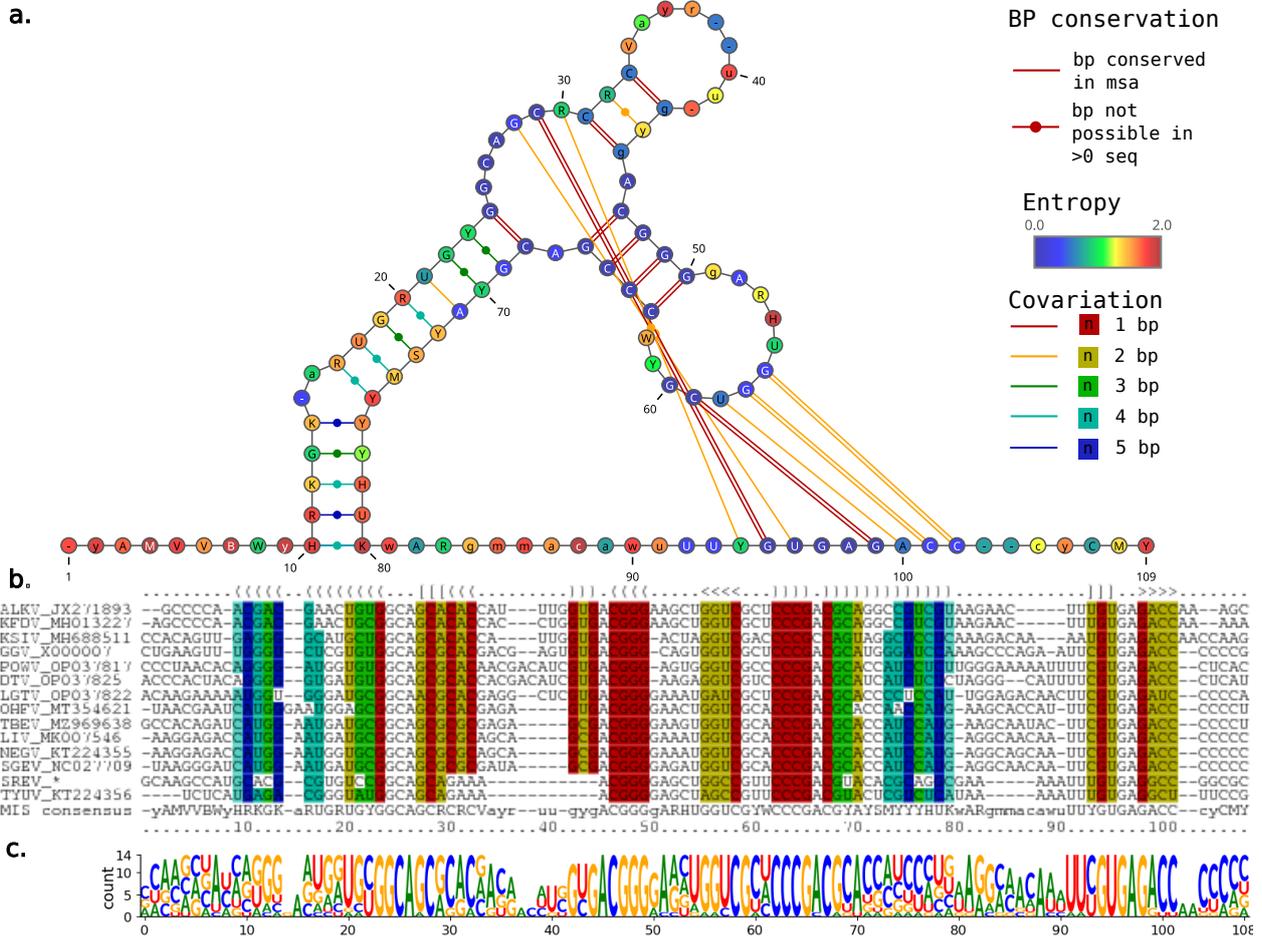
Figure 4.16: **Analysis of TBFV xrRNA1 Consensus Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA1 across 14 TBFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA1 sequence. The stack heights are proportional to the nucleotide frequency at each position.

60

regions around pk2 exhibit high variability.

The MSA in Figure 4.16 panel B. of xrRNA1 from 14 distinct viruses within the TBFV supports the consensus secondary structure predictions. The dot-bracket notation also includes pseudoknots. This MSA highlights the sequence diversity across different TBFV viruses and emphasizes the structural elements that could be essential for Xrn1 function. The α stem region of this MSA shows strong covariation, showing the evolutionary pressure to maintain base-pairing integrity, which is important for the structural stability of xrRNA1. This covariation highlights the importance of the α stem in the overall architecture of xrRNA1. The β stem, γ stem, and pseudoknot regions (pk1 and pk2) exhibit conserved base pairing with slight covariation. The γ stem plays a vital role in the tertiary structure formation of xrRNA1, as evidenced by its high degree of base pair conservation.

The sequence logo for the xrRNA1 sequences of the TBFV viruses MSA provides a quantitative illustration of nucleotide conservation Figure 4.16 panel C. By observing the positional, enumerated nucleotide stacks, it is confirmed that there is high conservation in region 1-10 and 55-63, corresponding to stem γ and pk1 and pk2. Additionally, these positions show a predominance of cytosine and guanine. The adenine at positions 46 and 67 are completely conserved, suggesting a vital role in the structural or functional aspects of the xrRNA1. Regions with a mix of nucleotides, such as those observed in positions 10-22 and 71-80, suggest a higher degree of sequence variability. This could show a region that tolerates mutations which lead to covariation as seen in the MSA and structural integrity, as seen in stem α. The presence of gaps in the alignment across the viruses is suggested by places where the total height of stacks is reduced, specifically at positions 34-44 and 81-90. This could suggest structural flexibility or regions that are less important to the conserved function of xrRNA1, such as the β loop or External M.

The consensus structure of TBFVs xrRNA2 Figure 4.17 panel A. shows us a similar structure compared to Figure 4.16 panel A. the consensus structure of xrRNA1 in TBFVs. The conserved pattern around the multi-loop core, stem γ, pk1, and pk2 is observed. Covariation is also evident in pseudoknots 1 and 2.

Stem α on the other hand shows high shannon entropy, while the high Shannon entropy in this region may suggest a high degree of sequence variability, the MSA in Figure 4.17 panel B. provides additional information by showcasing that not every

Figure 4.17: **Analysis of TBFV xrRNA2 Consensus Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA2 across 10 TBFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA2 sequence. The stack heights are proportional to the nucleotide frequency at each position.

sequence can form every base pair within this α stem.

4.2.2 TBFV 3D Structure Analysis

The 3D structure shown in Figure 4.18 panel A, predicted by simRNA and energy minimised using QRNAS, reveals a ring-like formation at the beginning of the DTV xrRNA1 structure. The core structure seems to be centred on the 5' region of the sequence. This ring comprises nucleotides from positions 61 to 76, aligned with the second base-pairing sequences of the α and the beginning of the External M region. PK1 and PK2 attach to the terminal segment of the molecule, possibly providing mechanical counterbalance or tension relief as the XRN1 enzyme exerts force during degradation. The β stem and hairpin loop, consisting of nucleotides 21-35, are positioned towards the back of the molecule and appear to have limited or no role in the direct resistance mechanism against XRN1 activity. It is possible that they are adjacent to the main structure without much interaction with the core resistance features.

4.2.3 TBFV xrRNA1 and xrRNA2 Comparison

This subsection presents an analysis of xrRNA1 and xrRNA2 within TBFV across 10 viruses that have at least two xrRNA. The main objective is to understand whether the observed structural features of these RNA elements are uniform or display high variability. Visual assessment is employed through boxplot representations that illustrate the length distribution of each structural component. We have also reinforced our visual findings by conducting Welch's t-test to calculate whether the differences in lengths are due to random variation or if they are statistically significant. Finally, we will compare the CM of each xrRNA to observe the way they cluster. This approach will enable us to evaluate the similarities and differences between xrRNA1 and xrRNA2, as we did in MBFV.

Figure 4.19 show the lengths of structural elements for MBFV xrRNA1 and xrRNA2 as boxplots. The data is separated into six categories: α stem, β stem, β loop, γ loop, γ stem, and External M. The lengths of α stem for both xrRNA1 and xrRNA2 vary slightly, with a median of 12 and 13 base pairs. The β stem exhibits
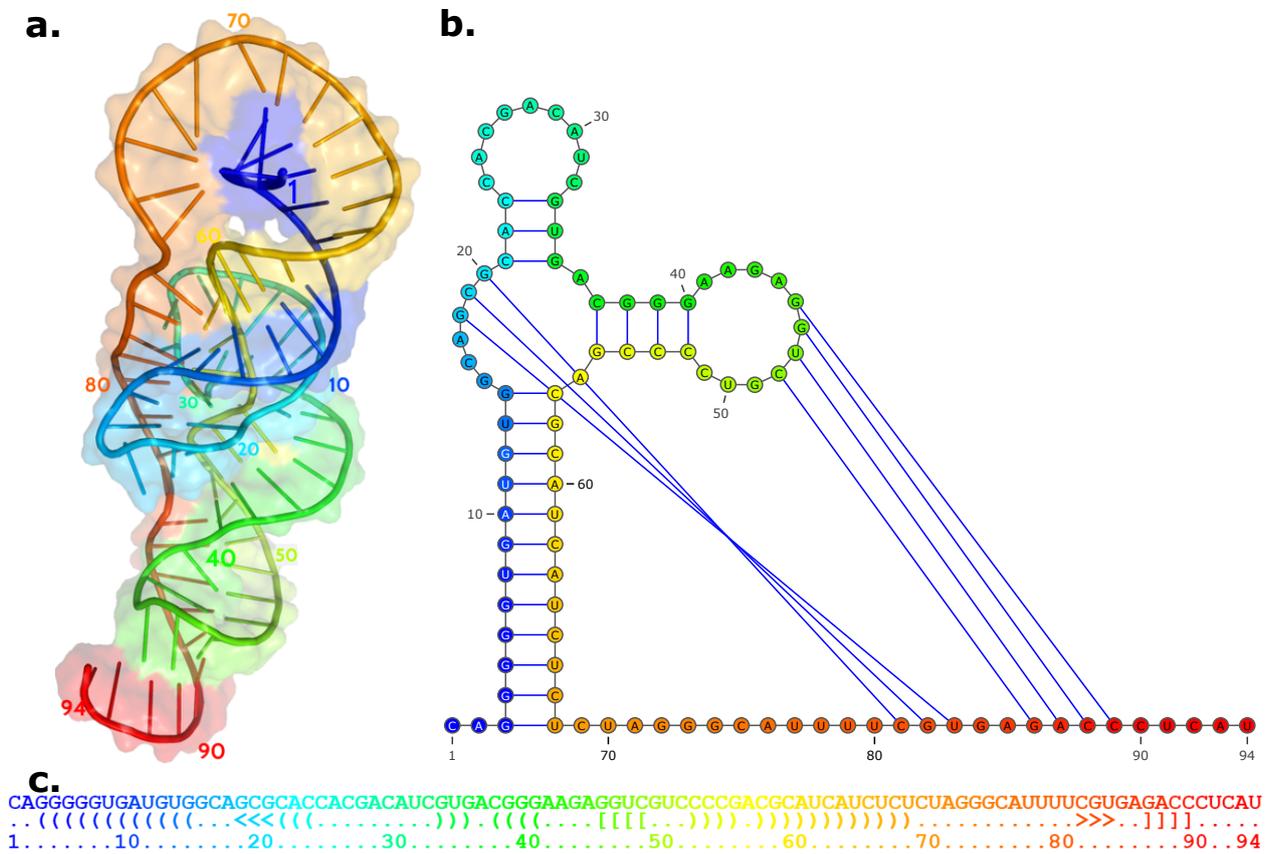
Figure 4.18: **3D structural analysis of DTV xrRNA1.** Panel A. shows the three-dimensional structure of DTV xrRNA1 in cartoon representation, as predicted by simRNA and subsequently energy-minimised using QRNAS. The visualisation uses a gradient rainbow colour scheme, starting with dark blue at the beginning of the sequence and progressing through the spectrum to red at the end of the sequence, to correlate nucleotide positions. A semi-transparent surface rendering, created using PyMOL, is superimposed on the cartoon model to provide a contour perspective of the molecular surface. The corresponding secondary structure is shown in Panel B., following the same colour scheme as the 3D structure, to facilitate correlation between the two-dimensional layout and the three-dimensional conformation. Panel C. displays the sequence of DTV xrRNA1 in the dot-bracket notation, with the secondary structure elements highlighted. The nucleotide sequence adheres to the same colour coding, ensuring a consistent visual reference throughout all panels and highlighting the continuity from sequence to structure.

a more varied range, with xrRNA1 having a smaller stem consisting of 3 base pairs and xrRNA2 having a median of 5 base pairs.

The β loop shows a similar pattern as the β stem, with xrRNA1 β loop being shorter by 4 nucleotides and xrRNA2 being longer with a median of 7 nucleotides, but both having a high variance. A behavioural contrast is observed in the γ elements. xrRNA1 has a shorter γ stem but compensates with an extended γ loop relative to xrRNA2. Both xrRNA1 and xrRNA2 show low variance in the γ stem and loop, which could suggest structural importance. External M shows the same compensation in the structural length as γ loop to γ stem. The xrRNA1 External M appears to be longer than xrRNA2, but xrRNA2 has a higher variance.
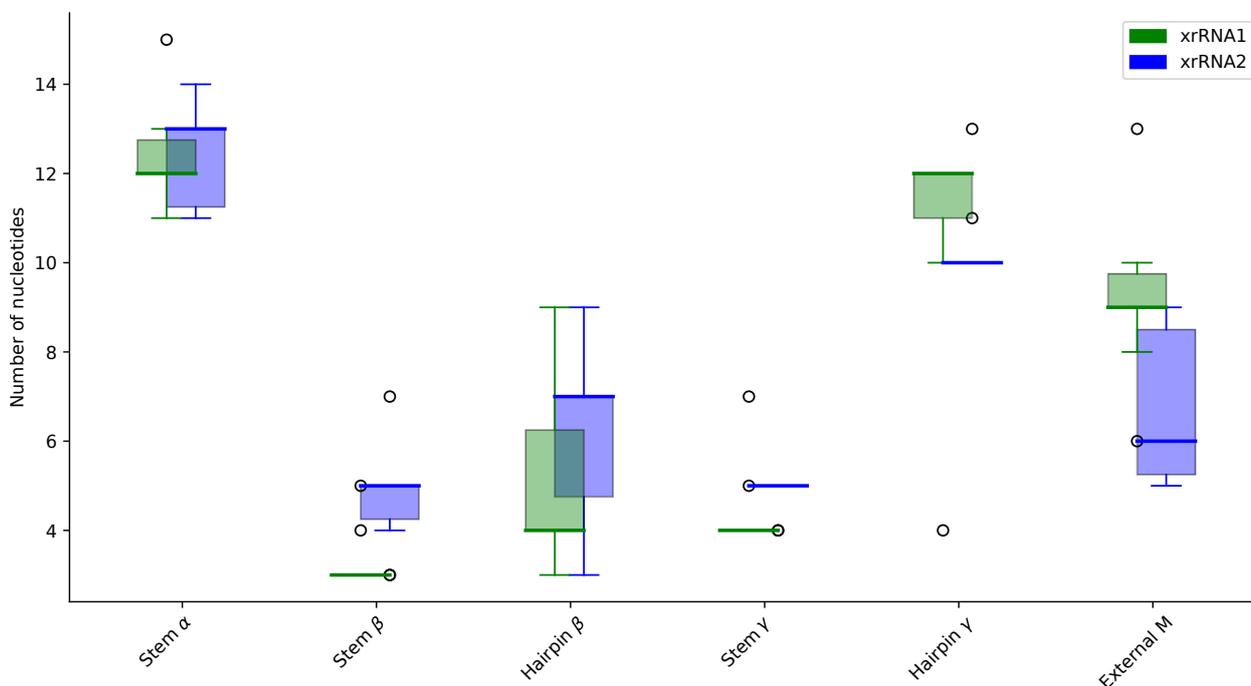

Figure 4.19: **Structural Element Length Distribution for TBFV xrRNA1 and xrRNA2.** Boxplots illustrating the distribution of lengths for different structural elements within TBFV xrRNA1 (in green) and xrRNA2 (in blue). The structural elements analyzed in this figure are the α stem, β stem, β loop, γ stem, γ loop, and External M. The thick line in each box represents the median length, the edges of the box are the interquartile range, and outliers are represented as individual points. These distributions provide a visual overview of the range and median sizes of each structural element across the two xrRNA species.

A Welch's t-test was conducted on all features to calculate if there were significant differences between xrRNA1 and xrRNA2, as shown in Figure 4.20. Only External M and stem β had a P-value smaller than 0.05, showing a statistical difference. The low p-value for Stem β and External M gives us a statistically significant difference in the lengths of the elements in xrRNA1 and xrRNA2. Every other value is well above the significance level of 0.05. It may be possible to merge xrRNA1 and xrRNA2 into a single group, as the stem β and External M components do not appear to have a significant effect on the defence mechanism of xrRNA, as observed in the 3D structure in Figure 4.18.
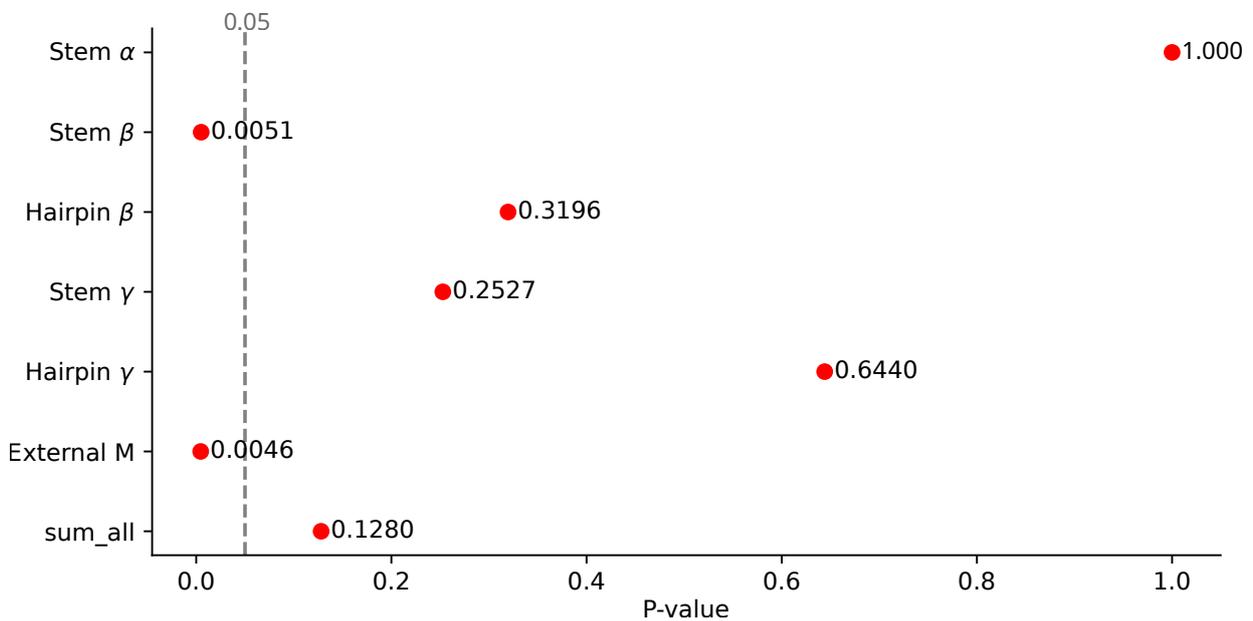

Figure 4.20: **Welchs t-test between length of xrRNA1 and 2 elements of TBFV.** Statistical Analysis of Nucleotide Length Variation in Structural Elements of xrRNA1 and xrRNA2. This figure presents the results of Welch's t-test comparing the number of nucleotides (length) in various structural elements between xrRNA1 and xrRNA2. The p-value measures the probability that an the mean of the two elements are the same. The lower the p-value, the less likely the observed data would be if there were no actual difference.

The observation that certain TBFV viruses, like SREV and TYUV, lack a stem β as seen in Figure 4.21 further consolidates this narrative.

A Link Score analysis was conducted to assess the similarities across all complete genomes of tick-borne flavivirus (TBFV) viruses. The findings, shown in Figure Fig-
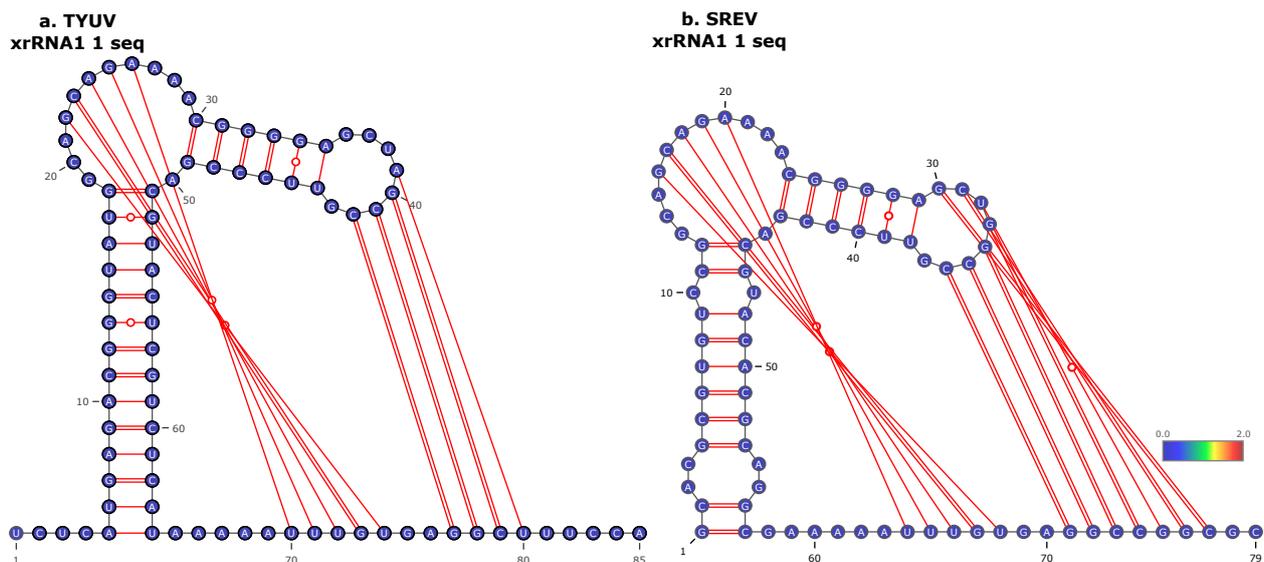
Figure 4.21: **SREV and TYUV secondary structure** prediction out of 1 sequence each

ure 4.22, predominantly align xrRNA1 and xrRNA2 within the groups, affirming the expectation of their close relationship. However, the analysis also reveals distinct patterns of similarity and divergence. SREV and TYUV are outliers, clustering closely together but distinctly apart from other viruses. This deviation could be due to the absence of the stem β in their xrRNA structures, which could potentially lower their overall similarity scores. MDFV and XiFV are also identified as outliers, reflecting the unique architectural features of their xrRNAs. ALKV and KFDV show a closer resemblance to the xrRNA2 cluster than to xrRNA1, as does KSIV xrRNA1. This suggests that these viruses may share more in common with the structural or functional aspects of xrRNA2 across TBFV. This trend could suggest a unique evolutionary trajectory. DTV and POWV exhibit a high degree of similarity, which reinforces the idea of conserved structural features within closely related viruses. This similarity likely reflects shared evolutionary pressures and functional requirements that have shaped the xrRNA architecture of these viruses.
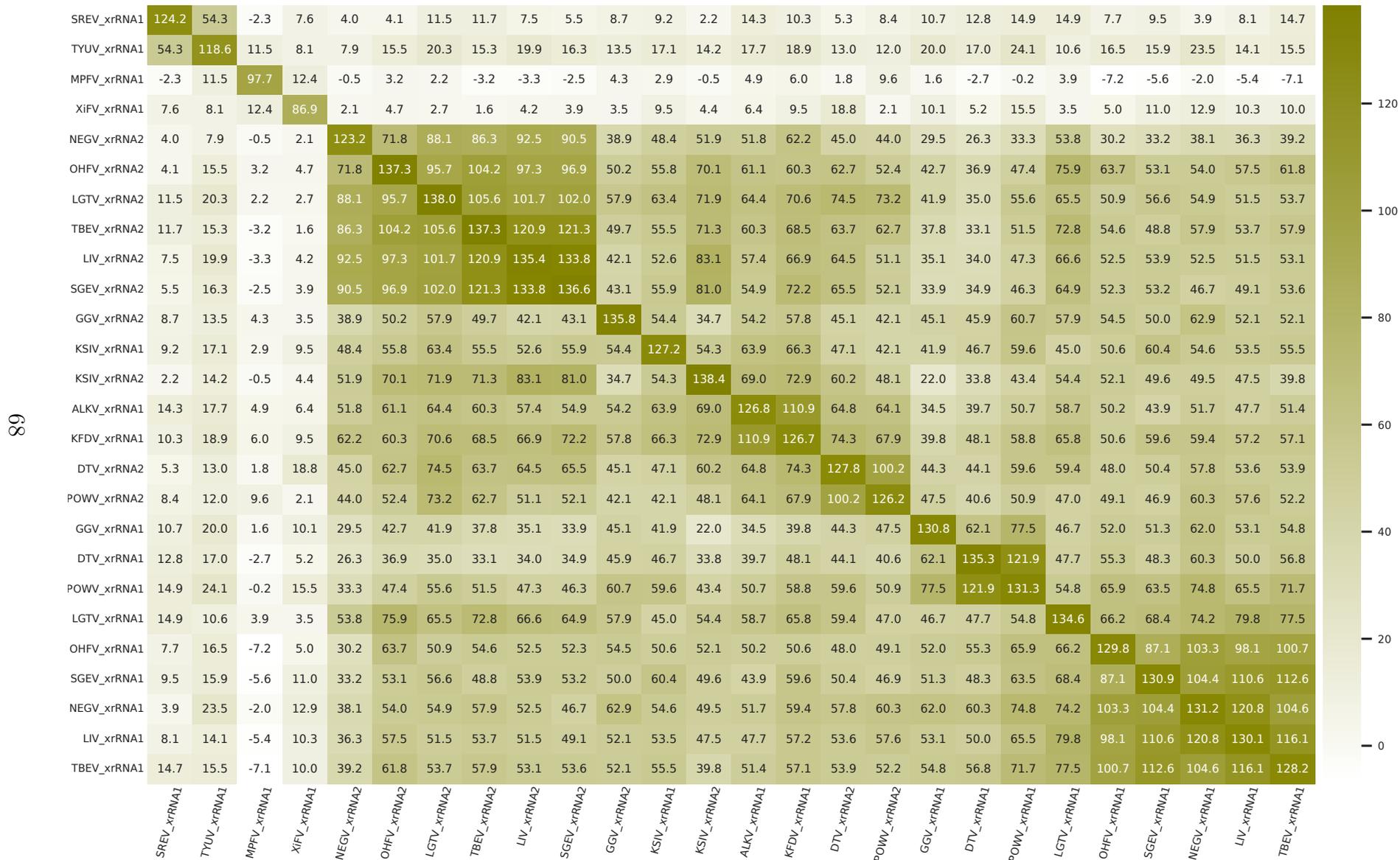
Figure 4.22: **Pairwise Similarity Score between all TBFV xrRNAs CM.** xrRNA similarity scores calculated from pairwise covariance model with CMCompare. Diagonal values show the maximum value that can be achieved. A hierarchical clustering with average linkage was performed, grouping xrRNAs based on the similarity of their covariance model scores.

4.2.4 TBFV xrRNA Design

This subsection establishes the foundation for future design efforts of TBFVs xrRNA. It is based on insights gained from the comparison of TBFV xrRNA1 and xrRNA2. The understanding of the structural complexities and length variations of xrRNA elements across different viruses could be used to develop artificial xrRNA. The design philosophy is based on three blocks: Firstly, the analysis of boxplot data that maps out the lengths of different structural elements within TBFV xrRNA1 and xrRNA2. This quantitative mapping provides a solid framework for modelling artificial xrRNA, supported by empirical data. Secondly, the correlation between the length of the elements. Finally, the sequence conservation in a flaviviral subgroup.

In the context of TBFV xrRNA, a symbolic representation from 20 consensus xrRNA sequences, as seen in Figure 4.23 reveals a conserved multiloop core. The β and γ stems initial base pairs and the two pseudoknots, pk1 and pk2, show a high degree of conservation, defined by a sequence identity exceeding 90% at these nucleotide positions. Artificial xrRNA designs must operate within these structural constraints. The conserved nature of the multiloop core implies limited flexibility for introducing variability in this region. However, potential for innovation is observed in the β loop, external M, and α stem, where the sequence data shows room for exploration and modification.

Designing artificial xrRNA requires an understanding of the interplay between different structural elements. For instance, the γ stem and γ hairpin loop exhibit a negative correlation, suggesting that changes in the length of one directly affect the other, as shown in Figure 4.24. This relationship underscores the importance of the overall length of the γ element in maintaining structural integrity and function. There are also negative correlations between the α stem and external M, as well as between the β stem and external M. These correlations suggest a balance is necessary to maintain the structural accuracy of the xrRNA while incorporating new characteristics. The correlation observed between stem β and external M challenges previous assumptions regarding the non-essential nature of stem β. This correlation suggests that the importance of external M may also be in question. It is possible that external M's role is not as important as we thought, and the correlation with stem β may arise from random variability rather than functional importance.
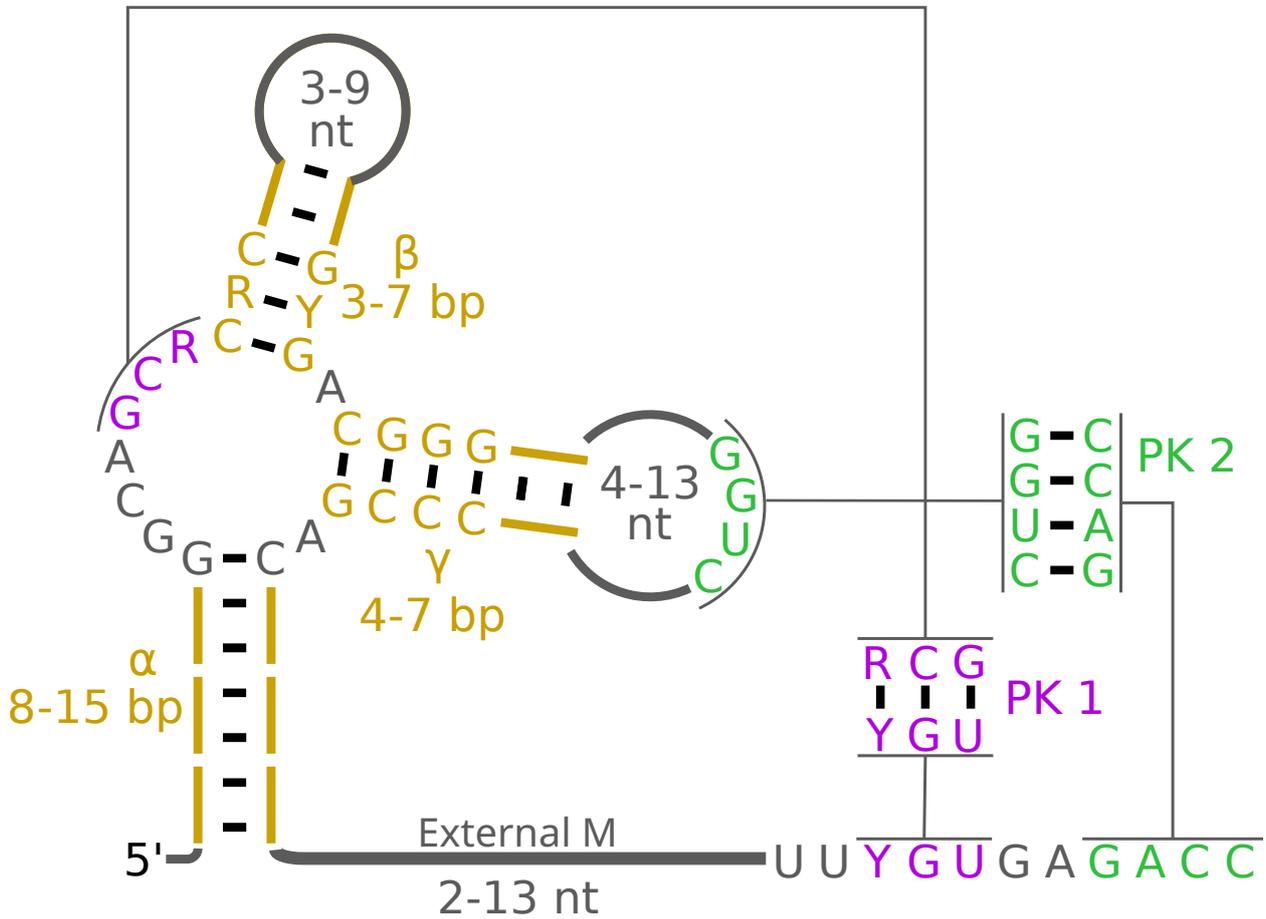
69

Figure 4.23: **Symbolic Secondary Structures of TBFV xrRNA.** The xrRNA is build out of 20 different consensus structures, one xrRNA1 and one xrRNA2 from each virus, calculated from the median elemental lengths from the boxplot analysis. The shown nucleotides represent highly conserved regions with at least 90% sequence identity in the MSA. Each structural element, identified by its functional annotation (α, β, γ, External M and pseudoknot pk1 and pk2), is quantified by the range of lengths (minimum to maximum) in nucleotides (nt) or base pairs (bp). Figure created with R2R by Leonhard Sidl.
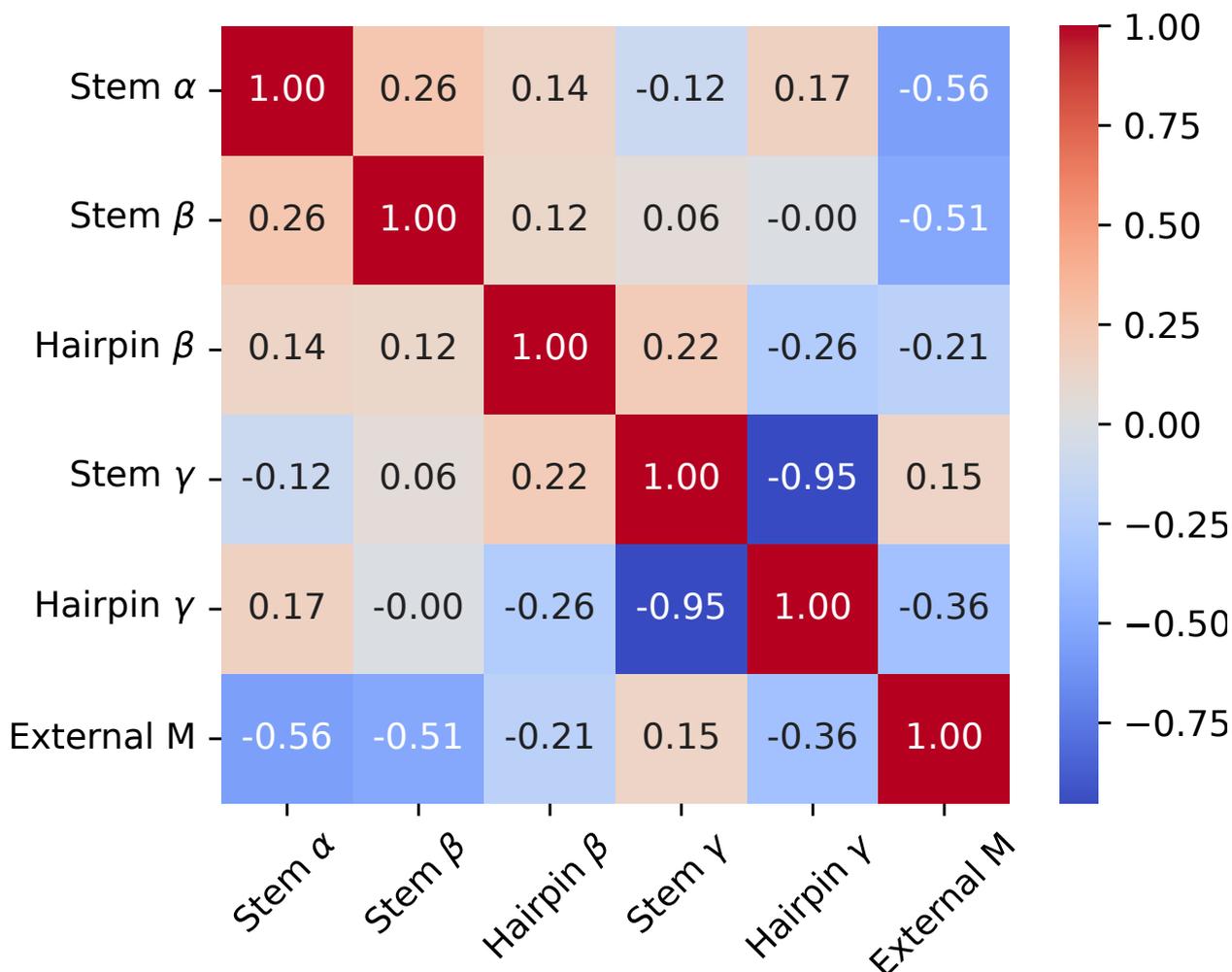
Figure 4.24: **Pearson Correlation Coefficients for Lengths of different xrRNA elements in TBFV** This table presents the calculated Pearson correlation coefficients for the lengths of 20 structural elements of xrRNA1 and xrRNA2, with two representatives from each virus. The coefficient ranges from -1 to 1 and provides insights into the linear relationship between the structural elements of the xrRNA. A positive number in the correlation matrix shows us an increase in the length of one structural element with an increase in the length of another, while a negative number shows us an inverse relationship, where the lengthening of one element gives us the shortening of another.

4.3   ISFV xrRNA Structure Analysis

4.3.1   dISFV xrRNA 2D Structure

In the field of ISFVs, dual-host ISFV exhibit a xrRNA structure that is similar to the before-described xrRNAs found in MBFV. The xrRNA of dISFVs retains the essential three-way junction that forms the core of the structure, as observed in panel A of Figure 4.25. The conserved structure is accompanied by α and γ stems that exhibit a high degree of sequence conservation, hinting at their critical structural and potentially functional roles within the xrRNA.

The dISFV xrRNA preserves the pseudoknot formations, pk1 and pk2. Co-variation in pk2 suggests its structural importance and evolutionary conservation. Additionally, the β loop, similar to its MBFV counterparts, exhibits a high degree of variability, suggesting a flexible region that may not be necessary but can accommodate variations without compromising the overall structural integrity. The function and stability of the dISFV xrRNA may depend on the conserved unpaired nucleotides, specifically a cytosine at position 23 and a uracil at position 3. These features are also present in the MBFV xrRNA structures, suggesting that these xrRNA elements are likely homologous across different flavivirus groups. The importance of the three-way junction, the conserved stems α and γ, and the roles of the pseudoknots in maintaining the structural integrity necessary for the xrRNA could be underscored by this homology.

4.3.2   cISFV xrRNA 2D and 3D Structure

During the investigation of classic insect-specific flaviviruses (cISFVs), which include Cell fusing agent virus (CFAV), Aedes flavivirus (AEFV), and Kamiti River virus (KRV), the consensus structure of xrRNA shows a familiar three-way junction similar to that of MBFV and dISFV xrRNAs. The structure includes important elements such as pk1 at the beginning, followed by stem α, stem β with a loop, and stem γ with a loop. The structure extends to where pk2 pairs towards the end of the sequence, as shown in Figure 4.26.

Upon thorough examination, distinct features are evident in the cISFV xrRNA structure when compared to its MBFV or dISFV counterparts. Specifically, pk1 ex-

BP conservation

bp conserved in msa

bp not possible in >0 seq

Entropy
0.0   2.0

Covariation
n 1 bp
n 2 bp
n 3 bp

b.

```
                <<.(((((((.............]].((((((...[[[[[[.[.]]]]]]>>]]]]]].]]]]]]..
EPEV_NC027999.1  agucaggdccaucggaaga--ugccaccggauaguguuagacggugcugccugcgaaca----
MMV_MF139576.1   agucaggccucaagacauug-agccaccggaugggacuagacggugcugccugcgcguccgcc
DONV_MZ357898.1  agucaggacccaaugugaccaccggaugcaacgcgccaccggaugauagugagacggugcugccugcgcguccgcc
CHAOV_JQ308185.1 agucaggccuaaa-------ugccaccggaugauagugagacggugcugccugcagcuaucac
PANFV_KY072986.1 agucaggaccgaucuuagau-agccaccggauga-gguagacggugcugucugcgaacc-----
LAMV_FJ606789.2  agucaggccuaaa-------ugccaccggaugauaggcuagacggugcugccugcagcuauc--
GUAV_MK908099.1  agucaggdccgcuug------uccaccggaugg-guuagacggugcugccugcgaacc-----
MIS consensus    AGUCAGGCCKMMMgaagaugHGCCACCGGAUGRkRBUAGACGGUGCUGCCUGCRVVYmyccc
                 .........10........20........30........40........50........60.
```
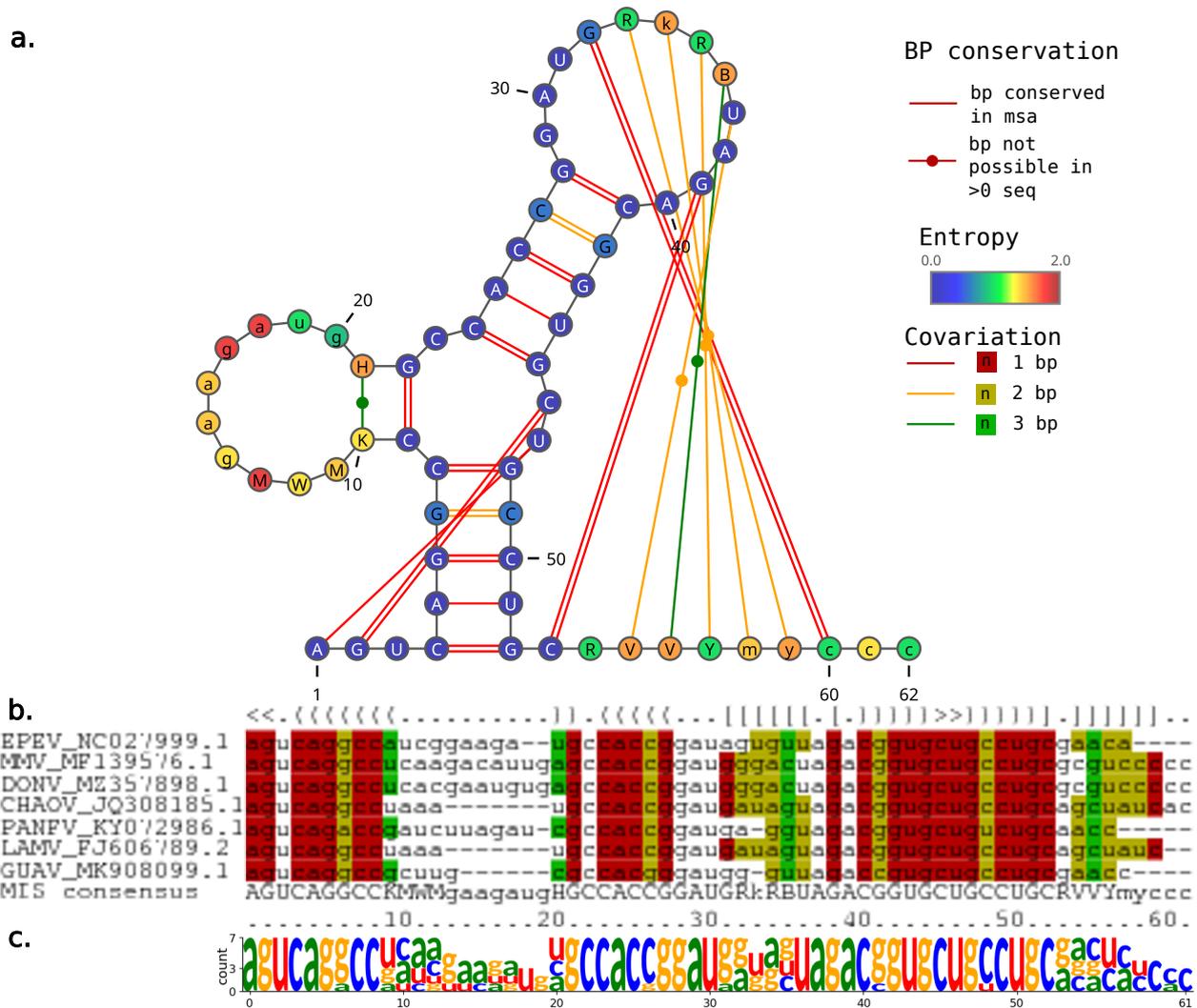
c.

Figure 4.25: **Analysis of dISFV xrRNA1 Secondary Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA1 across 7 dISFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. shows the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA1 sequence. The stack heights are proportional to the nucleotide frequency at each position.
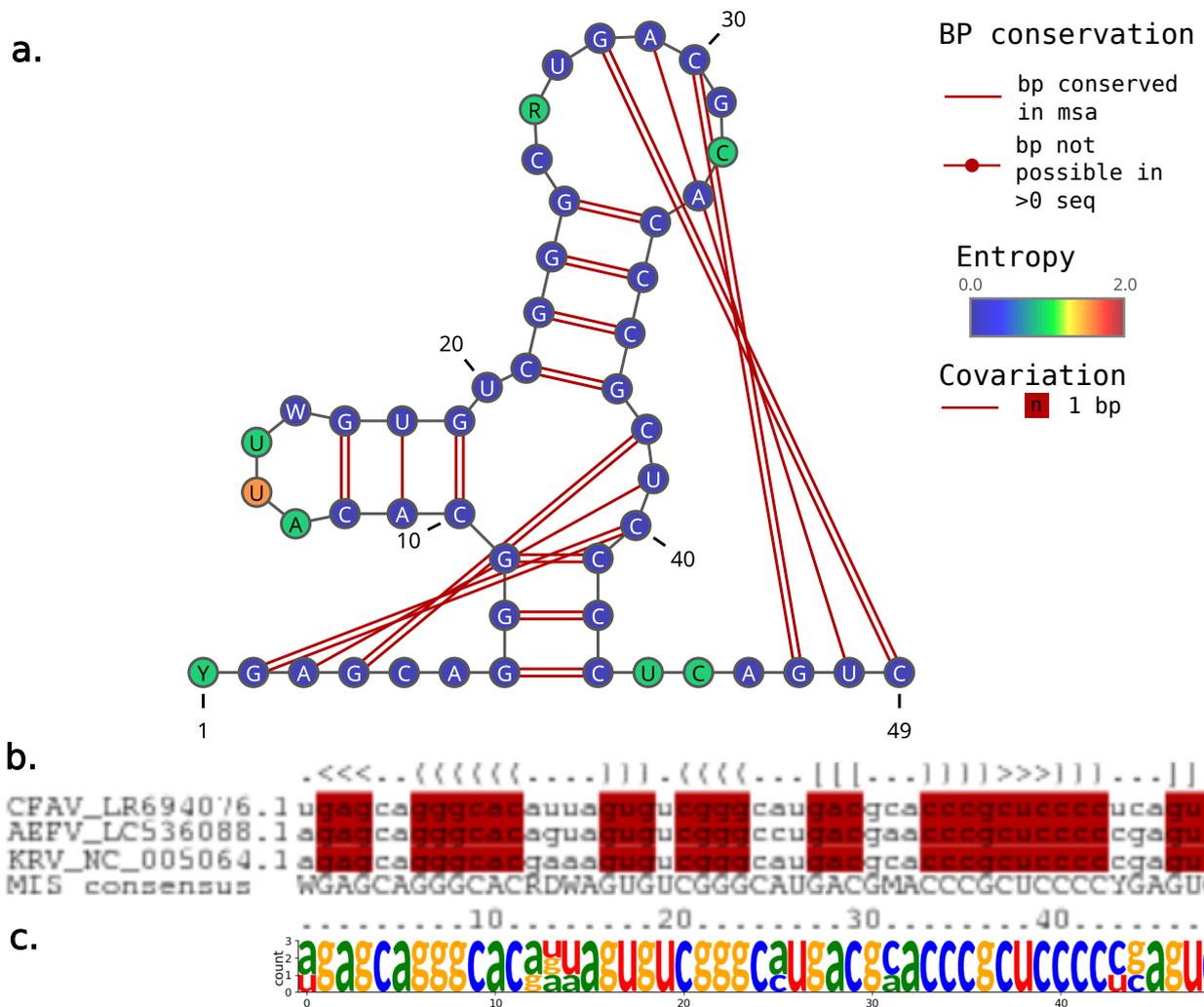
Figure 4.26: **Analysis of cISFV xrRNA1 Secondary Consensus Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA1 across 3 cISFV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA1 sequence. The stack heights are proportional to the nucleotide frequency at each position.

tends an additional base pair beyond what is typically observed in MBFV xrRNAs. In addition the unpaired region within the multiloop, where pk1 pairs, consists of three nucleotides. Additionally, the nucleotide conserved in the multiloop between stem β and γ is uracil. Prior to stem α, cytosine and adenine nucleotides are located at positions 5 and 6, respectively, suggesting the possibility of base pairing, as observed in MBFV xrRNAs. However, stem α differs from the typical MBFV structure, as it is shorter and consists of only three base pairs, each containing conserved guanine and cytosine pairs. Stem β retains three conserved base pairs, but the β hairpin loop shows regions of higher Shannon entropy, suggesting potential for functional adaptation across different host environments. Similarly, the areas preceding pk2 in the γ loop and adjacent to stem α exhibit variable regions, showcasing functional importance that may contribute to the xrRNA's adaptability and functionality. Exercise caution when interpreting the results of Shannon entropy due to the limited sample size of only three sequences used in this analysis, as no additional sequences were available at the time of writing.

Figure 4.27 panel A shows the 3D structure of Aedes Flavivirus (AEFV) xrRNA1, which was predicted by simRNA and refined through energy minimization with QRNAS. The visualization highlights a ring-like structure surrounding the sequence's 5' start. The ring is coloured dark blue at the start and gradually transitions through the colour spectrum to red at the opposite end. The AEFV xrRNA1 structure features pk1 at its core, which is believed to be essential for maintaining the molecule's three-dimensional architecture. The ring-like formation surrounding the core is composed of the second base pairs of stem γ and stem α, as well as pk1, creating a protective barrier for the core structure against XRN1.

Pseudoknot 2 (pk2) may serve as a vital structural element that anchors at the sequence's terminal end at position 44-48. This positioning suggests that pk2 consolidates the xrRNA structure, possibly by acting as a mechanical counterbalance or providing tension relief when confronted with the degradative pull of XRN1. The β stem and its loop appear to be detached from the central knot-like formation, adopting an isolated orientation. This observation shows that the β stem and loop may not have a role in the direct resistance mechanism against XRN1, unlike the core structural components.

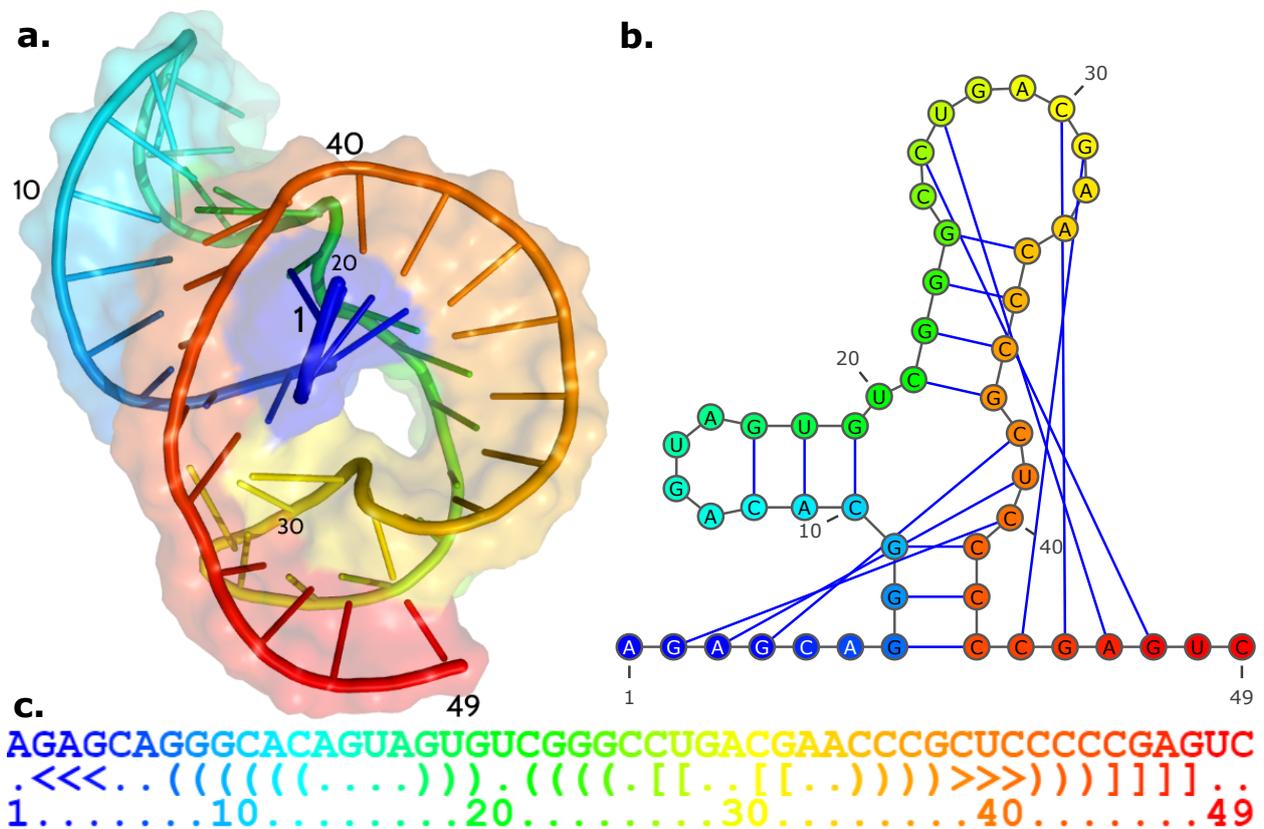Panel B in Figure 4.27 is aligned with the 3D visualization, displaying the corre-

Figure 4.27: **3D and 2D sSructural Analysis of AEFV xrRNA1.** Panel A. shows the three-dimensional structure in cartoon representation of AEFV xrRNA1 as predicted by simRNA and subsequently energy minimised using QRNAS. The visualisation uses a gradient rainbow colour scheme, starting with dark blue at the beginning of the sequence and progressing through the spectrum to red at the end of the sequence, to correlate to nucleotide positions. Superimposed on this cartoon model is a semi-transparent surface rendering that provides a contour perspective of the molecular surface. This rendering was done with PyMOL. Panel B. shows the corresponding secondary structure, following the same colour scheme as the 3D structure, facilitating correlation between the two-dimensional layout and the three-dimensional conformation. Panel C. shows the sequence of AEFV xrRNA1 in the dot-bracket notation, with the secondary structure elements highlighted. The nucleotide sequence follows the same colour coding, providing a consistent visual reference across all panels and emphasising the continuity from sequence to structure.

sponding secondary structure. A consistent color scheme is maintained for reference across dimensions, allowing for an intuitive correlation between the two-dimensional layout and the three-dimensional conformation. Panel C provides additional details on the sequence of AEFV xrRNA1 in dot-bracket notation, highlighting secondary structure elements.

4.4   NKV xrRNA Structure Analysis

During the exploration of no-known-vector (NKV) flaviviruses, such as Rio Bravo virus (RBV), Modoc virus (MODV), and Montana myotis leukoencephalitis virus (MMLV), it was observed that the consensus xrRNA structure slightly resembles that of TBFVs, rather than MBFVs or ISFVs. The xrRNA structure comprises a three-way junction at its core, surrounded by key elements such as stem α, stem β with a loop, and stem γ with a loop. Additionally, there is a multiloop consisting of six unpaired nucleotides. As shown in Figure 4.28, the xrRNA contains pk1, which spans from three nucleotides within the multiloop to the middle of the sequence's terminal unpaired region. This arrangement is similar to that found in TBFV xrRNAs. Subsequently, pk2 emerges from the γ loop and extends to the end of the sequence, reinforcing the structural and potentially functional similarities with TBFV xrRNAs. The multiloop between stem β and γ contains a conserved adenine nucleotide, which also exhibits structural similarities with TBFV xrRNAs. However, the region between stem γ and α deviates from this pattern by lacking the unpaired and conserved adenine, which is a distinctive feature in the NKV xrRNA structure.

A potential for hairpin formation in the individual sequences of the analyzed viruses is shown in the unpaired region following stem α, specifically at positions 55 to 67. However, this feature is not preserved in the consensus structure.

Figure 4.29 shows the 3D structure of Rio Bravo virus (RBV) xrRNA1, highlighting a yellow ring-like structure that encircles the sequence's 5' start. The progression of nucleotide positions is marked by a colour transition from dark blue to red along the sequence length.

PK1 and PK2 may play a major role in anchoring at the terminal end of the sequence, contributing to the structural consolidation of the xrRNA. This could
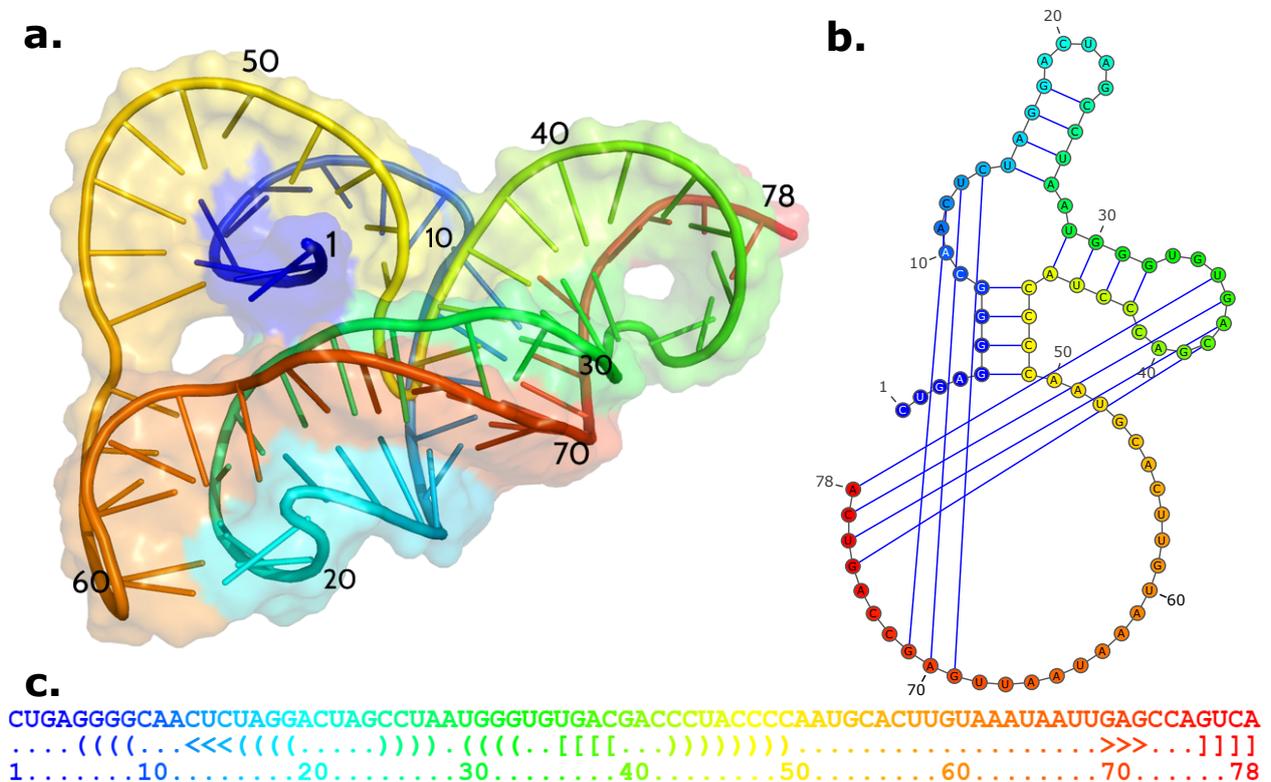
77

Figure 4.28: **Analysis of NKV xrRNA1 Secondary Consensus Structure, Sequence Alignment, and Nucleotide Conservation.** Panel A. shows the consensus secondary structure of xrRNA1 across 3 NKV viruses, with nucleotides and basepairs colored based on positional Shannon Entropy and covariation. Panel B. is the MSA of these xrRNA1 sequences, highlighting covariation and the most informative sequence per IUPAC nomenclature. Panel C. displays a sequence logo that quantitatively presents the conservation and variability of nucleotides across the consensus xrRNA1 sequence. The stack heights are proportional to the nucleotide frequency at each position.

a.

b.

c.

CUGAGGGGCAACUCUAGGACUAGCCUAAUGGGGUGUGACGACCCUACCCCAAUGCACUUGUAAAUAAUUGAGCCAGUCA
....((((...<<<((((.....))))).((((..[[[[...))))))))...................>>>...]]]]
1.......10........20........30........40........50........60........70......78

Figure 4.29: **3D and 2D Structural Analysis of RBV xrRNA1.** Panel A. shows the three-dimensional structure in cartoon representation of RBV xrRNA1 as predicted by simRNA and subsequently energy minimised using QRNAS. The visualisation uses a gradient rainbow colour scheme, starting with dark blue at the beginning of the sequence and progressing through the spectrum to red at the end of the sequence, to correlate to nucleotide positions. Superimposed on this cartoon model is a semi-transparent surface rendering that provides a contour perspective of the molecular surface. This rendering was done with PyMOL. Panel B. shows the corresponding secondary structure, following the same colour scheme as the 3D structure, facilitating correlation between the two-dimensional layout and the three-dimensional conformation. Panel C. shows the sequence of RBV xrRNA1 in the dot-bracket notation, with the secondary structure elements highlighted. The nucleotide sequence follows the same colour coding, providing a consistent visual reference across all panels and emphasising the continuity from sequence to structure.

offer mechanical stability or aid in resistance against the XRN1. In contrast to other xrRNA seen in MBFV, ISFV or TBFV, the ring and core structure in RBV xrRNA1 suggests a different spatial arrangement, with the 'knot' feature not being as centralized.

Panel B of the figure presents the secondary structure of RBV xrRNA1 using the established color scheme to ensure coherence across visual representations. This alignment enables a direct comparison and understanding of the structural layout in two dimensions versus the three-dimensional conformation. Panel C shows the sequence details of RBV xrRNA1 using dot-bracket notation to highlight secondary structure elements. The colour coding in RBV xrRNA1 provides a clear guide for navigating its structure, reinforcing the continuity between its sequence and visual representation.

Chapter 5

Discussion

5.1   Research Objectives

This research's objective was to conduct a computational analysis of xrRNA structures within the genus Flavivirus. The investigation focused on the distinct subgroups: MBFV, TBFV, ISFV and NKV. This Master Thesis aimed to show similarities and differences in and across these subgroups by trying to understand the nuances of xrRNA structures and sequences, in 2D but also in 3D. We wanted to understand whether the xrRNA structure within a virus, referred to based on their position in the genome, as xrRNA1 and xrRNA2, should be treated as separate models or merged into a unified consensus structure. This distinction should be important for accurately performing statistical analysis of xrRNAs structural elements to create guidelines for generating artificial xrRNA.

5.2   xrRNA 2D Structure

During my research into the 2D structure of xrRNA in the genus flavivirus, I predicted successfully the consensus structure of MBFV, TBFV, ISFV and NKV. It shows that the presence of a typical three-way junction, a structural motif consistently highlighted in previous studies cited. The core of the three-way junction is defined by the α, β and γ stems across all flavivirus subgroups examined. This arrangement highlights the evolutionary conservation of xrRNA structure within the

genus, despite the different host ranges and transmission vectors of each subgroup. Notably, a multi-loop region at the core of this junction shares at least one conserved nucleotide within each flavivirus subgroup, suggesting some functional importance. The core of the multiloop and bps close to it from stem α and stem γ appear to be sequence conserved or show high covariation suggesting structural importance. My research shows the identification of a short pk1 extending from the multiloop towards the beginning of the sequence in MBFV and ISFV, and towards the middle end of the sequence in TBFV and NKV. This positional variability suggests subgroup-specific adaptations or functionalities that are not yet fully understood. In addition, a long pk2 extending from the γ loop to the end of the sequence is consistently observed in all flavivirus subgroups. The omnipresence of this long pk2 in different flaviviruses highlights its likely important role in the structural stability of the xrRNA.

An intriguing aspect of our analysis focuses on the characterisation of stem β and its adjacent loop β within the consensus xrRNA structures of all flaviviral subgroups. These regions exhibited high sequence variability.

The observed variability in stem β and loop β could on the one hand mean that such diversity might suggest that these elements do not play a structurally conserved role in the fundamental mechanism of xrRNA resistance to XRN1. If their primary function were central to xrRNA's ability to evade host defences, one might expect to find greater sequence conservation, similar to the conserved nucleotide observed in the multi-loop region across flavivirus subgroups. On the other hand, the variability observed in stem β and loop β may hint at an evolutionary adaptation mechanism that allows flaviviruses to fine-tune their interactions with a wide range of host organisms. This flexibility may be vital to the ability of the virus to navigate and exploit the host cellular environment for replication and survival. The divergence in sequence may reflect evolutionary pressures exerted by different host immune systems, suggesting that while the core function of xrRNA remains constant, the specifics of its interaction with host factors may be finely tuned to optimise viral persistence and transmission. This hypothesis is further supported by consideration of the wider context of viral evolution, where sequence variability is often correlated with functional adaptability. In this light, stem β and loop β may serve as variable regions that allow flaviviruses to maintain a balance between

the structural integrity required for the protective role of xrRNA and the flexibility required for host adaptation.

The study of xrRNA structures within TBFV has revealed an additional layer of complexity in our understanding of the β elements. Within the TBFV subgroup, two specific structures, SREV and TYUV, stand out for their complete absence of stem β. This absence is significant because it differs from the consensus structure observed in other flaviviruses. The absence of stem β in SREV and TYUV presents a unique case for the structural diversity within flavivirus xrRNA. This suggests that the xrRNA mechanism for resisting XRN1 can function effectively without this structural element in certain flavivirus subgroups. This discovery challenges the idea that there is a universally conserved structural blueprint for xrRNA in all flaviviruses and highlights the evolutionary adaptability of these viral RNA molecules. From an evolutionary perspective, the lack of stem β in these TBFV xrRNAs may suggest an adaptation to the distinct selective pressures faced by these viruses. This adaptation could be due to the specific host immune environments that SREV and TYUV have evolved to exploit, requiring a deviation from the structural norms observed in other flaviviruses. Alternatively, this could be a case of convergent evolution, where different evolutionary paths have led to a similar functional outcome. This highlights the adaptability and resilience of the flavivirus genome.

Predicting accurate consensus structures for xrRNA has been challenging due to limitations in the computational tools used like RNAalifold, which do not account for pseudoknots. Pseudoknots appear to be an important element in the overarching structure of xrRNA. Advanced predicting techniques are necessary to incorporate tertiary base pair interactions such as pseudoknots and tertiary base pairs, to gain better insights into xrRNA architecture. Our findings confirm the classification proposed by [69], which dictates two primary categories of xrRNA structures within the Flavivirus genus: one shared between MBFV and ISFV, and the other common to TBFV and NKV.

5.3   xrRNA 3D Structure

During our investigation into the 3D structures of xrRNA across the four flavivirus (FV) subgroups, we observed a consistent presence of a ring-like structure.

This structure is similar to the one reported in MBFVs ZIKV, in the study by [56]. The given structure, which threads the unpaired beginning of the sequence through a loop, acts as a mechanical barrier against XRN1. This supports the hypothesis proposed by [53] regarding its functional significance in viral RNA stability and immune evasion.

A clear distinction was observed in the core components of the ring-like structure between the subgroups. In the case of MBFV and ISFV, the pk1 motif is central within the knot, whereas in TBFV and NKV, the stem α assumes this core role. This variation suggests that the structural fundamentals of the ring-like structure are conserved across flaviviruses, but with differences in the specific elements contributing to the core. A possible solution to reconcile these structural differences could be to redefine the nomenclature of stem α and pk1 in TBFV and NKV to match that of MBFV and ISFV. This would emphasize the structural and functional homology among these subgroups, as proposed in [70].

The use of Monte Carlo simulations through simRNA offers a reliable method for predicting the 3D structure of xrRNA at a macro level, especially in visualising the ring-like structure. However, the complex topology of xrRNA molecules, the importance of a magensium ion as seen in [55], combined with the inherent nature of Monte Carlo simulations, introduces uncertainty regarding the detailing of tertiary interactions. Although we have identified the overarching structure with confidence, it is still difficult to pinpoint specific interactions, such as base triples. In the well studied MBFVs ZIKV xrRNA, the presence of two base triples has been documented, as highlighted in the work of [71] and [51]. Although these structural features are known, their consistent reconstruction using simRNA does not always workout. Sometimes we are able to accurately predict the presence of these known base triples, while at other times we encounter alternative, previously unidentified, unconsistent base triples. Given this variability in the prediction of base triples in MBFV xrRNA structures with simRNA, it becomes less useable to extend this approach to the identification of base triples in the xrRNA structures of TBFV, ISFV and NKV. Despite this methodological limitation, there remains a theoretical basis for expecting the presence of base triples within these other flavivirus subgroups. This expectation is based on the observation that numerous sequence elements and structural configurations, involve those conserved unpaired nucleotides, that are

conserved across these subgroups. Such conservation suggests the potential for base triples to contribute to the structural integrity and functional mechanisms of xrRNA across different Flavivirus subgroups.

5.4   xrRNA Positional Divergence

The structural length differences between xrRNA1 and xrRNA2 in MBFVs suggest that they should be treated as separate entities. Statistical analysis show significant differences in the length of the structural elements, supporting a bifurcated modeling approach to more accurately asses their structural nuances. Work done by Wolfinger2021 [49] challenges the grouping of MBFV xrRNA based on their positional order, which show outliers like the DENV4 xrRNA1, which is more closely related to the xrRNA2 than xrRNA1, also shown in the Figure 4.11 comparing the CM link scores. Also in contrast, the data for TBFV does not strongly support a separation into xrRNA1 and xrRNA2, leading to a unified analysis for this subgroup. The complexity and variability of xrRNA structures within Flaviviruses are highlighted by this opposite behaviour. Therefore a tailored computational approach is necessary for each subgroup. In addition to the xrRNA1 and xrRNA2 categories, the CMCompare analyses revealed outlier xrRNA structures in MBFV and TBFV. This discovery indicates that xrRNA has a wider range of structural diversity than previously believed. Further investigation is needed to find out if there are new structural categories or if the xrRNA mechanism is not as strict on structural lengths and only requires a ring-like structure, which would make this positional classification obsolete.

5.5   Artificial xrRNA Design

My work has taken the first steps towards creating artificial xrRNA by demonstrating the feasibility of modifying xrRNA structures while retaining their characteristic functionalities. By altering nucleotides within the pk2 region of the AROAV xrRNA, a MBFV, it is possible to adjust its duplex energy levels without disrupting the xrRNA's ability to fold into its desired 2D structure. These modifications

preserve the formation of the 3D ring-like structure of xrRNA, underscoring the robustness of the xrRNA architectural blueprint.

To advance artificial xrRNA design, we will use the symbolic structures identified within MBFV (Figure 4.12) and TBFV (Figure 4.23) as templates for generating synthetic xrRNA molecules. These structures have highlighted sequence-conserved nucleotides that are vital components in the artificial design process. Incorporating these conserved elements ensures that the engineered xrRNA maintains the essential functionalities and interactions that are characteristic of its natural counterparts.

To improve the modification of structural lengths, it is pragmatic to adopt a consensus structure from a representative virus as a blueprint for the initial design iteration. This strategy allows for the creation of an artificial xrRNA based on proven structural parameters, thereby increasing the likelihood of successful folding and functional performance. Subsequent iterations can experiment with variations in structural lengths, guided by the observed ranges (minimum, maximum, and median) within natural xrRNA molecules.

To refine the design process further, it is recommended to develop an analytical framework to understand the interplay between different structural elements. A functional model that predicts optimal configurations for artificial xrRNA can be calculated by using the correlation between the lengths of the various structural components. This model would enable the design of xrRNA structures with novel length combinations, potentially unveiling new functionalities or enhancing existing ones.

Advancements in artificial xrRNA design show great promise for a range of applications. By mimicking and manipulating xrRNA structures, scientists can gain insights into viral immune evasion mechanisms and explore new therapeutic inventions. The journey towards mastering artificial xrRNA design is in its early stages, and the potential for discovery and innovation in this field is vast and largely untapped. As we continue to unravel the complexities of xrRNA architecture and function, the possibilities for leveraging these insights in virology and beyond continue to expand.

Chapter 6

Conclusion

This thesis presents an overview of the xrRNA structure across flaviviruses. The complex interactions of nucleotides in the xrRNA are demonstrated through 2D and 3D structural prediction. The two pseudoknots, along with stem α and γ build the knot-like structure around the beginning of the structure and are conserved across all flaviviruses. It remains to be shown in future studies if stem β is integral to the structure's solidity. Sequence alignment, structure prediction, and covariation were integrated into the analysis to identify and confirm key structural motifs and interactions in the xrRNA. This study lays the groundwork for future investigations of the xrRNA. The detailed understanding of the xrRNA structure could provide a guideline for the design of artificial xrRNA with potential for therapeutic applications. In the future, we should be able to pre-test our artificial xrRNA in molecular dynamics to test if the mechanical anisotropy, and therefore resistance to XRN1, is still present before conducting a costly and time-consuming enzyme assay in the wet lab.

Bibliography

[1] Xunxun Wang, Shixiong Yu, En Lou, Ya-Lan Tan, and Zhi-Jie Tan. Rna 3d structure prediction: Progress and perspective. *Molecules*, 28(14):5532, July 2023.

[2] Mirela Andronescu, Vera Bereg, Holger H Hoos, and Anne Condon. Rna strand: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9(1), August 2008.

[3] Christopher Barrett, Fenix W Huang, and Christian M Reidys. Sequence–structure relations of biopolymers. *Bioinformatics*, 33(3):382–389, September 2016.

[4] Amal S. Abu Almakarem, Anton I. Petrov, Jesse Stombaugh, Craig L. Zirbel, and Neocles B. Leontis. Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Research*, 40(4):1407–1423, November 2011.

[5] David W Staple and Samuel E Butcher. Pseudoknots: RNA structures with diverse functions. *PLoS Biology*, 3(6):e213, June 2005.

[6] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte fuer Chemie Chemical Monthly*, 125(2):167–188, February 1994.

[7] R. Giegerich. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843–4851, September 2004.

[8] Serge Léger, Maria Beatriz Walter Costa, and Dan Tulpan. Pairwise visual comparison of small RNA secondary structures with base pair probabilities. *BMC Bioinformatics*, 20(1), May 2019.

[9] S Rauscher, C Flamm, C W Mandl, F X Heinz, and P F Stadler. Secondary structure of the 3'-noncoding region of flavivirus genomes: comparative analysis of base pairing probabilities. *RNA*, 3:779–791, 1997.

[10] Monir Hajiaghayi, Anne Condon, and Holger H Hoos. Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, 13(1), February 2012.

[11] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1), November 2011.

[12] S M Freier, R Kierzek, J A Jaeger, N Sugimoto, M H Caruthers, T Neilson, and D H Turner. Improved free-energy parameters for predictions of RNAduplex stability. *Proceedings of the National Academy of Sciences*, 83(24):9373–9377, December 1986.

[13] David H. Mathews, Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, 101(19):7287–7292, May 2004.

[14] Mirela Andronescu, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–i28, July 2007.

[15] Stephan H Bernhart, Ivo L Hofacker, Sebastian Will, Andreas R Gruber, and Peter F Stadler. Rnaalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9(1), November 2008.

[16] Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment mod-

eling: methods and applications. *Briefings in Bioinformatics*, 17(6):1009–1023, November 2015.

[17] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, September 2007.

[18] Sebastian Will, Tejal Joshi, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Locarna-p: Accurate boundary prediction and improved detection of structural rnas. *RNA*, 18(5):900–914, March 2012.

[19] Dmitri J. Gritsun, Ian M. Jones, Ernest A. Gould, and Tamara S. Gritsun. Molecular archaeology of flaviviridae untranslated regions: Duplicated RNA structures in the replication enhancer of flaviviruses and pestiviruses emerged via convergent evolution. *PLoS ONE*, 9(3):e92056, March 2014.

[20] Thomas D. Schneider and R.Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, 1990.

[21] Sean R. Eddy and Richard Durbin. Rna sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, 1994.

[22] Zizhen Yao, Zasha Weinberg, and Walter L. Ruzzo. Cmfinder—a covariance model based RNAmotif finding algorithm. *Bioinformatics*, 22(4):445–452, December 2005.

[23] Andrew D. Johnson. An extended iupac nomenclature code for polymorphic nucleic acids. *Bioinformatics*, 26(10):1386–1389, March 2010.

[24] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, July 1948.

[25] Marco Pietrosanto, Marta Adinolfi, Andrea Guarracino, Fabrizio Ferrè, Gabriele Ausiello, Ilio Vitale, and Manuela Helmer-Citterich. Relative information gain: Shannon entropy-based measure of the relative structural conservation in RNAalignments. *NAR Genomics and Bioinformatics*, 3(1), February 2021.

[26] Zhendong Liu, Yurong Yang, Dongyan Li, Xinrong Lv, Xi Chen, and Qionghai Dai. Prediction of the RNA tertiary structure based on a random sampling strategy and parallel mechanism. *Frontiers in Genetics*, 12, January 2022.

[27] Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother, and Janusz M. Bujnicki. Simrna: a coarse-grained method for RNA folding simulations and 3d structure prediction. *Nucleic Acids Research*, 44(7):e63–e63, December 2015.

[28] Bradley Blitvich and Andrew Firth. Insect-specific flaviviruses: A systematic review of their discovery, host range, mode of transmission, superinfection exclusion potential and genomic organization. *Viruses*, 7(4):1927–1959, April 2015.

[29] Goro Kuno, Gwong-Jen J. Chang, K. Richard Tsuchiya, Nick Karabatsos, and C. Bruce Cropp. Phylogeny of the genus flavivirus. *Journal of Virology*, 72(1):73–83, January 1998.

[30] Keren Halabi and Itay Mayrose. Mechanisms underlying host range variation in flavivirus: From empirical knowledge to predictive models. *Journal of Molecular Evolution*, 89(6):329–340, May 2021.

[31] Theodore C. Pierson and Michael S. Diamond. The continued threat of emerging flaviviruses. *Nature Microbiology*, 5(6):796–812, May 2020.

[32] Kymberly A. Gyure. West nile virus infections. *Journal of Neuropathology amp; Experimental Neurology*, 68(10):1053–1060, October 2009.

[33] Lai Meng Looi and Kaw Bing Chua. Lessons from the nipah virus outbreak in malaysia. *Malaysian Journal of Pathology*, 29(2):63–67, 2007.

[34] Sadie J. Ryan, Colin J. Carlson, Erin A. Mordecai, and Leah R. Johnson. Global expansion and redistribution of aedes-borne virus transmission risk with climate change. *PLOS Neglected Tropical Diseases*, 13(3):e0007213, March 2019.

[35] Rachel Lowe, Christovam Barcellos, Patrícia Brasil, Oswaldo Cruz, Nildimar Honório, Hannah Kuper, and Marilia Carvalho. The zika virus epidemic in

brazil: From discovery to future implications. *International Journal of Environmental Research and Public Health*, 15(1):96, January 2018.

[36] Elena Gianchecchi, Virginia Cianchi, Alessandro Torelli, and Emanuele Montomoli. Yellow fever: Origin, epidemiology, preventive strategies and future prospects. *Vaccines*, 10(3):372, February 2022.

[37] Soon Hoe Ho, Jue Tao Lim, Janet Ong, Hapuarachchige Chanditha Hapuarachchi, Shuzhen Sim, and Lee Ching Ng. Singapore's 5 decades of dengue prevention and control—implications for global dengue control. *PLOS Neglected Tropical Diseases*, 17(6):e0011400, June 2023.

[38] Alaa Badawi, Russanthy Velummailum, Seung Gwan Ryoo, Arrani Senthinathan, Sahar Yaghoubi, Denitsa Vasileva, Emma Ostermeier, Mikayla Plishka, Marcel Soosaipillai, and Paul Arora. Prevalence of chronic comorbidities in dengue fever and west nile virus: A systematic review and meta-analysis. *PLOS ONE*, 13(7):e0200200, July 2018.

[39] Xijing Qian and Zhongtian Qi. Mosquito-borne flaviviruses and current therapeutic advances. *Viruses*, 14(6):1226, June 2022.

[40] Gábor Kemenesi and Krisztián Bányai. Tick-borne flaviviruses, with a focus on powassan virus. *Clinical Microbiology Reviews*, 32(1), December 2018.

[41] Irani Ratnam, Karin Leder, Jim Black, and Joseph Torresi. Dengue fever and international travel. *Journal of Travel Medicine*, 20(6):384–393, November 2013.

[42] Syed Soheb Fatmi, Rija Zehra, and David O. Carpenter. Powassan virus—a new reemerging tick-borne disease. *Frontiers in Public Health*, 5, December 2017.

[43] Rong Zhao, Meiyue Wang, Jing Cao, Jing Shen, Xin Zhou, Deping Wang, and Jimin Cao. Flavivirus: From structure to therapeutics development. *Life*, 11(7):615, June 2021.

[44] Mayra Diosa-Toro, K. Reddisiva Prasanth, Shelton S. Bradrick, and Mariano A. Garcia Blanco. Role of rna-binding proteins during the late stages of flavivirus replication cycle. *Virology Journal*, 17(1), April 2020.

[45] Suchetana Mukhopadhyay, Richard J. Kuhn, and Michael G. Rossmann. A structural perspective of the flavivirus life cycle. *Nature Reviews Microbiology*, 3(1):13–22, January 2005.

[46] Andrii Slonchak and Alexander A. Khromykh. Subgenomic flaviviral rnas: What do we know after the first decade of research. *Antiviral Research*, 159:13–25, November 2018.

[47] Wy Ng, Ruben Soto-Acosta, Shelton Bradrick, Mariano Garcia-Blanco, and Eng Ooi. The 5' and 3' untranslated regions of the flaviviral genome. *Viruses*, 9(6):137, June 2017.

[48] Roman Ochsenreiter, Ivo Hofacker, and Michael Wolfinger. Functional RNA structures in the 3' UTR of tick-borne, insect-specific and no-known-vector flaviviruses. *Viruses*, 11(3):298, March 2019.

[49] Michael T Wolfinger, Roman Ochsenreiter, and Ivo L Hofacker. Functional RNA structures in the 3 utr of mosquito-borne flaviviruses. In *Virus Bioinformatics*, pages 65–100. Chapman and Hall/CRC, 2021.

[50] Matthew J. Szucs, Parker J. Nichols, Rachel A. Jones, Quentin Vicens, and Jeffrey S. Kieft. A new subclass of exoribonuclease-resistant RNA found in multiple genera of flaviviridae. *mBio*, 11(5), October 2020.

[51] Rachel A. Jones, Anna-Lena Steckelberg, Quentin Vicens, Matthew J. Szucs, Benjamin M. Akiyama, and Jeffrey S. Kieft. Different tertiary interactions create the same important 3d features in a distinct flavivirus xrRNA. *RNA*, 27(1):54–65, October 2020.

[52] Ivar W. Dilweg, Assia Bouabda, Tim Dalebout, Alexander P. Gultyaev, Peter J. Bredenbeek, and R. C. L. Olsthoorn. Xrn1-resistant RNA structures are well-conserved within the genus flavivirus. *RNA Biology*, 18(5):709–717, October 2020.

[53] Quentin Vicens and Jeffrey S. Kieft. Shared properties and singularities of exoribonuclease-resistant rnas in viruses. *Computational and Structural Biotechnology Journal*, 19:4373–4380, 2021.

[54] Rhys Parry, Andrii Slonchak, Lewis J. Campbell, Natalee D. Newton, Humberto J. Debat, Robert J. Gifford, and Alexander A Khromykh. A novel tamanavirus (flaviviridae) of the european common frog (rana temporaria) encodes a divergent class 1b xrn1-resistant RNAelement. July 2023.

[55] Xiaolin Niu, Qiuhan Liu, Zhonghe Xu, Zhifeng Chen, Linghui Xu, Lilei Xu, Jinghong Li, and Xianyang Fang. Molecular mechanisms underlying the extreme mechanical anisotropy of the flaviviral exoribonuclease-resistant RNAs (xrRNAs). *Nature Communications*, 11(1), October 2020.

[56] Rhese Thompson, Derek Carbaugh, Joshua Nielsen, Ciara Witt, Rita Meganck, Atul Rangadurai, Bo Zhao, Jeffrey Bonin, Nathan Nicely, William Marzluff, Aaron Frank, Helen Lazear, and Qi Zhang. Lifetime of ground conformational state determines the activity of structured rna. May 2023.

[57] Xiaolin Niu, Ruirui Sun, Zhifeng Chen, Yirong Yao, Xiaobing Zuo, Chunlai Chen, and Xianyang Fang. Pseudoknot length modulates the folding, conformational dynamics, and robustness of xrn1 resistance of flaviviral xrRNAs. *Nature Communications*, 12(1), November 2021.

[58] Martin Jinek, Scott M. Coyle, and Jennifer A. Doudna. Coupled 5 nucleotide recognition and processivity in xrn1-mediated mrna decay. *Molecular Cell*, 41(5):600–608, March 2011.

[59] Jerome Molleston and Sara Cherry. Attacked from all sides: RNA decay in antiviral defense. *Viruses*, 9(1):2, January 2017.

[60] Wendy Ullmer and Bert Semler. Diverse strategies used by picornaviruses to escape host RNA decay pathways. *Viruses*, 8(12):335, December 2016.

[61] Xianwen Zhang, Yuhan Li, Yingyi Cao, Ying Wu, and Gong Cheng. The role of noncoding RNA in the transmission and pathogenicity of flaviviruses. *Viruses*, 16(2):242, February 2024.

[62] Kevin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975, April 2009.

[63] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. *Multiple Alignment of DNA Sequences with MAFFT*, page 39–64. Humana Press, 2009.

[64] Juliusz Stasiewicz, Sunandan Mukherjee, Chandran Nithin, and Janusz M. Bujnicki. Qrnas: software tool for refinement of nucleic acid structures. *BMC Structural Biology*, 19(1), March 2019.

[65] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, 40(1):82–92, 2002.

[66] Eric P. Nawrocki, Diana L. Kolbe, and Sean R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, March 2009.

[67] Florian Eggenhofer, Ivo L. Hofacker, and Christian Höner zu Siederdissen. CM-Compare webserver: comparing RNA families via covariance models. *Nucleic Acids Research*, 41(W1):W499–W503, 05 2013.

[68] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, January 2010.

[69] Andrea MacFadden, Zoe O'Donoghue, Patricia A. G. C. Silva, Erich G. Chapman, Rene C. Olsthoorn, Mark G. Sterken, Gorben P. Pijlman, Peter J. Bredenbeek, and Jeffrey S. Kieft. Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nature Communications*, 9(1), January 2018.

[70] Ivar W. Dilweg, Anya Savina, Susanne Kothe, Alexander P. Gultyaev, Peter J. Bredenbeek, and Rene C.L. Olsthoorn. All genera of flaviviridae host a conserved xrn1-resistant RNA motif. *RNA Biology*, 18(12):2321–2329, April 2021.

[71] Benjamin M. Akiyama, Hannah M. Laurence, Aaron R. Massey, David A. Costantino, Xuping Xie, Yujiao Yang, Pei-Yong Shi, Jay C. Nix, J. David Beckham, and Jeffrey S. Kieft. Zika virus produces noncoding rnas using a multi-pseudoknot structure that confounds a cellular exonuclease. *Science*, 354(6316):1148–1152, December 2016.

List of Figures

List of Tables

Table Appendix 1: Mosquito-borne flaviviruses (MBFV)

Abbreviation	Full Name
AROAV	Aroa virus
BAGV	Bagaza virus
BANV	Banzi virus
DENV1	Dengue virus type 1
DENV2	Dengue virus type 2
DENV3	Dengue virus type 3
DENV4	Dengue virus type 4
JEV	Japanese encephalitis virus
KOKV	Kokobera virus
KOUV	Koutango virus
KUNV	Kunjin virus
KVEV	Kedougou virus
NMV	Ntaya virus
SLEV	Saint Louis encephalitis virus
TMUV	Tembusu virus
UGSV	Uganda S virus
USUV	Usutu virus
WESSV	Wesselsbron virus
WNV	West Nile virus
YFV	Yellow fever virus
ZIKV	Zika virus

Table Appendix 2: Tick-borne flaviviruses (TBFV)

Abbreviation	Full Name
ALKV	Alkhurma virus
DTV	Deer tick virus
GGV	Gadgets Gully virus
KAMV	Kama virus
KFDV	Kyasanur Forest disease virus
KSIV	Karshi virus
LGTV	Langat virus
LIV	Louping ill virus
MPFV	Meaban virus
NEGV	Negishi virus
OHFV	Omsk hemorrhagic fever virus
POWV	Powassan virus
SGEV	Spanish goat encephalitis virus
SREV	Saumarez Reef virus
TBEV	Tick-borne encephalitis virus
TYUV	Tyuleniy virus
XiFV	Xinyang flavivirus

Table Appendix 3: Insect-specific flaviviruses (ISFV)

Abbreviation	Full Name
AEFV	Aedes flavivirus
BinJV	Binjari virus
BJV	Barkedji virus
CFAV	Cell fusing agent virus
DONV	Donggang virus
EPEV	Ecuador Paraiso Escondido virus
GUAV	Guangdong mosquito virus
KRV	Kamiti River virus

Table Appendix 4: No known vector flaviviruses (NKV)

Abbreviation	Full Name
MMLV	Montana myotis leukoencephalitis virus
MODV	Modoc virus
RBV	Rio Bravo virus