The application of nucleic acid interaction structure prediction

by

Tara Newman
B.Sc., University of Victoria, British Columbia, Canada, 2021

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science

in the Department of Computer Science

The application of nucleic acid interaction structure prediction

by

Tara Newman
B.Sc., University of Victoria, British Columbia, Canada, 2021

Supervisory Committee

Dr. Hosna Jabbari, Advisor, Supervisor
(Department of Computer Science)

Dr. Ulrike Stege, Departmental Member
(Department of Computer Science)

ABSTRACT

Motivation: Understanding how nucleic acids interact is essential for understanding their function. Controlling these interactions, for example, can allow us to detect diseases and create new therapeutics. During quantitative reverse-transcription polymerase chain reaction (qRT-PCR) testing, having nucleic acids interact as designed is essential for ensuring accurate test results. Accurate testing is an important consideration during the detection of COVID-19, the disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Results: I introduced the program DinoKnot (Duplex Interaction of Nucleic acids with pseudoKnots) that follows the hierarchical folding hypothesis to predict the secondary structure of two interacting nucleic acid strands (DNA/RNA) of similar or different type. DinoKnot is the first program that utilizes stable stems in both strands as a guide to find the structure of their interaction. Using DinoKnot, I predicted the interaction structure between the SARS-CoV-2 genome and nine reverse primers from qRT-PCR primer-probe sets. I compared these results to an existing tool RNAcofold and highlighted an example to showcase DinoKnot's ability to predict pseudoknotted structures. I investigated how mutations to the SARS-CoV-2 genome may affect the primer interaction and predicted three mutations that may prevent primer binding, reducing the ability for SARS-CoV-2 detection. Interaction structure results predicted by DinoKnot that showed disruption of primer binding were consistent with a clinical example showing detection issues due to mutations. DinoKnot has the potential to screen new SARS-CoV-2 variants for possible detection issues and support existing applications involving DNA/RNA interactions, such as microRNA (miRNA) target site prediction, by adding structural considerations to the interaction to elicit functional information.

Contents

List of Acronyms

A, C, T, G, U Adenine, Cytosine, Thymine, Guanine, Uracil

cDNA complementary DNA

CCR5 human immune- functioning C-C chemokine receptor 5

COVID-19 coronavirus disease 2019

Ct cycle threshold

DinoKnot Duplex Interaction of Nucleic acids with pseudoKnots

DNA deoxyribonucleic acid

MFE minimum free energy

mRNA messenger RNA

miRNA miRNA

NMR nuclear magnetic resonance

PCR polymerase chain reaction

-1 PRF programmed - 1 ribosomal frameshifting

qRT-PCR quantitative reverse transcription polymerase chain reaction

RNA ribonucleic acid

RT reverse transcriptase

VARNA Visualization Applet for RNA

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Hosna Jabbari, for giving me this opportunity, believing in me, and supporting my research.

This thesis is dedicated to Lance Lansing. Thank you for helping me in my bioinformatics journey, for supporting me, for making me laugh, and for everything.

PREFACE

The preliminary results of this thesis were published in the IEEE International Conference on Healthcare Informatics 2021, in which Tara Newman was the first author [8]. The long paper presented in the conference was written by Tara Newman and Dr. Hosna Jabbari. Hiu Fung Kevin Chang worked on the development of the DinoKnot algorithm during a co-op work term.

Chapter 1

Introduction

Interactions between DNA and RNA molecules are fundamental to many processes, including clinical testing for Coronavirus Disease 19 (COVID-19), and as part of a host's response to the disease [1, 7, 9, 10]. Binding of DNA/RNA molecules typically occurs between complementary nucleotide sequences. However, it can be important to consider the structure of DNA/RNA molecules at the expected binding site and its flanking area. For example, if the area on an RNA molecule where a complementary nucleic acid strand is expected to bind is part of a stable intramolecular structure, it is important to consider the energetics of the interaction. This determines if it would be more energetically favourable for the RNA to remain in the intramolecular structure or to bind to a complementary DNA/RNA molecule.

COVID-19 is the disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); a positive sense RNA virus most commonly detected clinically using quantitative reverse-transcription polymerase chain reaction (qRT-PCR) on samples collected by nasopharyngeal swabs [11]. qRT-PCR involves the use of primer-probe sets consisting of small DNA oligonucleotides. During qRT-PCR, the reverse primer first binds to the positive sense RNA genome so that an enzyme involved in the reaction called reverse transcriptase (RT) can use the primer to generate the complementary DNA (cDNA) of the negative sense strand [12].

The primer-probe set used to detect SARS-CoV-2 depends on which qRT-PCR assay is used. At the beginning of the COVID-19 pandemic, the China Center for Disease Control (China CDC), United States CDC (US CDC), Charité Institute of Virology, Universitätsmedizin Berlin (Charité), and Hong Kong University (HKU) developed assays to detect SARS-CoV-2 that all target different areas of the genome [1]. The DNA-RNA interactions that occur during this process are between the reverse

primer of the primer-probe set and the RNA genome.

During SARS-CoV-2 infection, interactions also occur between host microRNAs (miRNAs) and the SARS-CoV-2 genome [9]. miRNAs are short RNA sequences, approximately 22 nucleotides in length, that are involved in post-transcriptional regulation of other RNAs [13]. Considering the structure of these interactions may elicit a deeper investigation into their function. If these miRNAs are acting as an immune response to clear the virus or are being exploited by the virus for its own benefit, understanding their mechanism of action could inspire potential therapeutics to either promote or prevent the interaction.

To study the structure of such interactions, I present DinoKnot, a program that given two nucleic acid strands predicts their interaction structure. I aimed to determine whether the RNA structure of the SARS-CoV-2 genome affects the binding of the reverse primers in the qRT-PCR assay and whether this correlated qualitatively to the analytical efficiencies and sensitivities shown experimentally by Vogels *et al.* [1]. I further predicted how mutations in the gene areas targeted by the primer-probe sets change the interaction structure, potentially affecting the primer/probe sensitivity for SARS-CoV-2 detection. Finally, I used DinoKnot to present the predicted interaction structures between a human miRNA and the top three predicted binding sites on the SARS-CoV-2 genome, considering how the area flanking the target site affects the interaction structure. I discuss future work in the design of nucleic acid based testing and combining DinoKot with existing miRNA target site prediction tools to add structural considerations to the interaction.

1.1 Thesis Objective

This thesis aims to present applications of DinoKnot. Building upon the existing single RNA structure prediction program Iterative HFold [2], DinoKnot is the first tool to predict the interaction structure of nucleic acid strands of similar or different type (DNA/RNA), while also considering pseudoknotted structure. The objective of this thesis are as follows:

1. to investigate the minimum free energy interaction structures between the SARS-CoV-2 genome and reverse primers from qRT-PCR primer-probe sets used to detect COVID-19. The purpose of predicting the interaction structures is to determine how the RNA structure of the genome and mutations may affect the

interaction, and thus the ability to detect COVID-19.

2. to investigate how intramolecular structures may impact interactions that occur between miRNAs and their target sites when predicting the most energetically favourable interaction structures.

1.2 Thesis Contributions

Described in this thesis are the following contributions I have made:

1. I introduce DinoKnot (Duplex Interaction of Nucleic acids with pseudoKnots), a program that aims to overcome the limitations of existing methods that do not consider intramolecular or pseudoknotted structure.

2. I use DinoKnot to predict the interaction structures between the SARS-CoV-2 genome and nine reverse primers from primer-probe sets that were experimentally validated at the beginning of the COVID-19 pandemic and determine structural insights that may explain the reduced analytical efficiency of a primer-probe set.

3. I compare DinoKnot to the closest existing tool, RNAcofold [14], that also predicts interaction structures between nucleic acids of similar or different type but only considers the interaction site when predicting the structure and does not consider pseudoknots.

4. I predict how mutations in the primer/probe binding region may affect the interaction structure, and potentially reduce the ability of the primer-probe set to detect SARS-CoV-2.

5. I investigate a clinical report of detection issues by comparing the predicted interaction structures of both the SARS-CoV-2 reference genome and variant sequence to determine how mutations may affect the primer/probe interaction to the target sequence.

6. I use DinoKnot to predict the interaction structure of a miRNA (miR-2932) with the top three binding sites predicted on the SARS-CoV-2 genome, considering how the area flanking the target site affects the interaction structure.

Chapter 2

Background

In this chapter, I first introduce the two types of nucleic acids, DNA and RNA. I describe how these nucleic acids can from structure and introduce RNA secondary structure prediction. I discuss the computational methods available and review existing tools for both single and interacting nucleic acid structure prediction. I highlight the current limitations of these methods. I then describe examples of each possible type of nucleic acid interaction (DNA-DNA, RNA-RNA, DNA-RNA) and their biological importance, such as their applications in biotechnology.

2.1 Nucleic Acids

The two types of biological nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA encodes the genetic blueprint (genome) of an organism with both protein-coding and non-protein-coding regions. Genes are sections of the genome that encode a functional product. Through a process called transcription, genes can be copied into an RNA molecule. (Note: RNAs are also transcribed from sections of the genome with no known functions [13].) RNAs transcribed from protein-coding regions of DNA are known as messenger RNAs (mRNA) and are used to create proteins through a process called translation using cellular machinery called ribosomes. RNAs transcribed from non-protein-coding regions of DNA are known as non-coding RNAs (ncRNA) and have varying types and functions in cellular processes [15]. RNA can also encode the genome of some viruses.

Nucleic acids are made up of repeating nucleotide molecules. Nucleotides consist of a nitrogenous base and a phosphate-sugar backbone. DNA contains a deoxyribose

sugar and RNA contains a ribose sugar. Both DNA and RNA strands have directionality, with one end labelled as the 5' end and the other labelled as the 3' end. The nitrogenous bases are categorized as purines and pyrimidines. DNA and RNA are made up of the same purine bases Adenine (A) and Guanine (G) and the pyrimidine base Cytosine (C). However, DNA uses the pyrimidine base Thymine (T) and RNA uses the pyrimidine base Uracil (U). Base pairing occurs when hydrogen bonds form between the nitrogenous bases. Generally, purines form base pairs with pyrimidines. Under the rules of Watson-Crick base pairing, DNA molecules form $G.C$ and $A.T$ base pairs and RNA molecules form $G.C$ and $A.U$ base pairs [16]. Here "." represents a pairing of the two bases. However, RNA can also form a *wobble* base pair where G pairs with U ($G.U$) [16]. $A.T$, $A.U$ and $G.U$ base pairs form two hydrogen bonds. $G.C$ base pairs form three hydrogen bonds, which makes the pairing more stable. Intramolecular base pairing occurs between nucleotides on the same nucleic acid strand and intermolecular base pairing occurs between nucleotides on separate strands. Intermolecular base pairs can occur between DNA-DNA, RNA-RNA, and DNA-RNA, causing nucleic acid interactions.

2.2 RNA Secondary Structure Prediction

In addition to base complementarity, the structure of nucleic acid molecules has an impact on their interactions, and thus their function. It is important to consider the free energy of intermolecular and intramolecular base pair interactions to determine whether it is energetically favourable for an interaction to occur. This section has an emphasis on RNA structure prediction since DNA typically exists in a stable double helix structure but note that the principles of structure prediction are essentially the same for single stranded DNAs.

An RNA is a single stranded molecule with two distinct ends, namely 5' and 3'. An RNA molecule is represented by a sequence, S, of its four bases, A, C, G, and U, arranged on a line (representing the backbone) from 5' (left) to 3' (right) ends, highlighted as the RNA sequence in Fig 2.1. The nucleotide sequence is referred to as the primary structure [17]. The *secondary structure* is formed when the single stranded molecule folds back onto itself by forming *intramolecular* base pairs, represented as arcs connecting base pairs in Fig 2.1. The length of the RNA molecule is denoted by n and each base of the RNA sequence is referred to by its index i, $1 \leq i \leq n$. Complementary bases bind (form hydrogen bonds) and form base pairs

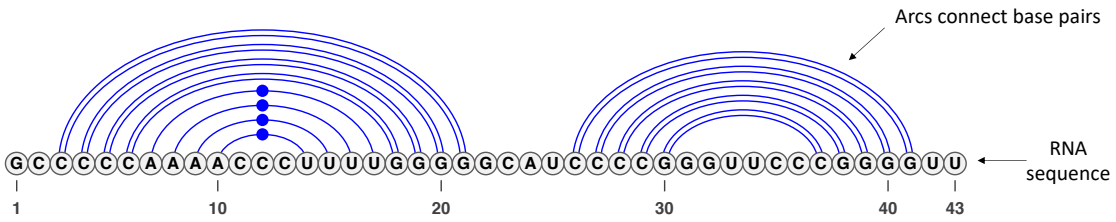

Figure 2.1: **Representation of RNA structure.** The RNA sequence represents the primary structure. Base pairing of nucleic acids forms the secondary structure, represented as arcs between base pairs. When arcs do not cross over each other (ie. nested base pairs), this is a pseudoknot-free structure. The pairing of multiple bases in a row within the arcs represents a stem structure.

($A.U$, $C.G$, and $G.U$). A *secondary structure*, R, is then defined as a set of base pairs $i.j$, $1 \leq i < j \leq n$; $i.j$ and $k.j$ can belong to the same set if and only if $i = k$ i.e., each base may pair at most with one other base. If $i.j$ and $k.l$ are two base pairs of a secondary structure, R, such that $1 \leq i < k < j < l \leq n$, then $i.j$ crosses $k.l$. A *pseudoknotted* secondary structure refers to a structure with such crossing base pairs, represented in Fig 2.2 where the bases highlighted in pink cross between arcs. A *pseudoknot-free* secondary structure refers to a structure without crossing base pairs, as shown in Fig 2.1 where bases pairs only form nested base pairs within arcs.

The RNA structure forms because it is energetically favourable for bases to form paired helices. Different base pairing patterns in a secondary structure define different loop types. The main structural features that form are stems, hairpin loops, interior loops, multi-loop branches, and bulges [17]. Stems are stretches of continuous base pairs which stabilize the structure of the molecule by lowering the free energy. In Fig 2.1, the nested arcs represent a stem structure. The remaining loop structures contain unpaired bases that increase the free energy of the structure. Note that the principles of structure prediction are essentially the same for single stranded DNAs.

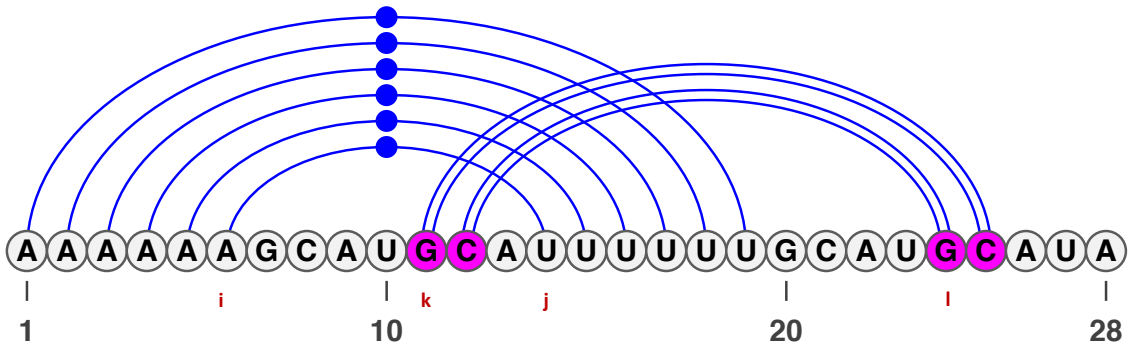

Figure 2.2: **Representation of a pseudoknotted structure.** A pseudoknot occurs when base pairs ($i.j$ and $k.l$) cross over each other, highlighted by the pink bases that form base pairs that cross over another stem of base pairs. This allows the RNA to form a more compact structure.

Laboratory methods of visualizing the structure of nucleic acids include X-ray crystallography and Nuclear Magnetic Resonance (NMR) [17]. Due to the time and cost requirements of these techniques, it is significantly more feasible to predict the structure using computational methods [17]. Prediction of secondary structure gives researchers insight on how to interpret experiment results and can guide new experimentation into the function of RNA [18].

Seetin and Mathews [18] provide an in depth overview of available RNA structure prediction methods. In summary, there are computational methods for single-sequence and multiple-sequence secondary structure prediction [18]. Multiple-sequence structure prediction involves predicting structures that are conserved within two or more RNA sequences [18]. Methods of this type are most relevant when comparing RNA sequences that are evolutionarily conserved. However, many RNAs have no known related sequences (e.g., novel RNAs) and thus multiple-sequence structure prediction methods cannot be used [18]. The methods discussed in this thesis use single-sequence structure prediction, which is capable of predicting the structure from the RNA sequence alone. Single-sequence structure prediction often uses free energy minimization to predict the secondary structure with the lowest free energy (the most stable structure) [18]. Other methods include predicted maximum expected accuracy structures, which calculate base pair probabilities to predict structures with the highest base pair probabilities, and suboptimal structure prediction, which predicts structures that are not the most energetically favourable [18]. This thesis focuses on single-sequence structure prediction methods that find the minimum free energy (MFE) structure. The most energetically favourable structure is the most stable and will be the most prevalent (in the highest concentration) when the RNA is in an equilibrium state [18]. The assumptions of the RNA structure prediction model focuses solely on the hydrogen bonds that form between base pairs [19].

The formula for the free energy is:

$$\Delta G = \Delta H - T\Delta S$$

where ΔG is the Gibbs Free energy, ΔH is the enthalpy, T is the temperature and ΔS is the entropy. The free energy of a structure is minimized by hydrogen bonding, which is why $G.C$ base pairs are the most stable.

When finding the energetically most stable (i.e., MFE) secondary structure from the base sequence using computational secondary structure prediction, each loop is

assigned an energy value. These energy values are known as energy parameters. Some parameter sets have been derived directly from experiments, and others are extrapolated based on experimentally determined values. Energy parameters are strand type specific, i.e., similar loops in an RNA molecule have different assigned energy than the ones in a DNA molecule because DNA and RNA are chemically different. Existing MFE structure prediction methods find the minimum free energy structure for a given sequence from the pool of all possible structures. Methods for MFE pseudoknot-free structure prediction use dynamic programming to find the MFE structure [20, 21]. These methods take the RNA sequence as input and find the most energetically favourable base pairs that make up the MFE structure. Dynamic programming breaks down the problem into smaller sub-problems and uses recursion in order to find the optimal solution (the MFE structure) [17]. Since prediction of the MFE pseudoknotted structure is NP-hard [22, 23] (cannot be solved in polynomial time complexity) and even inapproximable [24] (the solution cannot be approximated), methods for pseudoknotted MFE structure prediction focus on a restricted class of structures [25, 26, 27, 28, 29].

2.2.1 RNA Secondary Structure Prediction Tools

In this section, the MFE RNA secondary structure prediction methods related to this thesis are discussed.

Iterative Hfold

Iterative HFold [2] is a dynamic programming algorithm that predicts the MFE, potentially pseudoknotted, structure of a single stranded RNA using a pseudoknot-free structure approximation as input. Iterative HFold [2] uses four methods to predict the MFE structure. One method uses HFold [28], which follows the hierarchical folding hypothesis (discussed further in Chapter 3.1), adding pseudoknotted base pairs to the pseudoknot-free approximation structure if it is energetically favourable to do so. The other three methods add or remove base pairs from the input structure in order to find the final structure with the lowest free energy [2].

RNAfold

RNAfold [5] is a dynamic programming algorithm from the ViennaRNA Package 2.0 that predicts the MFE structure of a single stranded RNA. RNAfold [5] does

not require an input structure but is only able to output pseudoknot-free structures. The ViennaRNA Package contains a suite of command-line tools with varying functions/parameters. This package contains many different programs solve many different RNA secondary structure prediction programs, such as predicting the secondary structure of suboptimal (non-MFE) structures, of interacting nucleotides, and more [5].

Both Iterative Hfold and RNAfold have $\mathcal{O}(n)^3$ run times. The advantage of using Iterative HFold is that it can predict possibly pseudoknotted structures within the same time complexity as RNAfold, which can only predict pseudoknot-free structures. The 95% confidence interval for the accuracy of Iterative HFold with HotKnots hotspots as the input structure on predicting pseudoknotted structures is (72.83%,83.37%) and (74.93%, 80.26%) for pseudoknot-free structures [2]. RNAfold 2.0 has a sensitivity of 0.739 (73.9%), a specificity of 0.792 (79.2%), a MCC of 0.763 and an F measure of 0.761 [5].

Secondary Structure Visualization

RNA structure prediction software packages often output the secondary structure of an RNA in dot-bracket notation [30]. Here, a "." represents a base that is unpaired. A "(" is assigned to the beginning index of a base pair (i) and a ")" is assigned to the end index of a base pair (j), for each i and j base pair of the nucleic acid. Fig 2.3 shows an RNA secondary structure and the corresponding dot-bracket notation which represents that structure. The resulting dot-bracket notation of the structure is a string consisting of dots and brackets the length of the RNA sequence representing which bases pair or remain unpaired. VARNA [4] (Visualization Applet for RNA) is a java-based visualization tool that takes as input the sequence and structure of an RNA and outputs the secondary structure visualization using one of four drawing algorithms [4]. Of most relevance is the linear algorithm, which visualizes the RNA backbone as a horizontal line and base pairs are represented as arcs connecting the bases of the backbone, as was represented in Fig 2.1, 2.2, and 2.3 [4]. Fig 2.4 compares the linear algorithm arc diagrams to nucleic acid view which visualizes how the RNA folds together.

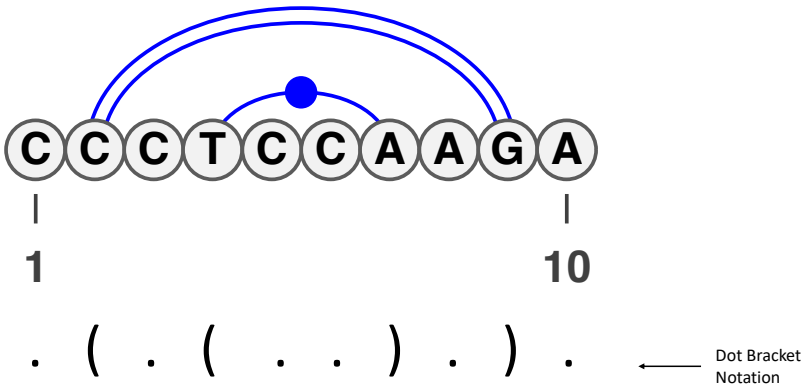

Figure 2.3: **Representation of dot-bracket notation.** A "." represents an unpaired base and a "(" and ")" represent the pairing of two bases.

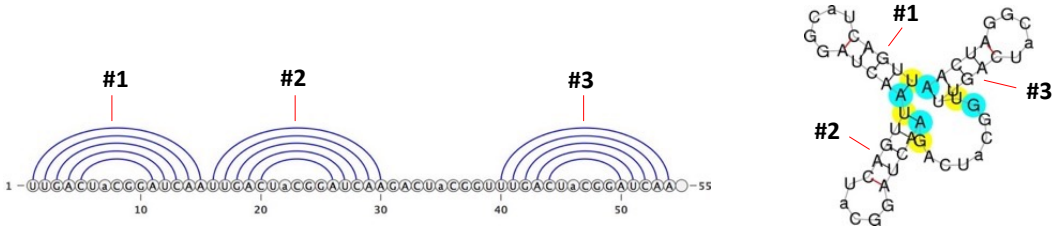

Figure 2.4: **Representation of RNA secondary structure in linear arc diagram view (left) versus nucleic acid view (right).** The three nested arcs of base pairs form the three stem structures when folded together with loops of unpaired bases in between.

2.2.2 Nucleic Acid Interaction Structure Prediction

The interaction of intermolecular base pairing between nucleic acids is a duplex nucleic acid interaction, in which both nucleic acid strands can be structured. Their structures can change upon interaction with one another to accommodate formation of more stable base pairings. Many tools exist to predict interactions between nucleic acids of the same type (DNA-DNA or RNA-RNA) [3, 31, 32, 33, 34, 35]. Fewer tools exist that can predict the interactions between nucleic acids of the same or different type (DNA-RNA). The focus of this section is narrowed to these methods and I discuss their limitations.

RNAcofold

RNAcofold is a program from the ViennaRNA package 2.0 that predicts the interaction structure between two RNA molecules by concatenating them together and outputting the pseudoknot-free common MFE interaction structure [5]. The ViennaRNA package version 2.1.9h implements RNA/DNA hybrid support for this program, allowing for the prediction of DNA-DNA, RNA-RNA, and DNA-RNA interaction structures [14]. RNAcofold [14] is not able to predict pseudoknotted structures.

RNAduplex

RNAduplex is another program from the ViennaRNA package 2.0 that has RNA/DNA hybrid support in the ViennaRNA package v2.1.9h preliminary version [5, 14]. RNAduplex considers the structure after hybridization of the two RNA/DNA sequences [36]. Therefore, it ignores the intramolecular structures and multi-branch loops when considering potential binding sites [5]. Like RNAcofold, RNAduplex is also unable to predict interaction structures with pseudoknots.

Limitations

A limitation of the tools discussed in this section (RMAcofold and RNAduplex) is that they do not consider the intramolecular structure of the nucleic acids prior to interaction. Intramolecular structures are important to consider because even if a sequence on one nucleic acid is complementary to another nucleic acid, if this potential binding site is part of a stable stem structure, it may not be accessible or energetically favourable for the other nucleic acid to bind to that site. As well, these tools do not

consider pseudoknotted structures which allow the RNA to fold into a more compact structure by allowing more base pairing.

Pseudoknots have important biological functions, such as involvement in gene expression. One example is that pseudoknots are known to cause programmed -1 ribosomal frameshifting (-1 PRF) [37, 38]. During -1 PRF, the pseudoknotted structure directs the ribosome to slip back one base during translation, resulting in mRNA decay in order to regulate gene expression [38]. -1 PRF is common in many viruses [39, 40]. SARS-CoV-2 and other related viruses use -1 PRF as a mechanism for the expression of the *ORF1b* gene in order to have successful viral replication [37, 40]. Pseudoknots are also involved in the mechanisms of self-cleaving ribozymes and in self-cleaving of introns during splicing [40]. As well, human telomerase mRNA has a pseudoknotted structure and mutations to this structure are known to cause multiple diseases [40]. Listed here are only a few examples of the biological importance of pseudoknots. Considering interactions between pseudoknotted structures and other DNA/RNA molecules is important for giving structural insights into these types of molecular functions [38, 39, 40, 37].

Both of these factors, intramolecular and pseudoknotted structures, may have an impact on the nucleic acid interaction. Interactions between nucleic acids are not static structures at proposed binding sites. When there is competition for binding, the structure of the entire molecule may have an impact when determining if it is energetically favourable for an interaction to occur. In Chapter 3, a method is introduced that aims to address the limitations of existing methods.

2.3 Nucleic Acid interactions

This section aims to highlight the biological and clinical importance of nucleic acid interactions between DNA-DNA, RNA-RNA and DNA-RNA molecules.

2.3.1 DNA-DNA interactions

Within a cell, DNA generally exists as a double stranded helix of two complementary DNA strands, also called a duplex [41]. The DNA duplex is complementary in that the bases on one strand completely pair, in order, to the bases on the other strand. This allows the DNA to form a stable structure that can be compacted via supercoiling, where the double stranded DNA can be wound tightly together [42]. In the context

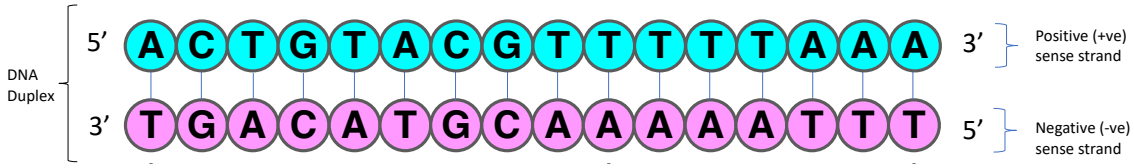

Figure 2.5: **Representation of a DNA duplex.** The strand in the 5' to 3' direction is called the positive sense strand and the stand in the 3' to 5' direction is called the negative sense strand

of a gene, one of the DNA strands is called the positive sense strand, also known as the coding strand. Fig 2.5 depicts a DNA duplex and highlights the positive and negative strand. The sequence of this strand is used during transcription to generate an RNA copy (i.e., this strand encodes the genetic information). The other strand is called the negative sense strand, also known as the non-coding strand.

An application in biotechnology involving DNA-DNA interactions is polymerase chain reaction (PCR). PCR is a laboratory technique used to amplify sections of DNA [43]. PCR involves the use of two short DNA strands (oligonucleotides), called a forward and reverse primer. The forward primer is designed to bind to a complementary region on the negative sense strand and the reverse primer binds to a complementary region on the positive sense strand [43]. Since the double-stranded DNA is a stable structure, high temperatures are used to separate the DNA-duplex into individual strands to allow for the shorter primers to bind to their target region [43]. Once the primer binds to its target region, an enzyme called DNA polymerase is able to generate a copy of the DNA strand. This cycle is repeated until the target section is amplified enough times to allow for detection as demonstrated in Fig 2.6.

2.3.2 RNA-RNA interactions

While the main role of DNA is to store genetic material, RNA is much more complex in that it is involved in the creation of proteins, the regulation of gene expression, and many other diverse functions [13]. One way of classifying RNA is into the three categories of protein-coding associated RNA (e.g. , mRNA), regulatory RNA, and parasitic RNA [13]. Only a subset of RNAs will be discussed in this thesis. The reader is refered to Dai *et al.* for an in depth review on the varying types of RNAs and their functionalities [13].

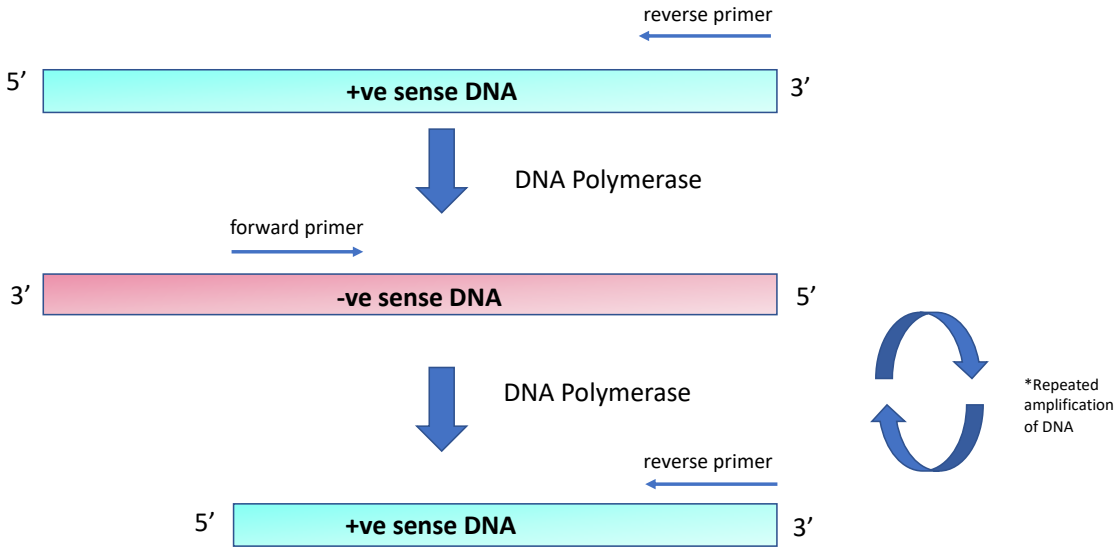

Figure 2.6: **Interaction of the primers to the DNA strands during PCR amplification.** The reverse primer interacts with the positive sense strand and the forward primer interacts with the negative sense strand in order to amplify a section of DNA for detection.

Regulatory RNA and Parasitic RNA

An example of a regulatory RNA is microRNA (miRNA). miRNA is a short sequence of RNA (22 nucleotides) that is involved in the post-transcriptional regulation of mRNA [13]. The first 2-8 nucleotides at the 5' end of the miRNA is known as the seed region. miRNAs regulate target RNA with sequences that are complementary to their seed region [9]. Interactions between miRNAs and mRNA can prevent the interactions required for translation or can cause the mRNA to be degraded so that it is unable to be translated into a protein [44, 13]. This is used as a method for gene expression regulation [13].

RNA viruses, such as SARS-CoV-2, are an example of parasitic RNA [13]. Viruses are capable of infecting host cells and hijacking cellular machinery in order to replicate their viral genome. Interactions between viruses and miRNAs have also been shown to have an important role during viral infection [10, 9]. miRNA interactions may have positive effects for the host (i.e., part of the host's defense mechanisms) or have positive effects for the virus (i.e., exploited for the benefit of the virus) [10]. miRNAs have been shown to target sites on RNA viruses such as human immunodeficiency virus (HIV), hepatitis B and C viruses, influenza H1N1 and Rhinovirus [9]. Similar to how miRNA interactions can lead to degradation of mRNA for gene expression regulation, miRNA interactions may also lead to degradation of the RNA virus in order to stop the viral infection. However, this mechanism has also been exploited by HIV and hepatitis B virus to keep the the viral count low in order to evade detection by the immune system so that the viral infection can persist within a cell [9]. The role of miRNAs during SARS-CoV-2 infection has been of interest for potential biomarkers and therapeutics [45, 46]. Many potential target sites have been identified using in silico prediction methods and some interactions have been experimentally validated [9, 7]. In particular, miR-2392 was predicted to be upregulated in patients infected with SARS-CoV-2 and to interact with the *NSP2*, *NSP3* and *E* gene regions of the SARS-CoV-2 genome [7]. Through experimental studies, miR-2392 has been shown to suppress mitochondrial gene expression, increase inflammation and glycolysis, and promote symptoms of SARS-CoV-2 infection [7]. Furthermore, miR-2932 was detected in patients infected with SARS-CoV-2 and not detected in uninfected individuals [7].

miRNA Target Site Prediction

While many tools exist to predict interactions between RNA molecules, this section focuses specifically on those designed for interactions between miRNAs and their target sites. miRNA most commonly targets the 3' untranslated region (UTR) of the mRNA (ie. part of the mRNA sequence that isn't translated into a protein) [44]. Since the binding of miRNAs is determined by its seed sequence, many tools specifically identify target sites in the 3' UTR of mRNAs with sequence complementarity to this seed region [47]. Some tools also take into account the thermodynamic stability of the interaction between the miRNA and potential target sites [47]. RNAhybrid [48] is a tool that works similarly to RNAduplex [5] in that it considers only the hybridization site first to determine the MFE of the interaction. RNAhybrid [48] allows the user choose whether or not to include a seed region as a requirement for target site prediction. Other tools also consider whether a target region is conserved across different species [44]. The tool miRanda [6] uses seed sequence similarity, thermodynamic stability, and evolutionary conservation to predict target sites. When considering the free energy of the target site, only the free energy of the interaction between the miRNA and the target site is reported; the intramolecular structures surrounding the target site are not considered.

RNAhybrid [48] and miRanda [6] have both been recently used to identify potential binding sites of human miRNAs to the SARS-CoV-2 genome [9, 7]. The same limitations discussed in section 2.2.2 also apply to the these tools. Specifically in the case of miRNA target site prediction tools, considering the thermodynamic stability of the interaction and seed sequence similarity at the target site are important factors but theses tools do not consider the intramolecular structures of the target site and miRNA prior to interaction. Both these intramolecular structures and pseudoknotted structures may prevent the predicted interaction if it is more energetically favourable to remain in these structures than to form base pair interactions. Other tools exist that do consider the accessibility of the target site, meaning that it takes into account existing intramolecular structures at the target site [47]. For example, PITA [49] is a tool that adds the free energy required to break the intramolecular bonds of the target site prior to interaction to the thermodynamic stability of the interaction with the miRNA [47, 49]. PITA also takes into account regions flanking the target site when determining site accessibility [47, 49]. However, some of the most commonly used miRNA target site prediction tools, such as miRanda [6, 47], are not designed

to predict intramolecular structures. Since there are many different tools that offer different solutions to target site prediction, scientists may have certain experimental specifications that requires the use of existing tools that do not consider intramolecular structures. The method I describe in Chapter 3, has the potential to be combined with any existing miRNA target site prediction tools that already predict functional target sites. This method would add an additional level of verification of energetically favourable binding and address the limitations of existing methods that do not consider site accessibility or pseudoknotted structures.

2.3.3 DNA-RNA Interactions

Interactions between DNA and RNA molecules, or DNA-RNA hybrids, occur during fundamental cellular processes such as transcription and DNA replication [50]. DNA-RNA hybrids were shown in yeast to be involved in the formation of heterochromatin, directed by ncRNAs involved in gene silencing [51, 52]. RNA has also been shown to interact with DNA duplexes in a structure called an R loop as part of the CRISPR/Cas pathway used by bacteria and archaea for adaptive immunity against foreign DNA [51, 53, 54]. CRISPR/Cas also has the potential to be used for treatment of diseases such as cancer and autoimmune disorders through gene editing [54].

DNA-RNA interactions also occur between the SARS-CoV-2 genome and the reverse primer during quantitative reverse transcriptase PCR (qRT-PCR). qRT-PCR is similar to PCR, except that the process begins with an RNA molecule that is copied into a complementary DNA (cDNA) molecule via the enzyme reverse transcriptase (RT). The cycle then repeats with DNA amplification of the target sequence as described in Section 2.3.1.

qRT-PCR

qRT-PCR involves the use of forward and reverse primers, a probe, reverse transcriptase (RT) and DNA polymerase. The forward and reverse primer DNA oligonucleotides bind to complementary sequences in order to amplify a section of the SARS-CoV-2 genome. The reverse primer interacts with complementary sequences on both the positive sense SARS-CoV-2 RNA genome and the positive sense cDNA transcript as shown in Fig 2.7. Therefore, the reverse primer is involved in both RNA-DNA and DNA-DNA interactions. The forward primer interacts with the negative sense cDNA

transcript.

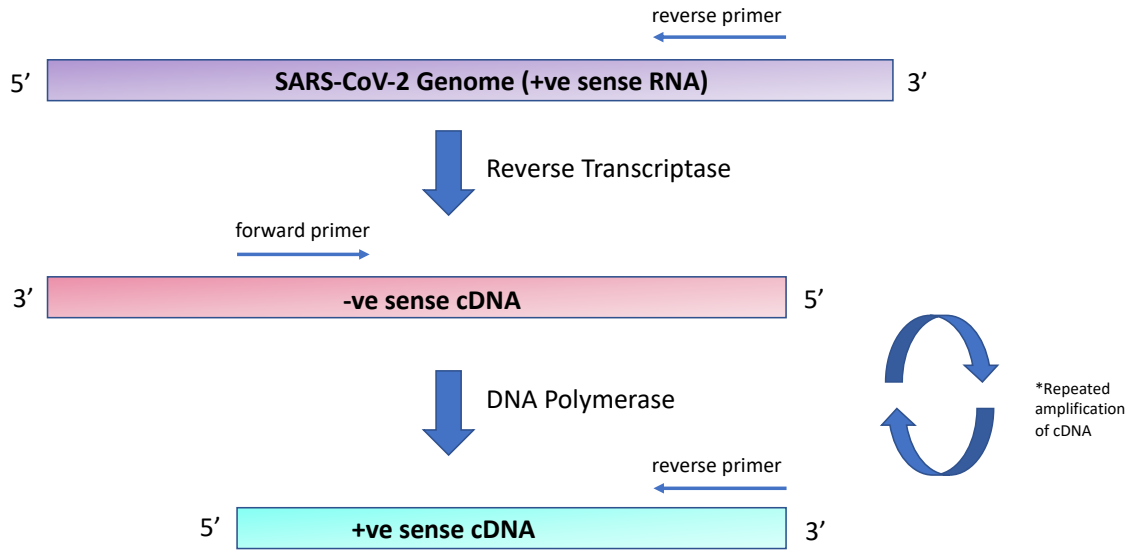

Figure 2.7: **Interaction of the primers to the RNA/cDNA strands during qRT-PCR amplification.** The reverse primer first binds to the target complementary sequence on the SARS-CoV-2 positive (+ve) sense RNA genome. The reverse transcriptase then generates the negative (-ve) sense complementary DNA (cDNA) strand. The forward primer then binds to the negative sense cDNA strand and the DNA polymerase generates the positive sense cDNA strand. The reverse primer binds to the complementary target sequence on the positive sense cDNA and the DNA polymerase generates a new negative sense cDNA strand. This process repeats for strand amplification during qRT-PCR.

The probe is a fluorescent-labelled DNA oligonucleotide that is detected when the probe binds to its target sequence, resulting in a positive test if the fluorescence is higher than the background signal. The cycle threshold (Ct) value is the number of qRT-PCR amplification cycles it takes for the fluorescence to be detected above the background signal [55]. The probe may bind to either the positive or negative sense cDNA transcript, depending on how it is designed. Therefore, the forward primer and probe are only involved in DNA-DNA interactions. RT is an enzyme that uses an RNA as a template to generate a complementary DNA strand. As discussed in Section 2.3.1, DNA polymerase is an enzyme that uses a DNA template to generate a complementary DNA strand.

2.3.4 Importance

At the most basic level, understanding nucleic acid interactions and their structure allows us to understand biological functions. Nucleic acids must interact in order for life to exist and studying these interactions has allowed us to understand fundamental processes such as DNA replication and transcription. Understanding nucleic acid interactions may also allow for the discovery of potential therapeutic target sites, such as in the case of interactions between human miRNAs and the SARS-CoV-2 genome [9]. Knowledge of the interactions between a host and virus may allow us to understand how these interactions could be positive, i.e., part of the host immune response, or negative, i.e., part of viral exploitation of the host for replication/infection [9]. Here, I use positive to mean a positive effect on the health of the host and negative to mean a negative effect on the health of the host. Therapeutics can then be designed to amplify positive interactions or prevent negative interactions. In Section 2.3.2, the negative effects of the interaction between miR-2392 and SARS-CoV-2 were discussed [7]. Computational methods were used to identify potential binding sites for miR-2392 with the SARS-CoV-2 genome [7]. A potential therapeutic has been developed that contains the antisense (complementary) sequence to miR-2923 [7]. The therapeutic therefore competes for binding to miR-2923 in order to prevent the negative effects of its interaction with SARS-CoV-2. The therapeutic was tested in human lung cells and showed an average of 85% viral inhibition [7]. When tested on animal models, the treatment group had a lower viral load compared to the control group, but the results were not statistically significant 2 days post-treatment [7]. This is an example of how understanding and controlling nucleic acid interactions allows us to identify potential therapeutics to treat disease. Nucleic acid therapeutics have been approved to treat diseases such as Duchenne muscular dystrophy, Hypercholesterolemia, and COVID-19 [56]. These therapies involve gene inhibition, addition, replacement or editing of the genome through nucleic acid interactions and have the potential to cure diseases caused by genetic mutations [56]. For processes such as CRISPR/Cas 9, if used as a therapeutic method, it is essential to understand and ensure nucleic acids interact as expected for proper gene editing and to prevent off-target effects.

In the case of laboratory techniques such as PCR and qRT-PCR, their success relies on the primers and probe interacting with the target DNA/RNA as designed in order to amplify the correct section of DNA/RNA. In the height of the COVID-19 pandemic, it was essential to have accurate test results so that infected individuals

isolated in order to prevent spreading the virus to others. With the presence of new SARS-CoV-2 variants, it is important to determine how mutations may affect the primer/probe interaction, especially mutations in the regions where the primers and probe are designed to bind. If the primer/probe is unable to bind to the mutated sequence, this could result in a false-negative test result since the DNA/RNA would not be successfully amplified. Nucleic acid interaction structure prediction can be a useful tool for identifying interaction sites and the structure to be able to study the functions of these interactions.

Chapter 3

DinoKnot

Chapter 2 described the fundamentals of RNA secondary structure prediction and highlighted the relevant existing tools. In this chapter the RNA secondary structure prediction problem is expanded to two interacting molecules. I then describe the method DinoKnot initally developed by Kevin Chang and Dr. Hosna Jabbari and explain how it overcomes the shortcomings of the existing methods on prediction of the secondary structure for two different types of nucleic acid strands.

Existing tools that predict structure of interaction in two molecules mostly focus on similar strands (i.e., DNA/DNA or RNA/RNA) [3, 31, 32, 33, 34, 35], and merely focus on the interaction site (i.e., ignoring the intramolecular structures) [57, 58, 5, 59, 60]. The tool DinoKnot (Duplex Interaction of Nucleic acids with pseudoKnots) aims to address both of these shortcomings. The interacting secondary structure of two nucleic acid sequences can be represented by concatenating the two strands together and keeping track of the gap between the two strands with a linker as demonstrated in Fig 3.1.

When two strands are of similar type, the energy calculation of the concatenated sequence will be similar to that of a single strand of the same type, except for loops containing the gapped region (as they are not true loops when sequences are not concatenated). When two strands of different types interact (i.e , a DNA strand binding to an RNA strand) the situation is more complicated as there are currently no comprehensive energy parameters known for loops formed between the two strands.

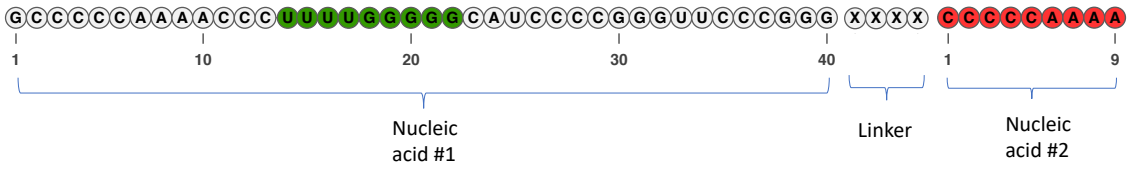

Figure 3.1: **Representation of interacting nucleic acid secondary structure prediction.** The sequence that is complementary to nucleic acid 2 is highlighted in green on nucleic acid 1.

3.1 DinoKnot

Initially developed by Hiu Fung Kevin Chang and Dr. Hosna Jabbari, DinoKnot follows the relaxed hierarchical folding hypothesis [2] for prediction of the minimum free energy (MFE) structure of two interacting nucleic acid strands. Following this hypothesis, an RNA molecule first forms simple pseudoknot-free base pairs before forming more complex and possibly pseudoknotted structures [61]. During this process some of the originally formed base pairs may open up to accommodate more stable pairings. Existing methods based on hierarchical folding, namely HFold [28] and Iterative HFold [2], focus on single RNA structure prediction. DinoKnot, to the best of our knowledge, the first program that follows the relaxed hierarchical folding hypothesis for prediction of pseudoknotted structure of two interacting nucleic acid molecules. Similar to HFold and Iterative HFold, DinoKnot handles a large class of pseudoknotted structures, which include a wide range of commonly found pseudoknotted structures, including H-type pseudoknots and kissing hairpins with arbitrary nested substructures. DinoKnot takes a pair of nucleic acid sequences as input and returns their interaction structure with its corresponding free energy value. Each sequence can be of type RNA or DNA. Note that the minimum free energy structure of two input strands may not involve any interaction if it is energetically more favourable for each sequence to form intramolecular base pairs.

The user can optionally provide a pseudoknot-free input structure (in addition to the input sequence) if such information is available to guide the prediction. If no input structure is provided by the user, DinoKnot will generate up to 20 pseudoknot-free secondary structures (i.e., energetically favourable stems) by default for each strand. Considering all possible combinations of these structures, DinoKnot creates up to 400 sequence-structure combinations for the two strands. For each sequence-structure combination, DinoKnot (1) finds a pseudoknot-free structure that when combined with the input structure provides the minimum free energy structure given the input structure; (2) explores iteratively adding and removing base pairs to and from the input structure in search of lower energy structures than that found in part (1). To achieve this, DinoKnot follows four methods (all biologically sound and similar to the underlying methods of Iterative HFold [2]). Following these steps, DinoKnot finds multiple structures (sorted by their free energy) for the interacting structures. The output structure (in dot-bracket format) is the minimum free energy structure among this set of structures. Note that DinoKnot's output structure is guided by

the originally found energetically favourable stems, and may not include the input structure for the given sequences.

DinoKnot employs the Andronescu et al. energy parameters of HotKnots V2.0 [62] for RNA structures and MultiRNAFold energy parameters [31, 63] for DNA structures. Energy parameters for perfect hybrid stacks (DNA/RNA) were obtained from [64, 65]. Similar to the work of Lorenz *et al.* [14], the energy parameters for loops formed between an RNA and a DNA molecule are estimated to be the average of similar loops formed intramolecularly in an RNA and a DNA molecule.

DinoKnot is freely available on Github at `https://github.com/HosnaJabbari/DinoKnot`

3.1.1 How to Use DinoKnot

After installation and configuration of DinoKnot, the user can run the DinoKnot MultiModel with the required arguments as highlighted in Fig 3.2. DinoKnot takes two sequences (–s1 and –s2) as input, along with their nucleic acid type (–t1 and –t2) to specify the required energy parameters. Optionally, the user can provide the pseudoknot-free structure input for –s1/–s2 if that information is available to guide the prediction.

An example of running DinoKnot with two interacting nucleotides is shown in Fig 3.3. In this example, –s1 is the RNA sequence of a section of the SARS-CoV-2 genome and –s2 is a qRT-PCR reverse primer. The resulting output Seq_0 states the sequence of –s1 and –s2 concatenated with a linker, as was demonstrated in Fig 3.1. Restricted_0 states the most energetically favourable stems for –s1 and –s2. Result_0 states the minimum free energy interaction structure in dot-bracket notation for –s1 and –s2. Finally, Energy_0 is the free energy of the structure.

The resulting interaction structure can then be visualizing by inputting Seq_0 and Result_0 into VARNA, as presented in Fig 3.4. The –s1 and –s2 sequences are concatenated like in Fig 3.1, with the qRT-PCR primer (–s2) highlighted in red and the complementary sequence on the SARS-CoV-2 genome (–s1) highlighted in green. In this example, the qRT-PCR reverse primer binds completely to the expected binding site in green, as shown through the arcs connecting the red and green highlighted regions.

```
#### How to use:
    Arguments:
        DinoKnot:
            --s1 <sequence1>
            --r1 <restricted_structure1>
            --s2 <sequence2>
            --r2 <restricted_structure2>
            --t1 <type_for_sequence1>
            --t2 <type_for_sequence2>

        Remarks:
            Required arguments:
            1. --s1 <sequence1>, --s2 <sequence2>, --t1
<type_for_sequence1>, --t2 <type_for_sequence2>



    Sequence requirements:
        containing only characters GCAUT

    Structure requirements:
        -pseudoknot free
        -containing only characters ._(){}[]
        Remarks:
            Restricted structure symbols:
                () restricted base pair
                _ no restriction

    Type options:
        DNA
        RNA
```

Figure 3.2: **DinoKnot argument specifications.**

```
Command Line INPUT:


./DinoKnotMultimodel --s1

“GUGGGGGACAACCAAUCACUAAUUGUGUUAAGAUGUUGUGUACACACACUGGUACUGGUCAGGCAAUAACAGUUACACCGGAAGCCAAUAUGGAUCAAGAAU

CCUUUGGUGGUGCAUCGUGUUGUCUGUACUGCCGUUGCCACAUAGAUCAUCCAAAUCCUAAAGGAUUUUGUGACUUAAAAGGUAAGUAUGUACAAAUACCUAC

AACUUGUGCUAAUGACCCUGUGGGUUUUACACUUAAAAACACAGUCUGUACCGUCUGCGGUAUGUGGAAAGGUUAUGGCUGUAGUUGUGAUCAACUCCGCGAA

CCCAUGCUUCAGUCAGCUGAUGCACAAUCGUUUUUAAACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGU

AUACAGGGCUUUUGACAUCUACAAUGAUAAAGUAGCUGGUUUUGCUAAAUUCCUAAAAACUAAUUGUUGUCGCUUCCAAGAAAAGGACGAAGAUGACAAUUUA

AUUGAUUCUUACUUUGUAGUUAAGAGACACACUUUCUCUAACUACCAACAUGAAGAAACAAUUUAUAAUUUACUUAAGGAUUGUCCAGCUGUUGCUAAACAUG

ACUUCUUUAAGUUUAGAAUAGACGGUGACAUGGUACCACAUAUAUCACGUCAACGUCUUACUAAAUACACAAUGGCAGACCUCGUCU” —s2

“ACGATTGTGCATCAGCTGA” ——t1 RNA ——t2 DNA


OUTPUT:


Seq:

GUGGGGGACAACCAAUCACUAAUUGUGUUAAGAUGUUGUGUACACACACUGGUACUGGUCAGGCAAUAACAGUUACACCGGAAGCCAAUAUGGAUCAAGAAUC

CUUUGGUGGUGCAUCGUGUUGUCUGUACUGCCGUUGCCACAUAGAUCAUCCAAAUCCUAAAGGAUUUUGUGACUUAAAAGGUAAGUAUGUACAAAUACCUACA

ACUUGUGCUAAUGACCCUGUGGGUUUUACACUUAAAAACACAGUCUGUACCGUCUGCGGUAUGUGGAAAGGUUAUGGCUGUAGUUGUGAUCAACUCCGCGAAC

CCAUGCUUCAGUCAGCUGAUGCACAAUCGUUUUUAAACGGGUUUGCGGUGUAAGUGCAGCCCGUCUUACACCGUGCGGCACAGGCACUAGUACUGAUGUCGUA

UACAGGGCUUUUGACAUCUACAAUGAUAAAGUAGCUGGUUUUGCUAAAUUCCUAAAAACUAAUUGUUGUCGCUUCCAAGAAAAGGACGAAGAUGACAAUUUAA

UUGAUUCUUACUUUGUAGUUAAGAGACACACUUUCUCUAACUACCAACAUGAAGAAACAAUUUAUAAUUUACUUAAGGAUUGUCCAGCUGUUGCUAAACAUGA

CUUCUUUAAGUUUAGAAUAGACGGUGACAUGGUACCACAUAUAUCACGUCAACGUCUUACUAAAUACACAAUGGCAGACCUCGUCUXXXXXACGATTGTGCAT

CAGCTGA


Restricted_0:
_____________________________________________________________________________________________________________
_____________________________________________________________________________________________________________
_____________________________________(((((((_____________)))))))_____________________________________________
_____________________________________________________________________________________________________________
_____________________________________________________________________________________________________________
_____________________________________________________________________________________.....__(((_____)))
)_____


Result_0: [[[[[.[[[[.[....[[[.....]]]....].]]]].[[[[[[....[[[[.[[[[[..[[[.......]]]..]]]]]..]]]]....
[[[.......]]]...[[[[[[[[[..........]]]]]]]]].[[[[[....[[.[[[[..
[[[[[[....]]]]]].]]]]].....]]]]]...]]]]]]....]]]]].............[[[...[[[[[[[[[.............[[[[[[...
[[[.[[[[[.....]]]]]..]]]...]]]]]][[[[[....]]]]]....]]]]]]]]]..]]].[[[[[[[[[[[[[[[[[[....[.
[[[[[[[[[[(((((((((..........)))))))))].[[[[[[[[[[.........]]]].]]]]]]..[[[.[[[[[...
[[.......]]..]]]]]..]]][[[[[.............]]]]].[[[[[[[[[[...[[[........]]]....]]]]]]]]]].[[[[[.[[[[[.....
[[[[[[[.[[[[.......]]]]]]]]]].......]]]]].]]]]].........[[[[[[[[[..[[[....[[[[....]]]].]]].]]]]]]]]]
[[[[...[[[[[.[[[[.[[[[[.......]]]]].]]]].]]]]]..]]]].........]]]]]]].]]].]..... ]]]]]]]]]]]]]]]]]]]]


Energy_0: -138.540000
```

Figure 3.3: **Example DinoKnot Input and Output.** Sequence 1 (–s1) and Sequence 2 (–s2) are highlighted in purple and red, respectively.

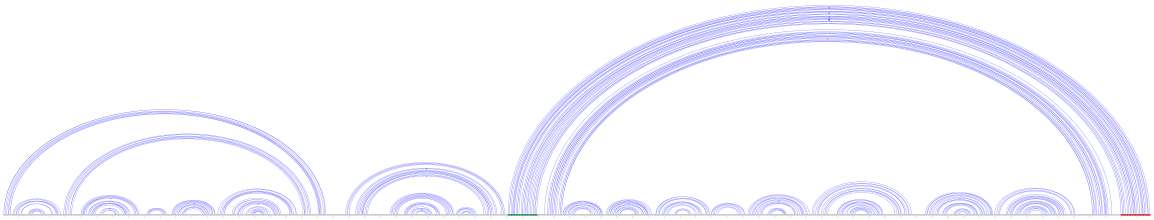

Figure 3.4: **VARNA [4] visualization of the resulting example DinoKnot output.** The qRT-PCR reverse primer (–s2) is highlighted in red, and the expected binding site (the complementary sequence) on the SARS-CoV-2 genome (–s1) is highlighted in green. This structure is pseudoknot-free.

Chapter 4

Materials and Methods

In this chapter, the system and experiment set up of the analyses applying DinoKnot are described.

4.1 qRT-PCR interaction structures

Vogels *et al.* compared the analytical efficiencies and sensitivities of the nine primer-probe sets used to detect COVID-19 in early 2020 [1]. Dinoknot can predict the structure of the interaction between the reverse primer and the SARS-CoV-2 genome. The predicted interaction structure can determine whether the RNA structure of the SARS-CoV-2 genome may affect the binding of the reverse primers in the qRT-PCR assay and whether this correlates qualitatively to the analytical efficiencies and sensitivities shown experimentally by Vogels *et al.* [1]. The focus is narrowed to the DNA-RNA interactions that occur during this process by studying how the reverse primer interacts with the RNA genome. Therefore, the interaction of the forward primer and the probes with the SARS-CoV-2 genome is not investigated because these oligonucleotides are involved in binding to the cDNA, not the positive sense RNA genome.

The SARS-CoV-2 reference genome NC_045512.2 was obtained from the National Center for Biotechnology Information GenBank database [66]. The sequence of the RNA transcripts used by Vogels *et al.* [1] to determine the analytical efficiencies and sensitivities of the primer-probe sets were input into the program Iterative HFold [2] to predict their individual secondary structures. This output is referred to as the secondary structure of the transcript prior to the interaction with the reverse primer. The

primer sequences were obtained from the list of World Health Organization (WHO) protocols to diagnose COVID-19 [67]. The locations where the primers bind on the reference genome NC_045512.2 and the corresponding RNA transcript area are stated in Table 4.1.

Table 4.1: **Primer binding location on the SARS-CoV-2 reference genome NC_045512.2**. The transcript location is the area from the reference genome input into DinoKnot to predict the reverse primer/SARS-CoV-2 RNA genome interaction structure. The transcript locations are based on lab protocols used by Vogels *et al.* [1].

Gene	Transcript location	Primer	Primer location
nsp10	13,122 - 13,825	CCDC-ORF1-F	13,342 - 13,362
		CCDC-ORF1-R	13,442 - 13,460
RdRp	15,094 - 15,976	RdRp-SARSr-F	15,431 – 15,452
		RdRp-SARSr-R	15,491 - 15,517
nsp14	18,447 - 19,294	HKU-ORF1-F	18,778 - 18,797
		HKU-ORF1-R	18,889 - 18,909
Envelope (E)	26,207 - 27,116	E-Sarbeco-F	26,269 - 26,294
		E-Sarbeco-R	26,360 - 26,381
Nucleocapsid (N)	28,068 - 29,430	CCDC-N-F	28,881 – 28,902
		CCDC-N-R	28,958 - 28,979
		HKU-N-F	29,145 - 29,166
		HKU-N-R	29,145 - 29,166
		2019-nCoV_N1-F	28,287 - 28,306
		2019-nCoV_N1-R	28,335 - 28,358
		2019-nCoV_N2-F	29,164 - 29,183
		2019-nCoV_N2-R	29,213 - 29,230
		2019-nCoV_N3-F	28,681 - 28,702
		2019-nCoV_N3-R	28,732 - 28,752

The whole RNA transcript region stated in Table 4.1 and the corresponding reverse primer were input into DinoKnot to determine the secondary interaction structure. Two of the reverse primers, HKU-ORF1-R and RdRp-SARS-R, contain degenerate bases R and S which means there are a mixture of oligonucleotides that contain different bases at that position [68]. This means that an A or G may be present in the position of the R degenerate base and a C or G may be present in the position of the S degenerate base. In these cases, all possible degenerate base combinations were predicted with DinoKnot. The dot-bracket output was visualized using VARNA [4] in arc format in which the RNA backbone is represented by a horizontal line and base pairs are presented as arcs that connect the two bases.

4.2 Comparing DinoKnot to RNAcofold

In order to compare DinoKnot to an existing tool, RNAcofold from the ViennaRNA package version 2.1.9h was used to predict the qRT-PCR interaction structures of

the 5 gene areas from the SARS-CoV-2 reference genome targeted by the reverse primers from Table 4.1. RNAcofold was chosen over RNAduplex since RNAduplex only considers the structure after hybridization of the two RNA/DNA sequences [36]. RNAcofold was used with the "–noconv" parameter, to not convert T to U's when considering the DNA molecule. The RNA sequence (the gene region) was inputted first, and the DNA sequence (the reverse primer) was inputted last. The resulting structure in dot-bracket output was visualized using VARNA [4].

4.2.1 Interaction involving a pseudoknotted structure

In order to compare the performance of DinoKnot and RNAcofold [14], an experimentally studied interaction involving a pseudoknot was investigated since the SARS-CoV-2 qRT-interaction sites did not involve any known pseudoknotted structures. An example of an interaction involving a pseudoknotted structure is between the human immune-functioning C-C chemokine receptor 5 (*CCR5*) mRNA and miR-1224. To investigate the interaction structure between the *CCR5* mRNA and miR-1224, the sequences were obtained from Belew *et al.* [38] and input into DinoKnot and RNAcofold [14]. The structure of *CCR5* mRNA prior to interaction was compared between Iterative Hfold [2] and RNAfold [5].

4.3 Mutations

Vogels *et al.* listed mutations in the primer/probe binding area of the SARS-CoV-2 genome that occur at a frequency of greater than 0.1% [1]. To predict if these mutations affect primer/probe binding, and thus the sensitivity of SARS-CoV-2 detection, the mutated sequence of the transcript along with the affected primer/probe was entered into DinoKnot. The RNA/DNA setting was used to predict the interaction between the mutated RNA transcript and the reverse primer. The DNA/DNA setting in DinoKnot was used to predict the interaction structure between the mutated cDNA transcript and the corresponding primer/probe. The dot-bracket output was visualized using VARNA [4] in arc format.

4.3.1 Variants of Concern

To investigate the effect of variants of concern on the reverse primer binding ability, a complete genome from each of the B.1.1.7 (Alpha), B.1.617.2 (Delta) and B.1.1.529

(Omicron) lineages were obtained from the National Center for Biotechnology Information GenBank database, (accession ID: MW487270.1 (Alpha), OK091006.1 (Delta) and OL672836.1 (Omicron)) [66]. A genome from the B.1.315 (Beta) lineage was obtained from the GISAID database (accession ID: EPI_ISL_860693) [69]. BLASTn [66] was used to align the 5 gene areas of the RNA transcripts from the NC_045512.2 reference genome against the MW487270.1, EPI_ISL_860693, OK091006.1 and OL672836.1 sequences to detect any mutations that occurred in these gene areas. To predict if these changes may affect primer binding, the gene areas of the variants with mutations, along with the affected primer, was input into DinoKnot.

4.3.2 Clinical report of variant causing N gene detection issues

Laine *et al.* reported a variant strain from the B.1.1.318 Pango lineage with mutations in the N gene region that caused this variant to not be detected by the CCDC-N primer-probe set [70]. The variant has three base pair mutations and three deletions in the region where the CCDC-N-F primer binds, as noted by Laine *et al.* to likely explain the detection issue [70]. To determine if DinoKnot would be able to predict any changes to the interaction structure as a result of the mutations, the sequence for the hCoV-19/Finland/FinD796H/2021 strain was obtained from GISAID (accession ID: EPI_ISL_1061414) [70, 69]. The cDNA of the N gene region and the CCDC-N-F primer were input into DinoKnot and the resulting dot-bracket output structure was visualized using VARNA [4].

4.4 miRNA - SARS-CoV-2 interaction structures

In order to assess DinoKnot's performance in other applications, I investigated the area of miRNA target site prediction. McDonald *et al.* [7] used the tool miRanda [6] to predict potential miR-2392 binding sites to the SARS-CoV-2 genome. DinoKnot was used to predict the nucleic interaction structure of miR-2392 and its target sites on the SARS-CoV-2 genome to determine how the intramolecular structure may affect the interaction. The nucleotide sequences of the top three binding sites for miR-2329 to the SARS-Cov-2 reference genome and the miR-2392 sequence was obtained from McDonald *et al.* [7]. The target site locations on the NC_045512.2 reference genome are 2130-2149bp (*NSP2*, *ORF1ab*), 7153-7172bp (*NSP3*, *ORF1ab*) and 26326-26346bp

(*E* gene) [7]. Each target site, along with the miR-2392 sequence was input into DinoKnot. As well, each target site, plus a 100bp flanking region added on either side was input into DinoKnot to take into account how the flanking region may affect the interaction structure. The dot-bracket output was visualized using VARNA [4].

Chapter 5

Results

In this chapter I present the interaction structure results of the SARS-CoV-2 reference genome with the nine reverse primers from the primer-probe sets. I then compare DinoKnot's predicted interaction structures to those predicted by RNAcofold [14] and give an example of an interaction structure to highlight the importance of considering pseudoknotted structures. I explore the effect of mutations on the interaction structure which may reduce the efficacy of these primer-probe sets and investigate a clinical report of a variant SARS-CoV-2 strain causing detection issues. Finally, I discuss the application of DinoKnot to miRNA target site prediction to consider the effect of intramolecular structures prior to interaction.

5.1 qRT-PCR Interaction Structures

DinoKnot predicted the interaction structures discussed in this section to occur between the reverse primer and the corresponding transcript locations stated in Table 4.1. Fig 5.1 and Fig 5.2 represents the interaction site only, with the reverse primer highlighted in red and the target sequence highlighted in green. All interaction structures are pseudoknot-free, except for the HKU-N-R and E-Sarbeco-R primers presented in Fig 5.2.

As confirmation of stability of primer binding, I further calculated the net free energy of primer binding, $\mathcal{E}$, as follows based on the energy values shown in Table 5.1. (Note: the free energy is reported in kcal/mol rather than kJ/mol for consistency with the outputs from the RNA secondary structure prediction programs).

$$\mathcal{E} = \Delta G_{intermolecular} - \sum \Delta G_{intramolecular}$$

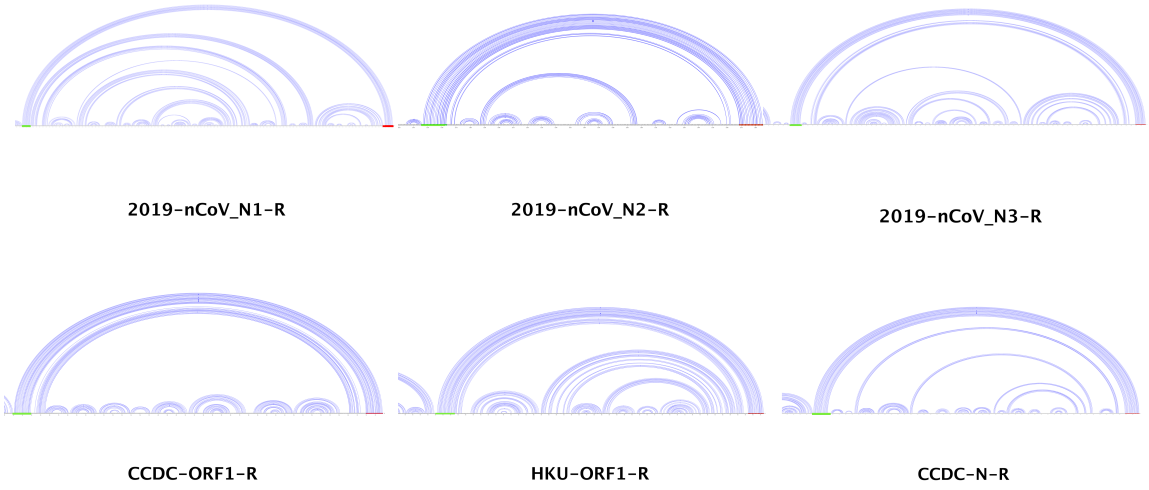

Figure 5.1: **Interaction structures predicted by DinoKnot of the SARS-CoV-2 transcript gene area targeted by the reverse primer.** The expected target region of the reverse primer is highlighted in green and the reverse primer sequence is highlighted in red.

The net free energy of primer binding is negative (i.e., binding is stable) in most cases except HKU-N-R when the primer is allowed to form intramolecular structure (i.e., at 37°C). In addition, the net free energy of binding is negligible in the case of E-Sarbeco-R at 37°C.

The efficiency of a primer refers to the fraction of cDNA transcript molecules that are copied during an amplification cycle, which is ideally 100% [71]. Vogels *et al.* experimentally found that all of the primer-probe sets had comparable analytical efficiencies that were all above 90% [1]. All primer-probe sets had comparable analytical sensitivities with a limit of detection of $100,000$ SARS-CoV-2 viral copies/mL, except for the RdRp-SARSr set, which had the lowest sensitivity [1].

All of the reverse primers were predicted by DinoKnot to interact with their expected target region. The 2019-nCoV_N1-R, 2019-nCoV_N3-R, CCDC-N-R and CCDC-ORF1-R primers fully paired to their target region as shown in Fig 5.1, where the base pair arcs connect the primer highlighted in red to the complementary sequence on the SARS-CoV-2 genome highlighted in green. This prediction agrees with the analytical efficiency and sensitivity results found by Vogels *et al.* [1].

Partial reverse primer mismatching: The HKU-ORF1-R and 2019-nCoV_N2-R primers in Fig 5.1 were predicted to pair to their target region with single base mismatches. The term mismatch or primer mismatch is used to mean that the primer

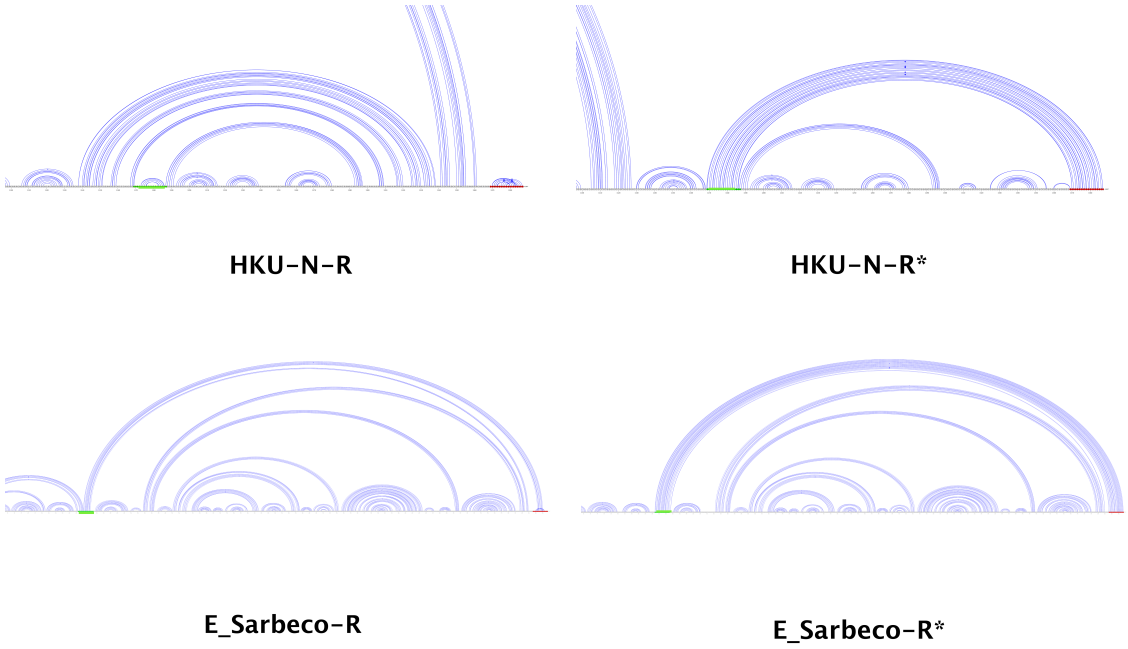

Figure 5.2: **Interaction structures predicted by DinoKnot of the SARS-CoV-2 transcript gene area targeted by the reverse primer.** The expected target region of the reverse primer is highlighted in green and the reverse primer sequence is highlighted in red. The E-Sarbeco-R* and HKU-N-R* primer structures were input into DinoKnot as unfolded to simulate the primer structure during the 95°C denaturation step of the qRT-PCR assay due to primer mismatch prediction under default conditions of 37°C.

Table 5.1: **Energies (kcal/mol) of RNA transcript, reverse primer and reverse primer/transcript interaction structures.** The transcript energy and primer energy is the minimum free energy (MFE) of the transcript and primer structures before the interaction. The DinoKnot interaction MFE is the energy of the interaction structure of the RNA transcript and reverse primer. The transcript minimum free energies were predicted by Iterative HFold [2], the primer minimum free energies were predicted by Simfold [3], and the interaction structure free energies were predicted by DinoKnot. The reported net free energy is the interaction structure MFE (intermolecular) minus the transcript and reverse primer energies combined (intramolecular).

Primer	Transcript MFE (kcal/mol)	Primer MFE (kcal/mol)	DinoKnot Interaction MFE (kcal/mol)	Net Free Energy (kcal/mol)
2019-nCoV_N1-R	-267.24	-5.90	-277.67	-4.53
2019-nCoV_N2-R	-267.24	0	-276.06	-8.82
2019-nCoV_N3-R	-267.24	-3.75	-272.3	-1.31
CCDC-N-R	-267.24	-1.30	-274.67	-6.61
CCDC-ORF1-R	-133.39	0.2	-138.54	-5.35
HKU-ORF1-R	-138.89	-3.10	-151.76	-9.77
HKU-N-R	-267.24	0*	-280.82	-13.58
		-3.30	-267.91	2.63
RdRP-SARSr-R	-150.03	-3.70	-162.37	-13.64
E-Sarbeco-R	-148.74	0*	-167.39	-18.65
		-1.50	-150.48	-0.24

*E-Sarbeco and HKU-N-R primer structure forced unfolded for primer to bind to targeted area.

did not bind to the expected region or expected base pair. A single base pair mismatch would be unlikely to affect the qRT-PCR test and the analytical efficiencies and sensitivities of these primer-probe sets were comparable to the other primer-probes tested experimentally [1].

HKU-N-R and E-Sarbeco-R: The HKU-N-R and E-Sarbeco-R primers did not pair as expected to their target region when DinoKnot was not given an input structure (i.e., when both strands were free to assume possible structures before interacting with one another), as shown in Fig 5.2. The HKU-N-R primer was predicted to bind to itself, rather than to its target region. When the HKU-N-R primer structure was forced to have no structure prior to interaction (ie. to be completely unfolded), the first base of the HKU-N-R primer at the 5' end did not bind to its expected nucleotide but the rest of the primer interacted with the target region as expected. Forcing the primer to be unfolded is a prediction of the primer structure after the qRT-PCR test denaturation step at 95°C in the work done by Vogels *et al.* [1]. The predicted interaction structure for the E-Sarbeco-R primer when DinoKnot was not given an input structure resulted in the bases at positions 8-15 in the reverse primer binding

as expected to the target region. However, bases 5-7 paired with bases 16-18 and the remaining bases of the 22bp primer remained unpaired. When the primer structure was inputedt into DinoKnot with no structure prior to interaction, as was done for the HKU-N-R primer, the E-Sarbeco-R primer fully paired to its target region. Both the HKU-N-R and E-Sarbeco primer-probe sets were shown experimentally by Vogels *et al.* [1] to have analytical efficiencies and sensitivities that are comparable to the other primer-probe sets. Therefore, the structures predicted with the HKU-N-R and E-Sarbeco primers having no structure prior to interaction are the more likely interaction structures since the primer is capable of binding to its target region which is required for successful amplification and subsequent positive COVID-19 test.

A limitation in the study of the qRT-PCR interaction structures is that DinoKnot's energy parameters are determined from experimental evidence performed at 37°C. During the qRT-PCR test, the temperature is changed between the denaturation, annealing and extension cycles. At the denaturation step, in this case when the temperature is 95°C, it is unlikely for the nucleic acid molecules, especially the short oligonucleotide primers, to retain their structure. At high temperatures (approximately above 60°C depending on the sodium concentration), a nucleic acid will be in a single stranded (linear) state with no structure [61]; this is why I have chosen to force the primer structure as unfolded in the cases of the HKU-N-R and E-Sarbeco-R primer mismatching under the default conditions. In the case of qRT-PCR, the effect of higher temperatures on the nucleic acid structure needs to be considered before concluding a predicted mismatch. Despite this limitation for the case of qRT-PCR testing, I hypothesize that if the primers are capable of binding to their expected target site, or are designed to be stable, at a lower temperature of 37°C, then binding will still be likely favourable at the higher annealing temperature (in this case 55°C) because at higher temperatures, there will be less intramolecular structure capable of forming to compete for primer binding.

RdRp-SARSr-R: The RdRp-SARSr-R primer contains two degenerate bases, R and S. This means that an A or G may be present in the position of the R degenerate base and a C or G may be present in the position of the S degenerate base. All possible combinations were given as input to DinoKnot and the predicted structures are shown in Fig 5.3, which represents the interaction site only.

The primer with the degenerate base combination that was predicted to have complete binding between the primer and its target site was when an A was input at the R position and a G was input at the S position (R = A with S = G). One base

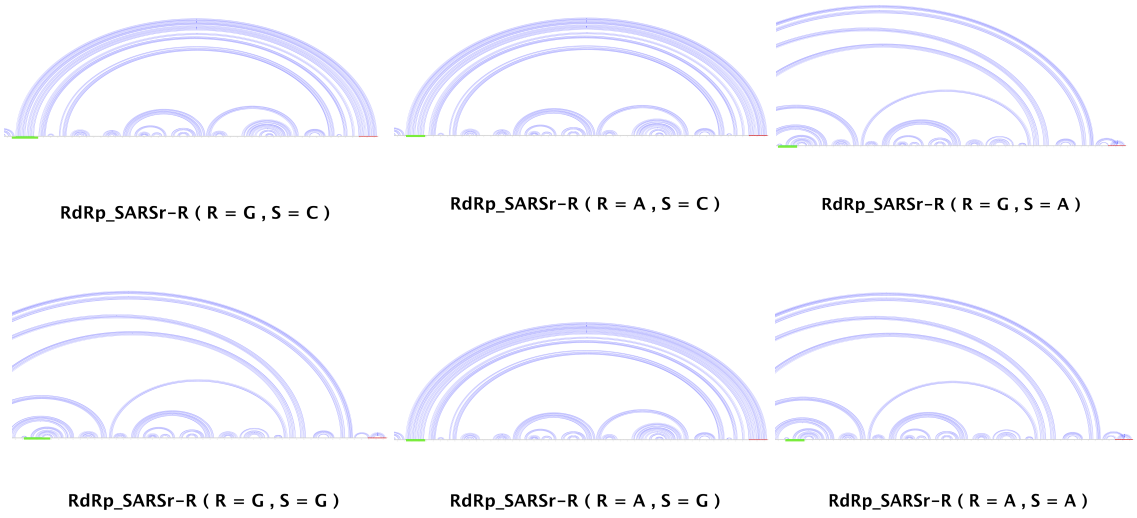

Figure 5.3: **Interaction structures of the RdRp-SARS-R primer predicted by DinoKnot with all of the possible base combinations from the degenerate primers.** The RdRp-SARS-R primer contains the degenerate bases R and S, which means an A or G may be present at the R position and a C or G may be present at the S position. All possible combinations were predicted. The S position was also input as an A to predict if this change may increase primer sensitivity.

pair mismatch was predicted when R = A or G with S = C. When R = G with S = G, DinoKnot predicted the primer to not bind to its expected binding site. During the qRT-PCR testing by Vogels *et al.*, the RdRp-SARSr primer-probe set had 6-10 Ct values higher than all other sets and had the lowest analytical sensitivity [1]. A lower sensitivity means that a primer-probe set may not be able to detect SARS-CoV-2 in patients with a low viral load. Vogels *et al.* proposed that changing the degenerate base S to an A in the reverse primer could increase the sensitivity of the primer-probe set [1]. To test this, the primer was given as input to DinoKnot with an A input at the S position. All combinations where S was input as an A resulted in the primer not binding to the target area. The result was the primer binding to itself and partial binding to the 5' end of the transcript, over 400 bases downstream from the target area. Therefore, changing the S to an A is unlikely to increase the sensitivity of the primer if considering the minimum free energy interaction structure predicted by DinoKnot. Based on these results, I hypothesize that the low sensitivity issue of the RdRp-SARSr primer-probe set is due to the predicted primer mismatch

when a G is present in the R and S position of the primer. If this mismatch (when R = G with S = G) were to occur, it would lower the RdRp-SARSr-R concentration in the primer-probe set that is capable of binding to its target region. Although the concentration of each base combination at the degenerate R and S positions is not stated in the protocol, if the assumption is made that the four possible combinations are present in equal amounts, this may explain the low sensitivity issue of the RdRp primer-probe set. Therefore, I suggest that changing the base at the R position to an A and the base at the S position to a G may increase the RdRp-SARSr primer-probe set sensitivity since this base combination was predicted to completely bind to its target region.

5.2 Comparing DinoKnot to RNAcofold

To compare DinoKnot's qRT-PCR interaction structure results to those of an existing tool, the interaction structures of the same nine reverse primers and their corresponding gene regions were predicted by RNAcofold [14]. The interaction structures predicted by RNAcofold are presented in Fig 5.4 and Fig 5.5.

The structure comparison is focused on the interaction site between the reverse primer and the gene transcript. DinoKnot and RNAcofold [14] predicted the same primer binding for the 2019-nCoV-N2-R, CCDC-N-R, HKU-ORF1-R and E-Sarbeco-R primers to their respective gene regions. There are minor base pair differences in the 2019-nCoV-N1-R, 2019-nCoV-N3-R and HKU-N-R primers.

HKU-ORF1-R: For the HKU-ORF1-R primer that contains a degenerate base R, DinoKnot predicts complete binding between the primer and its target site and RNAcofold predicts a single base pair mismatch at the 5' end of the primer when the degenerate base is predicted as an A. However, when the degenerate base R is predicted to be a G, RNAcofold predicts that the primer does not bind to its expected binding site, resulting in a complete primer mismatch. DinoKnot only predicts a single base pair mismatch. The interaction structure predicted by RNAcofold is unlikely since the HKU-ORF1 primer-probe set was shown by Vogels *et al.* [1] to have high analytical efficiency and sensitivity. The degenerate base in this reverse primer means that there is a certain concentration of primers with an A at this position and a certain concentration with a G at this position. The exact concentration is unspecified, but if the primers with a G at this position were unable to bind to the expected binding site, the primer-probe set would be expected to have a lower analytical efficiency and

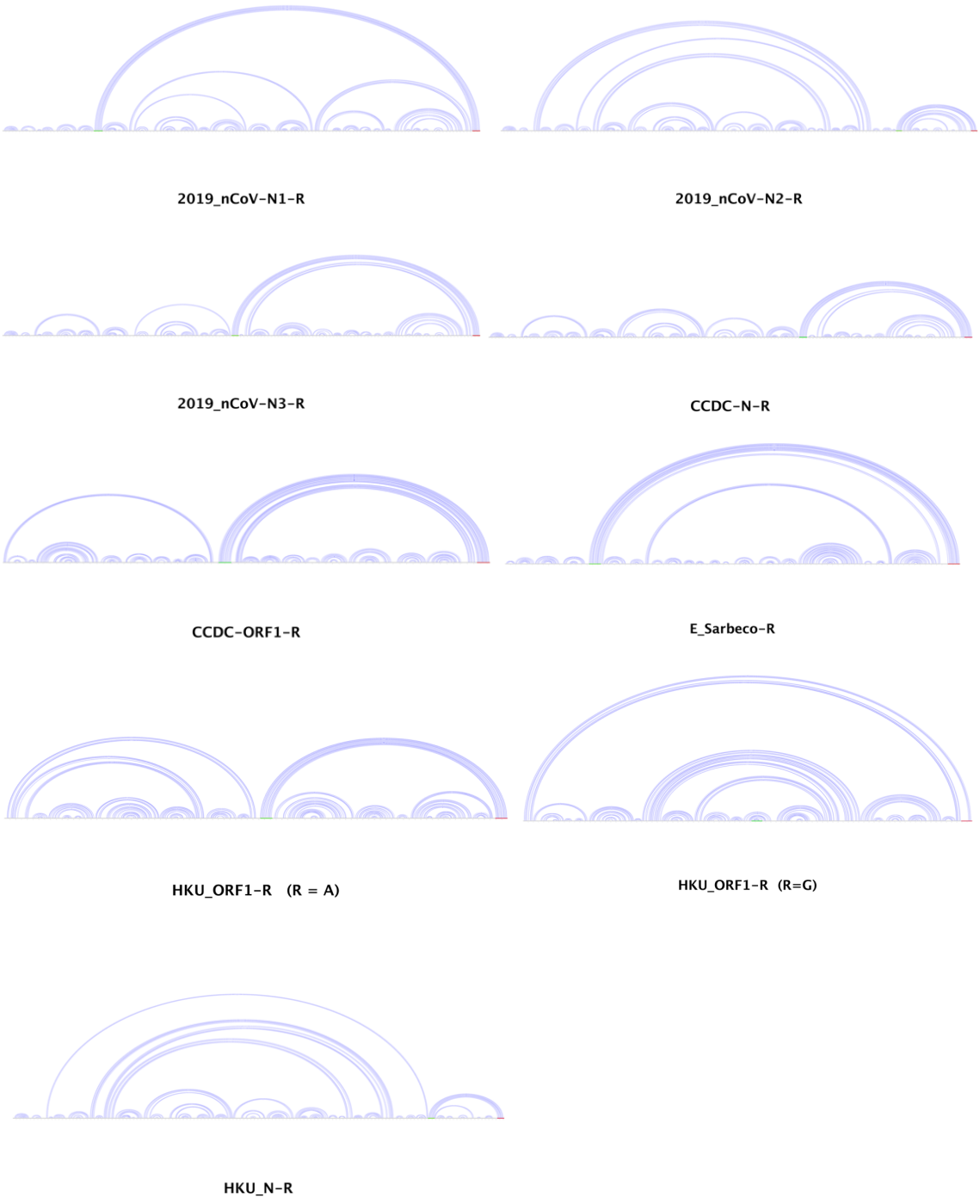

Figure 5.4: **qRT-PCR interaction structures predicted by RNAcofold.**

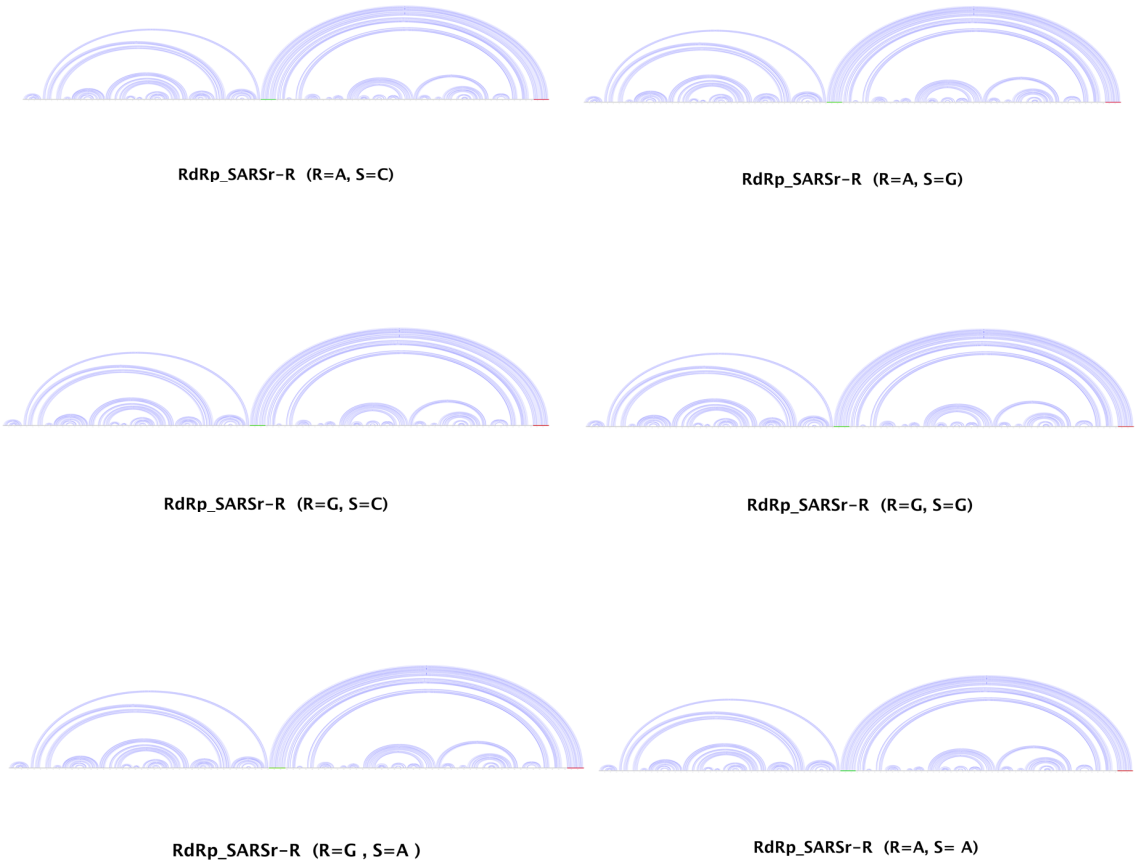

Figure 5.5: **Interaction Structures predicted by RNAcofold for the RdRp-SARSr-R primer.**

sensitivity than what was observed by Vogels *et al.* [1]. Therefore, the interaction predicted by DinoKnot for this reverse primer would be the more likely interaction structure.

RdRp-SARSr-R: For the RdRp-SARSr-R primer, DinoKnot and RNAcofold predicted the same primer binding in the cases where the degenerate primer R = A with S = G, and R = A or G with S = C. DinoKnot and RNAcofold predict different primer binding when R = G with S = G. RNAcofold predicts complete primer binding whereas DinoKnot predicts complete primer mismatching. As discussed, the RdRP-SARSr primer-probe set was shown by Vogels *et al.* [1] to have a lower analytical sensitivity compared to the other primer-probe sets. The interaction structure with the primer mismatch predicted by DinoKnot may give a structural indication for the lower sensitivity, which is not shown by the RNAcofold interaction structure.

When predicting the degenerate base S = A, as recommended by Vogels *et al.*, with R = A or G, RNAcofold predicted complete primer binding and DinoKnot predicted complete primer mismatching. Since the primers with S=A are not experimentally validated, there is no evidence to support the structures predicted by RNAcofold or DinoKnot to be able to make a comparison. However, the structures predicted by RNAcofold would agree with the suggestion by Vogels *et al.* [1] to replace the S degenerate base with A to increase the RdRP-SARSr primer-probe set sensitivity.

Overall, DinoKnot predicted primer mismatches not predicted by RNAcofold which are supported more by the experimental analytical efficiencies and sensitivities. The primer mismatch predicted by RNAcofold was not supported by the analytical sensitivities.

5.2.1 Interaction involving a pseudoknotted structure

In Section 5.2 it was shown that DinoKnot is comparable to RNAcofold in predicting primer binding to the expected target area on the SARS-CoV-2 genome. In this subsection I aim to highlight the advantage of using DinoKnot, which is the ability to predict pseudoknotted structures. Using DinoKnot, the MFE interaction structure involving a pseudoknotted structure between the human immune-functioning C-C chemokine receptor 5 (*CCR5*) mRNA and miR-1224 can be predicted. *CCR5* is involved in various immune system functions and is also a co-receptor involved in the entry of the human immunodeficiency virus (HIV) into CD4+ T cells [72, 38, 73]. The *CCR5* mRNA has a two-stemmed pseudoknotted structure that was believed to be part of programmed -1 ribosomal frameshifting (-1 PRF) [38]. During -1 PRF, the pseudoknotted structure directs the ribosome to slip back one base during translation, resulting in mRNA decay in order to regulate gene expression [38]. There is experimental evidence showing that miR-1224 interacts with the pseudoknotted *CCR5* mRNA structure from in vitro electrophoretic mobility shift assays, as well as in live cells using an affinity capture assay [38]. This interaction is also supported by NMR-labelling data [73]. It was hypothesized that the interaction with miR-1224 works to enhance the -1 PRF signal in a suggested triplex RNA structure [38]. However, recent counter evidence has shown that this frameshifting does not occur for *CCR5* so the exact function of this interaction is uncertain [39]. Fig 5.6, shows the interaction structure between the *CCR5* mRNA and miR-1224 predicted by DinoKnot and RNAcofold. DinoKnot predicts a pseudoknotted structure and predicts

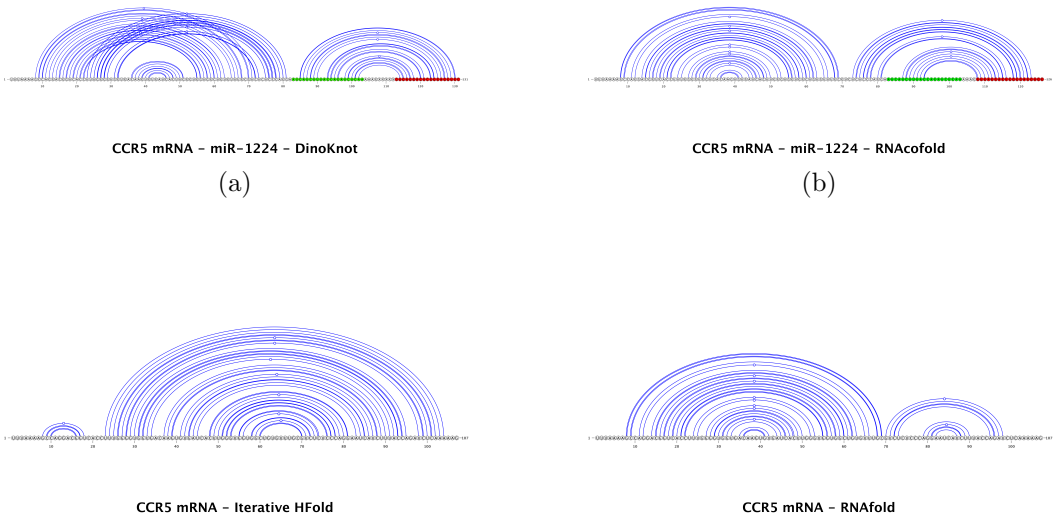

Figure 5.6: **Interaction structures of *CCR5* mRNA and miR-1224 predicted by a.) DinoKnot and b.) RNAcofold**. The miR-1224 sequence is highlighted in red and the putative binding site is highlighted in green. The secondary structure of the *CCR5* mRNA is predicted by c.) Iterative HFold [2] and d.) RNAfold [5] to show the predicted structure prior to interaction. DinoKnot predicts the miR-1224 interaction to stabilize a pseudoknotted structure.

miR-1224 to bind to one of the putative binding sites presented by Belew *et al.* [38].

RNAcofold predicts no pseudoknotted structure, as it is not designed to, and predicts partial binding of miR-1224 to one of the putative binding sites and partial binding upstream of this site on the *CCR5* mRNA. DinoKnot predicts binding one base upstream and one base downstream from a residue (Adenine (A) base) with observed dramatic chemical shifts indicating an interaction during NMR-labelling, whereas RNAcofold predicts binding 3 bases upstream from this residue [73]. When looking at the structure of the *CCR5* mRNA prior to interaction, neither Iterative Hfold [2] or RNAfold [5] predict a pseudoknotted structure (note: RNAfold [5] is not capable of predicting pseudoknotted structures). Iterative Hfold [2] and RNAfold [5] were chosen to predict the structure of the *CCR5* mRNA prior to interaction since these programs have the same underlying structure prediction methods as DinoKnot and RNAcofold, respectively.

DinoKnot predicts the interaction with miR-1224 to cause a pseudoknotted struc-

ture, which is consistent with the idea proposed by Belew *et al.*, i.e., that the interaction may work to stabilize a pre-existing structure since the interaction caused no differences in RNA modification patterns during selective 2'-hydroxyl acylation analysed by primer extension [38]; DinoKnot and Iterative HFold [2] predict that a pseudoknotted structure is not the MFE structure without the miR-1224 interaction. To investigate this hypothesis further, I compared the minimum energies of the structures. The interaction structure predicted by DinoKnot has a MFE of -29.45 kcal/mol. The Iterative Hfold MFE structure of *CCR5* has a free energy of -17.57 kcal/mol. HotKnots v2.0 [62] was used to compute the free energy (without dangling) of the pseudoknotted structure of the *CCR5* mRNA predicted by DinoKnot without the miR-1224 interaction (ie. the energy of the pseudoknotted structure and leaving the remaining bases of the *CCR5* mRNA as unpaired). The pseudoknotted structure predicted by DinoKnot was computed to have a free energy of -16.76 kcal/mol. The MFE of the *CCR5* mRNA structure predicted by Iterative Hfold [2] is only 0.75 kcal/mol more stable than the pseudoknotted structure (without the miR-1224 interaction) predicted by DinoKnot. DinoKnot predicts the interaction with miR-1224 to stabilize this structure with a MFE of -29.45 kcal/mol. The MFE of miR-1224 alone is predicted by Simfold [3] to be 0.0 kcal/mol. Therefore, the interaction is predicted to reduce the MFE by approximately 12 kcal/mol. A structure is considered significant when there is a stem of a minimum three base pairs.

Although neither DinoKnot or RNAcofold are capable of predicting the exact proposed structure, this example highlights a nucleic acid interaction involving a pseudoknotted structure that is appropriate for DinoKnot, but not for RNAcofold.

By this comparison in Section 5.2, I have shown that DinoKnot is comparable to an existing tool but has the added benefit of considering the intramolecular structure of both strands prior to interaction and is capable of predicting pseudoknotted structures. Since DinoKnot is capable of predicting more complex pseudoknotted structures, this could elicit functional information of the interaction and its flanking region.

5.3 Applications of DinoKnot

5.3.1 The effects of mutations on expected interactions

DinoKnot can be used to predict how mutations to the sequence of the SARS-CoV-2 viral genome may affect the ability of primers to bind to their expected target area.

Vogels *et al.* looked at 992 clinical samples and identified mutations in the expected primer binding regions that could decrease the primer sensitivity for SARS-CoV-2 detection [1]. I used DinoKnot to predict how those mutations affect the ability of the primer/probe to bind to its target sequence. Mutations in the expected binding region of the forward primer and probe were DNA/DNA interactions and mutations in the reverse primer expected binding region show both the RNA/DNA interaction and the DNA/DNA interaction. The list of mutations are stated in Table 5.2, along with the interaction structure energies.

The resulting structures (interaction site only) for all mutations can be found in Fig A.1 of the appendix.

Mutations that cause no change to the interaction structure: The mutations in the primer binding region of the CCDC-N-F at base pair positions 28,881 and 28,882 and the 2019-nCoV_N3-F mutation at base pair position 28,688 were predicted to have no affect on the primer binding ability.

Mutations that cause partial mismatching: The following mutations resulted in a single mismatch between the primer/probe and the mutated base: CCDC-ORF1-P at base position 13,402, 2019-nCoV_N1-R at base position 28,344 and CCDC-N-F at base position 28,883.

The mutation in the primer binding region of CCDC-ORF1-F at position 13,358 affects the ability of the 5 bases at the 3' end of the primer to bind. The remaining 16 out of 21 bases of the forward primer are still able to bind to their target nucleotides.

The mutation in the probe binding region of 2019-nCoV_N1-P at position 28,311 is predicted to cause a mismatch of the first three bases at the 5' end of the probe. The remaining 21 nucleotides of the probe still bind to their target sequence.

Mutations that cause complete mismatching: The following mutations prevent any binding to the target sequence: the mutations in the primer binding region of E-Sarbeco-R at position 26,370, HKU-N-F at position 29,148, and 2019-nCoV_N3-R mutation at position 28,739. The interaction structures for these mutations that result in a disruption of binding are shown in Fig 5.7.

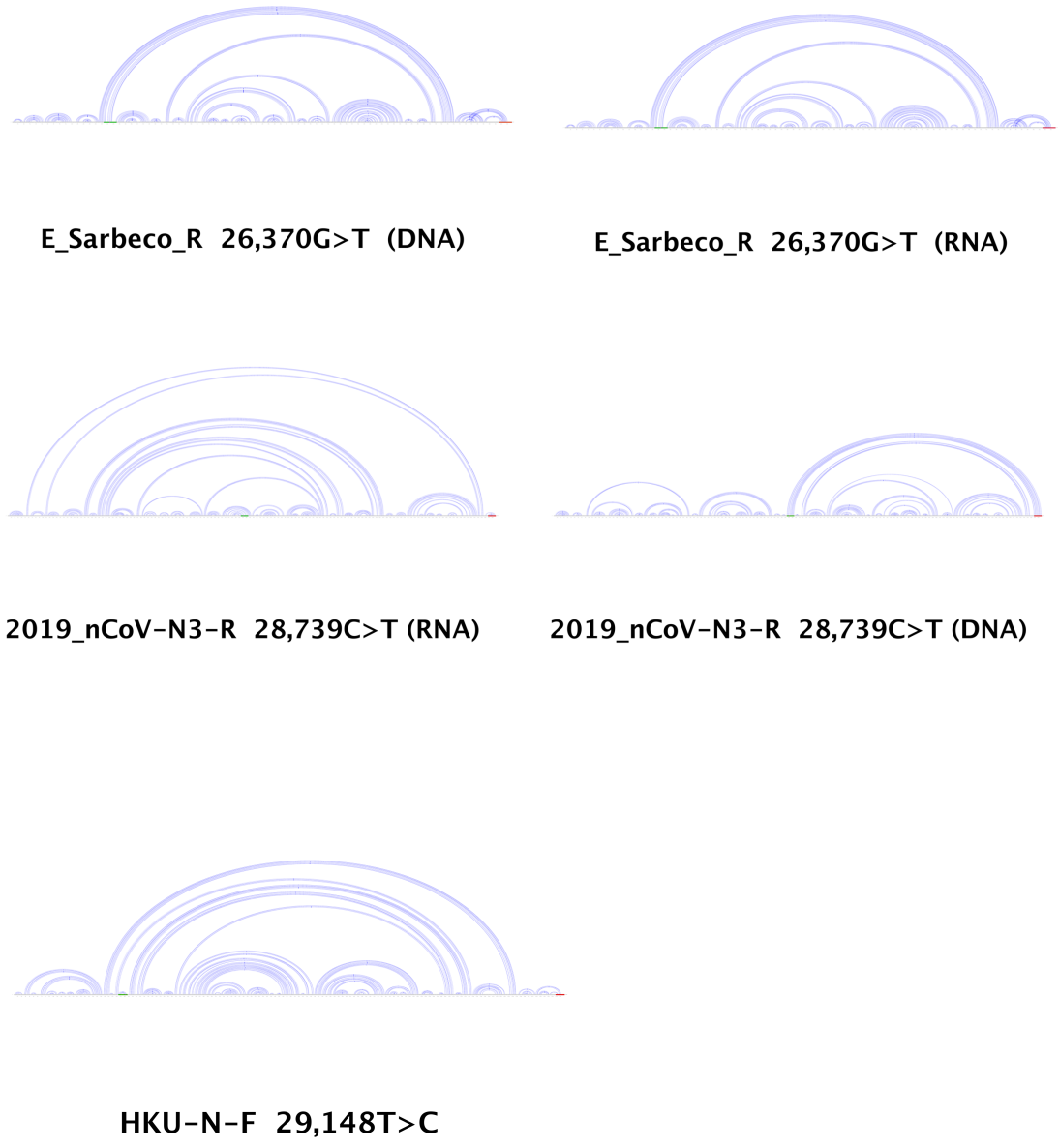

Figure 5.7: **Interaction structures of transcript regions containing mutations in the primer/probe binding region predicted by DinoKnot to disrupt primer/probe binding ability.** The expected target region is highlighted in green and the primer/probe sequence is highlighted in red. The mutated base is highlighted in pink.

Table 5.2: **Interaction structure energy differences predicted with mutations in the primer/probe binding region of the SARS-CoV-2 genome obtained from Vogels *et al.*** [**1**] . The interaction MFE was predicted by DinoKnot. The probe/transcript and forward primer/transcript interactions are DNA/DNA interactions since these oligonucleotides interact with the cDNA strands . Mutations in the reverse primer binding region include both DNA/DNA and DNA/RNA interaction since the reverse primer interacts with the SARS-CoV-2 genome along with the negative sense cDNA strand.

Primer-probe	Mutation in ref. genome	Mutation Position	DinoKnot Interaction MFE (kcal/mol)	Interaction type
CCDC-N-F	no mutation		-283.27	DNA/DNA
	G → A	28,881	-282.43	DNA/DNA
	G → A	28,882	-281.6	DNA/DNA
	G → C	28,883	-279.62	DNA/DNA
CCDC-ORF1-F	no mutation		-138.05	DNA/DNA
	C → T	13,358	-133.77	DNA/DNA
CCDC-ORF1-P	no mutation		-144.82	DNA/DNA
	T → G	13,402	-147.82	DNA/DNA
E-Sarbeco-R	no mutation		-167.39	RNA/DNA
	no mutation		-167.77	DNA/DNA
	G → T	26,370	-149.57	RNA/DNA
	G → T	26,370	-149.49	DNA/DNA
HKU-N-F			-279.36	DNA/DNA
	T → C	29,148	-275.41	DNA/DNA
2019-nCoV_N1-P	no mutation		-282.37	DNA/DNA
	C → T	28,311	-279.49	DNA /DNA
2019-nCoV_N1-R	no mutation		-277.67	RNA/DNA
	no mutation		-277.85	DNA/DNA
	C → A	28,344	-273.22	RNA/DNA
	C → A	28,344	-273.4	DNA/DNA
2019-nCoV_N3-F	no mutation		-280.31	DNA/DNA
	T → C	28,688	-279.86	DNA/DNA
2019-nCoV_N3-R	no mutation		-272.3	RNA/DNA
	no mutation		-273.04	DNA/DNA
	C → T	28,739	-267.95	RNA/DNA
	C → T	28,739	-268.57	DNA/DNA

In summary, out of the 11 mutations studied, DinoKnot predicted that 3 of the mutations caused no change to the interaction structure, 5 of the mutations caused partial mismatching between the primer and target site and 3 of the mutations caused complete mismatching between the primer and target site. Based on the interaction structures predicted by DinoKnot, the mutations in the primer binding regions of the E-Sarbeco-R, HKU-N-F, and 2019-nCoV_N3-R primers are the most likely to have the greatest effect on decreasing primer sensitivity for SARS-CoV-2 detection since the structures show no binding of the primer to the target area. All mutations resulted in an increase in the energy of binding, except for the mutation in the CCDC-ORF1-P binding region, which lowered the energy of binding. The energy increase means that binding is not as favourable as before and in environments where there is competition for binding, binding may not happen.

Variants of Concern

Mutations were detected in the Alpha, Beta, Delta and Omicron variants of concern in the gene areas presented in Table 5.3

Table 5.3: **Mutations in the gene areas of variants of concern compared to the reference genome NC_045512.2**.

Variant	Gene region with mutations	# Mutated Bases
Alpha	RdRp	1
	Nucleocapsid (N)	9
Beta	Nucleocapsid (N)	3
	nsp14	1
	Envelope (E)	2
Delta	RdRp	1
	Envelope (E)	1
	Nucleocapsid (N)	6
Omicron	RdRp	1
	Nucleocapsid (N)	14
	nsp10	1
	Envelope (E)	5

Despite these mutations, DinoKnot did not predict any change in the primer binding ability. The primer binding results predicted by DinoKnot were the same as presented for the reference genome. Since the variant sequences did not change the predicted primer binding, the in silico thermodynamic evidence supports the primers having the same efficiency as determined by Vogels *et al.* [1].

Clinical report of variant causing N gene detection issues

DinoKnot was used to predict the interaction between the CCDC-N-F primer and the EPI_ISL_1061414 sequence of the hCoV-19/Finland/FinD796H/2021 strain with mutations in the N gene region. The interaction structure presented in Figure 5.8 shows no binding of the CCDC-N primer to the EPI_ISL_1061414 sequence but complete binding to the reference NC_045512.2 sequence. DinoKnot was able to predict the disruption of primer binding which was shown through a clinical report that this variant was unable to be detected by this primer-probe set [70]. There was no change in the interaction structure between the reference and variant strain for the CCDC-N reverse primer and probe, which narrows down the detection issue to the mutations in the forward primer. This is consistent with Laine *et al.*'s explanation of the detection issue being likely due to mutations in the region where the CCDC-N-F primer binds [70]. Prior predictions were hypotheses on how mutations may prevent primer binding, but this result is supported by real world clinical data. This is a validation of how DinoKnot may be used to screen variants of concern to highlight potential

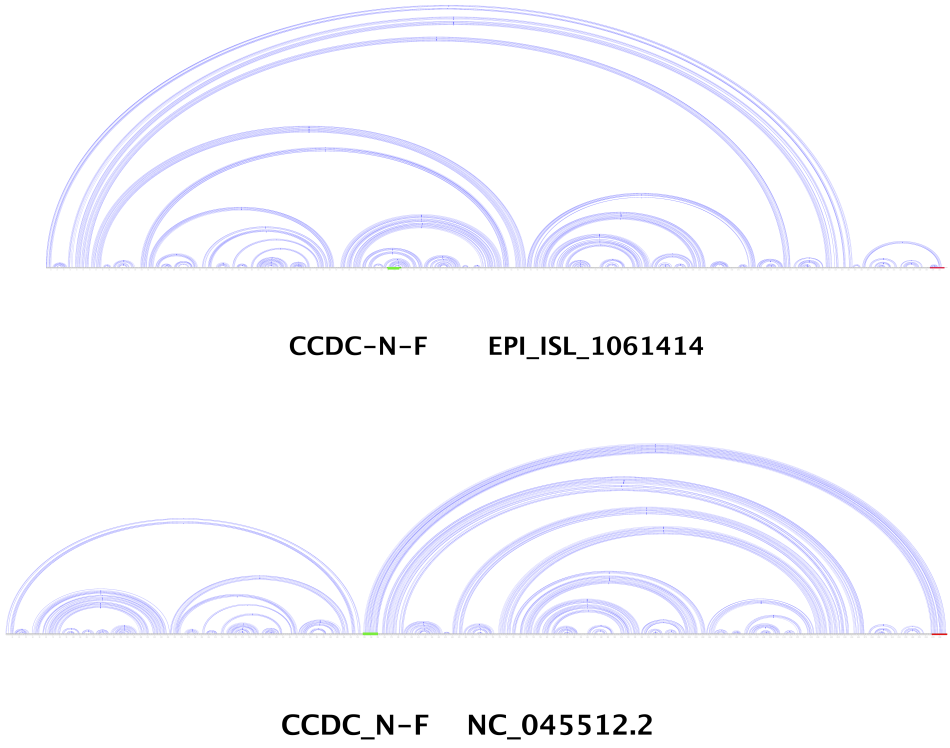

Figure 5.8: **Interaction of the CCDC-N-F primer with the EPI_ISL_1061414 variant strain compared to the reference genome.** The CCDC-N-F primer sequence and the cDNA of the N gene region was input into DinoKnot to determine the interaction structure. The CCDC-N-F primer is highlighted in red and the expected binding site is highlighted in green. The mutations of in the variant strain cause a disruption to the CCDC-N-F primer binding, compared to the reference genome NC_045512.2 which has complete primer binding.

detection issues to help guide laboratory experimentation and allocation of clinical resources, saving time and money from false negative test results.

5.3.2 miRNA - SARS-CoV-2 interaction structures

In Chapter 2, the interaction of miR-2392 with the SARS-CoV-2 genome and miRNA target site prediction tools were discussed. In this section the target sites on the SARS-CoV-2 genome predicted by McDonald *et al.* [7] using the tool miRanda [6] for miR-2932 are discussed and DinoKnot is used to investigate how intramolecular structures flanking the target site may impact the interaction site.

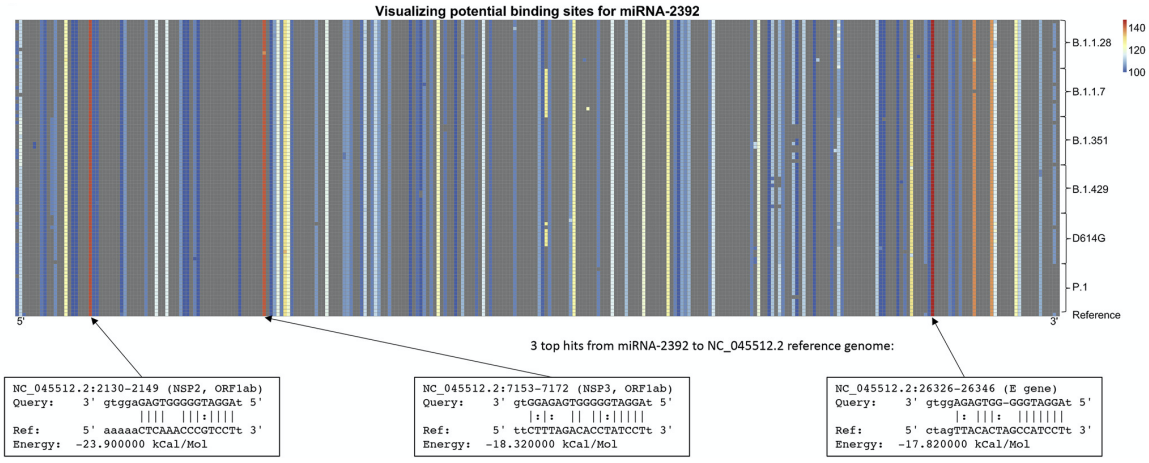

Figure 5.9: **Top three binding sites predicted by miRanda [6] of miR-2392 to the SARS-CoV-2 reference genome and evolutionary conservation of these sites in variant lineages.** This figure was obtained from McDonald *et al.* [7] Figure 2C under the terms of the Creative Commons Attribution-NonCommercial-No Derivatives License (CC BY NC ND) (Permission is not required for this non-commercial use).

Predicted Target Sites

McDonald *et al.* [7] identified the top three miR-2392 binding sites on the SARS-CoV-2 reference genome presented in Fig 5.9. These binding sites were in the *NSP2*, *NSP3*, and *E* gene locations of the genome and were also conserved between different lineages of the SARS-CoV-2 virus [7]. The tool miRanda [6] reports the predicted binding between miR-2392 and the target site, as well as the free energy of the interaction. The energy that is reported only considers the target site sequence, not the surrounding region. Using DinoKnot, the intramolecular structures of miR-2392 and the target site, as well as the region flanking the target site and their impact on the minimum free energy interaction can be investigated.

DinoKnot - Intramolecular Structures

In Fig 5.10, the interaction structure predicted by DinoKnot between the top three target sites and miR-2392 is presented, as well as the interaction between miR-2392 and the target site with a 100bp flanking region on both sides. When considering the interaction between miR-2392 and the target site, DinoKnot predicts the seed region of miR-2392 to interact with the target site which agrees with the miRanda results. However, when considering the intramolecular structure surrounding the target site,

DinoKnot does not predict binding to the target site for any of the three target sites. It is predicted to be more energetically favourable for miR-2392 to bind to locations either upstream or downstream from the target site. As well, the seed region of miR-2392, which is important for miRNA binding, is not predicted to bind to any site. All interaction structures considering the 100bp flanking region of the *NSP2*, *NSP3*, and *E* gene target sites predict pseudoknotted structures. Specifically, the target site for *NSP2* is predicted to be part of a pseudoknotted structure which may be preventing the miR-2932 interaction with the target site predicted by miRanda.

This example is meant to highlight how considering the intramolecular structure of the interacting nucleotides prior to interaction and potential pseudoknotted structures may impact the predicted binding sites. miRNA target prediction tools are designed for the predicting target sites with the factors that are important of miRNA binding, such as seed sequence complementarity. DinoKnot has the potential to be combined with these existing tools, such as miRanda, to add an additional layer of verification of target site prediction prior to laboratory validation. Since DinoKnot is capable of considering the intramolecular structure of nucleic acid strands prior to interaction, interaction structures predicted by DinoKnot could be used to eliminate target site predictions that have energetically unfavourable structures inside the human body. If the top binding sites predicted by miRanda or other similar tools were to be screened through DinoKnot with their flanking regions, those with intramolecular structures or pseudoknotted structures that may prevent the interaction could be considered as a factor to lower the reported score of the target site. Temperature would not be a limitation in this case, as human miRNA interactions inside the body occur at 37°C, the temperature that DinoKnot's energetic parameters are based on.

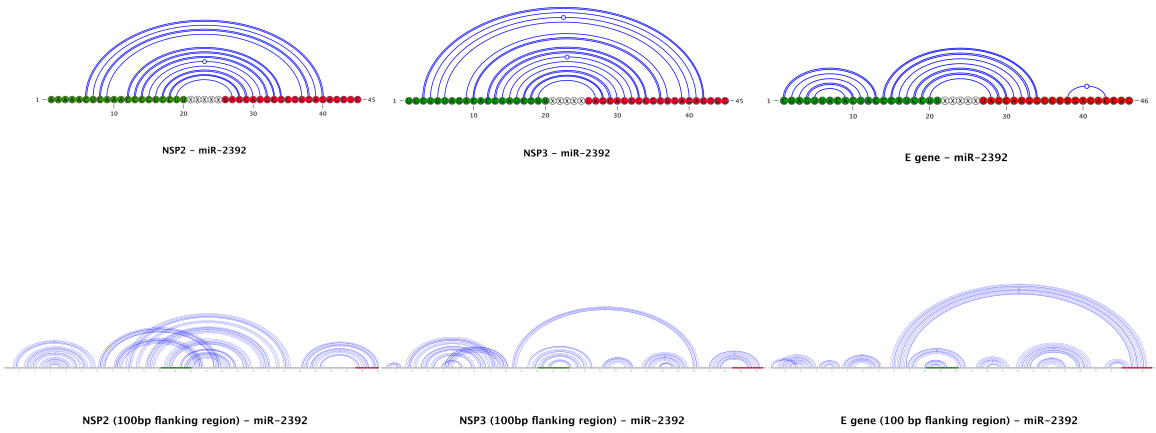

Figure 5.10: **miR2392 target site interaction structures.** The expected target site predicted by miRanda [6] is highlighted in green and the miR-2392 sequence is highlighted in red. The top row represents the interaction structure between miR-2392 and the NSP2, NSP3 and E gene target sites only. The bottom row represents the interaction structure between miR-2392 and the target site with 100 base pairs flanking either side of the target site to show how the intramolecular structure is predicted to impact miR-2392 binding.

Chapter 6

Conclusion and Future Work

This thesis aimed to investigate the minimum free energy (MFE) nucleic acid interaction structures that occur during qRT-PCR testing in order to determine if the SARS-CoV-2 intramolecular structure or genome mutations may affect the interactions required for COVID-19 testing. To study the structure of these interactions, I introduced DinoKnot, the first program to use the relaxed hierarchical folding hypothesis for prediction of the structure of two interacting nucleic acid molecules of the same or different type. Using DinoKnot, I studied the DNA/RNA interactions that occur between the reverse primers with the SARS-CoV-2 genome from 9 experimentally validated primer-probe sets. I compared DinoKnot's results to RNAcofold, an existing tool that considers only the interaction site between DNA/RNA molecules and pseudoknot-free structures. DinoKnot predicted interaction structures that were comparable to RNAcofold but were more consistent with the analytical efficiency and sensitivities of the primer-probe sets. DinoKnot's in silico predictions may give structural insights to the reduced analytical sensitivity of the RdRp-SARSr primer-probe set and laboratory experiments are required to determine whether changing the degenerate bases R and S to A and G, respectively, on the reverse primer will increase the primer sensitivity. Since the SARS-CoV-2 interactions did not involve a known pseudoknotted structure, I highlighted an example of the *CCR5* mRNA and its interaction with miR-1224 to demonstrate DinoKnot's ability to predict pseudoknotted structures. Therefore, DinoKnot allows for the prediction of biologically relevant structures that existing tools are not capable of predicting. Furthermore, I investigated how mutations in the primer-probe binding regions and in variants of concern may affect primer binding based on DinoKnot's in silico predictions of how these mutations change the interaction structure of the primer/probe with its target site.

I predicted mutations in the E-Sarbeco-R, HKU-N-F, and 2019-nCoV_N3-R primer binding regions that may prevent primer binding but found no change in interaction structure for a genome from each of the Alpha, Beta, Delta and Omicron variants of concern. DinoKnot's interaction structure predictions were consistent with a real-world clinical example where mutations in the N gene region were shown to prevent detection, validating how DinoKnot may be used to screen variants of concern to direct hypotheses on possible detection issues.

An additional objective of this thesis was to explore further applications of DinoKnot to miRNA target site prediction to consider how intramolecular structures and pseudoknots may impact the nucleic acid interaction. I presented the interaction structure of miR-2932 with the top three binding sites to the SARS-CoV-2 genome predicted by miRanda [6]. When predicting the interaction structures including the 100 nucleotide region flanking the target site, the most energetically favourable binding location was outside the target site in all three cases. This highlights the importance of considering the energetics of the target site intramolecular structure prior to interaction. While DinoKnot addresses the shortcomings of existing tools that predict nucleic acid interactions, DinoKnot has limitations in that the structures predicted are limited to what can be predicted by its underlying methods and energy parameters.

6.1 Future Work

DinoKnot is a tool that can be used in a broad range of applications involving DNA/RNA interactions. DinoKnot could be used in the design of other nucleic acid-based testing, such as an aptamer test. Aptamers are single stranded RNA or DNA nucleotides (10-100nt) that are able to bind to targets such as viruses and proteins [74]. An aptamer's binding specificity is ensured by their secondary and tertiary structure [74]. Aptamers are capable of a lower limit of detection than PCR testing. An RNA aptamer test designed for the SARS-CoV Nucleocapsid protein showed a detection limit of 2 pg/mL [75]. An aptamer test for Norovirus, a positive sense RNA virus, has a limit of detection of 200 viral copies/mL [76], which is lower than the qRT-PCR limit of detection of determined by Vogels *et al.* by a magnitude of 10^3 [1]. The lower limit of detection that is possible with aptamer tests would be beneficial in testing patients with low viral load, such as asymptomatic carriers. DinoKnot considers the structure of interactions, which is essential for aptamers, and could therefore

be used to predict apatmers with optimal binding to a target area in order to guide laboratory experimentation.

Finally, DinoKnot could be implemented as part of a pipeline with existing miRNA target site prediction tools as an additional screening tool prior to laboratory experimentation used to confirm miRNA interactions to target sites. Considering the structure of these interactions or the structure of the target site may also allow for a deeper investigation into the function of the miRNA binding to the SARS-CoV-2 virus. For example, as previously discussed, SARS-CoV-2 and other viruses use -1 programmed ribosomal frameshifting (-1 PRF) for replication and a pseudoknotted structure has been implicated as part of this function [37]. If a structure such as this were to be targeted by an miRNA, this would elicit functional information on how the miRNA works to clear the SARS-CoV-2 infection, in this hypothetical case by preventing viral replication. Therefore, DinoKnot could be used to investigate the function of miRNA-target site interactions, as well any other type of nucleic acid interactions.

Appendix A

Additional Information

The following page contains the interaction structures predicted by DinoKnot for all mutations discussed in Chapter 5.3.1.

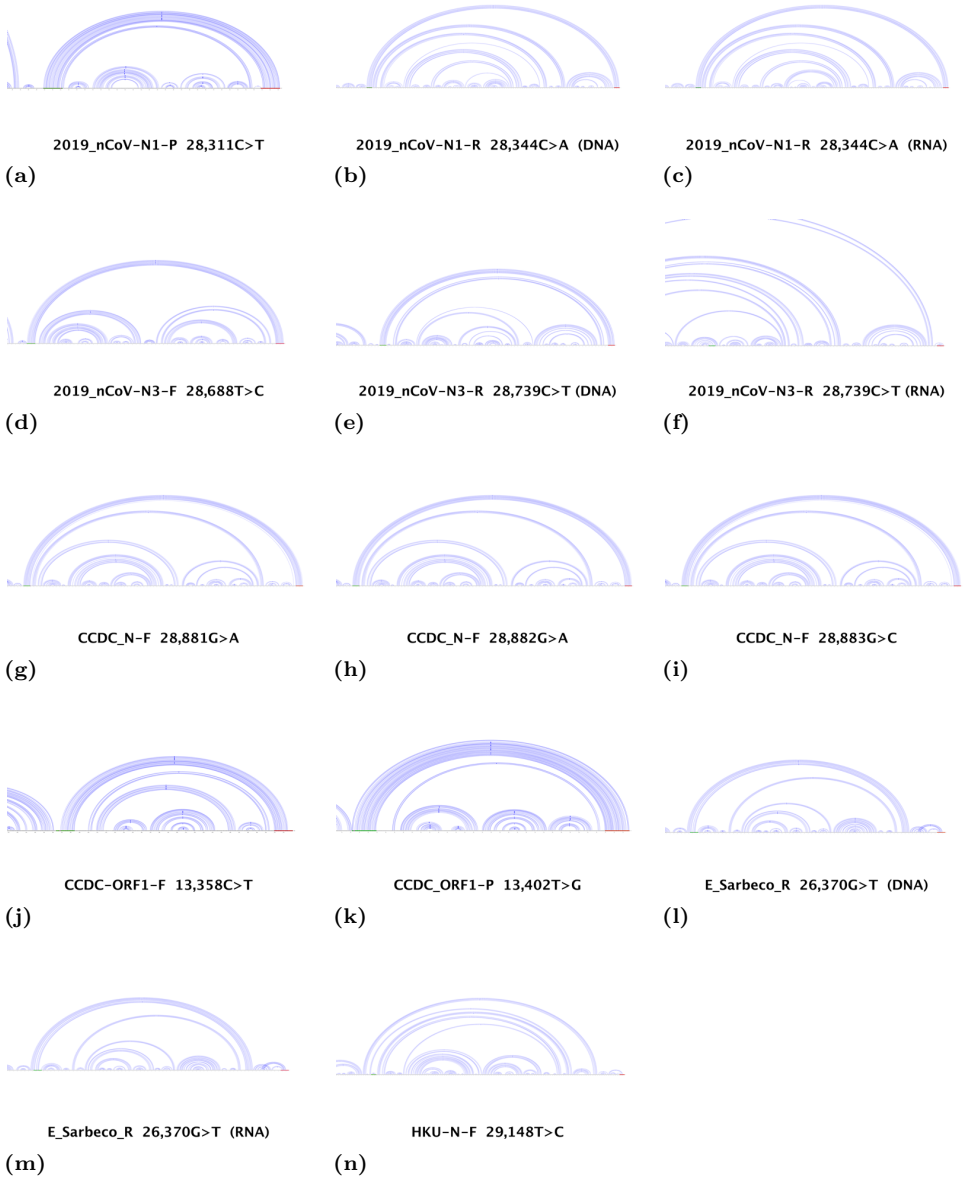

Figure A.1: **Interaction structures of primer/probes with the transcript regions containing mutations in the primer/probe binding region predicted by DinoKnot.** The expected target region is highlighted in green and the primer/probe sequence is highlighted in red. The mutated base is highlighted in pink.

Bibliography

[1] C. Vogels, A. Brito, A. Wyllie, and et al., "Analytical sensitivity and efficiency comparisons of sars-cov-2 rt–qpcr primer–probe sets," *Nat. Microbiol.*, vol. 5, no. 10, pp. 1299–1305, 2020.

[2] H. Jabbari and A. Condon, "A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures," *BMC Bioinform.*, vol. 15, no. 1, pp. 147+, 2014.

[3] M. Andronescu, R. Aguirre-Hernández, A. Condon, and H. Hoos, "Rnasoft: A suite of RNA secondary structure prediction and design software tools.," *Nucleic Acids Res.*, vol. 31, pp. 3416–3422, July 2003.

[4] K. Darty, A. Denise, and Y. Ponty., "Varna: Interactive drawing and editing of the RNA secondary structure.," *Bioinformatics*, pp. 1974–1975, 2009.

[5] R. Lorenz, S. Bernhart, C. Honer zu Siederdissen, H. Tafer, C. Flamm, P. Stadler, and I. Hofacker, "ViennaRNA package 2.0," *Algorithms Mol. Biol.*, vol. 6, no. 1, p. 26, 2011.

[6] A. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. Marks, "MicroRNA targets in drosophila," *Genome Biol.*, vol. 4, no. 11, pp. 1–27, 2003.

[7] J. T. McDonald, F. J. Enguita, D. Taylor, R. J. Griffin, W. Priebe, M. R. Emmett, M. M. Sajadi, A. D. Harris, J. Clement, J. M. Dybas, *et al.*, "Role of miR-2392 in driving SARS-CoV-2 infection," *Cell Rep.*, vol. 37, no. 3, p. 109839, 2021.

[8] T. Newman, H. F. K. Chang, and H. Jabbari, "In silico prediction of covid-19 test efficiency with dinoknot," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pp. 470–479, 2021.

[9] C. Siniscalchi, A. Di Palo, A. Russo, and N. Potenza, "Human microRNAs interacting with SARS-CoV-2 RNA sequences: Computational analysis and experimental target validation," *Front. Genet*, vol. 12, 2021.

[10] E. Girardi, P. López, and S. Pfeffer, "On the importance of host microRNAs during viral infection," *Front. Genet*, p. 439, 2018.

[11] W. Wang, Y. Xu, R. Gao, and et al., "Detection of SARS-CoV-2 in different types of clinical specimens.," *JAMA*, vol. 323, pp. 1843–1844, 2020.

[12] S. Bustin and R. Mueller, "Real-time reverse transcription PCR (qRT-PCR) and its potential use in clinical diagnosis.," *Clin Sci (Lond)*, vol. 109, pp. 365–379, 2020.

[13] X. Dai, S. Zhang, and K. Zaleta-Rivera, "Rna: interactions drive functionalities," *Mol. Biol. Rep*, vol. 47, no. 2, pp. 1413–1434, 2020.

[14] R. Lorenz, I. Hofacker, and S. Bernhart, "Folding RNA/DNA hybrid duplexes," *Bioinformatics*, vol. 28, pp. 2530–2531, 07 2012.

[15] J. S. Mattick and I. V. Makunin, "Non-coding RNA," *Hum. Mol. Genet.*, vol. 15, pp. R17–R29, 04 2006.

[16] G. Varani and W. H. McClain, "The g· u wobble base pair," *EMBO Rep.*, vol. 1, no. 1, pp. 18–23, 2000.

[17] I. K. Oluoch, A. Akalin, Y. Vural, and Y. Canbay, "A review on RNA secondary structure prediction algorithms," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pp. 18–23, IEEE, 2018.

[18] M. G. Seetin and D. H. Mathews, *RNA Structure Prediction: An Overview of Methods*, pp. 99–122. Totowa, NJ: Humana Press, 2012.

[19] M. Zuker, "On finding all suboptimal foldings of an rna molecule," *Science*, vol. 244, no. 4900, pp. 48–52, 1989.

[20] R. Nussinov and A. Jacobson, "Fast algorithm for predicting the secondary structure of single-stranded rna.," *PNAS*, vol. 77, pp. 6309–6313, Nov. 1980.

[21] M. Zuker and P. Stiegler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.," *Nucleic Acids Res.*, vol. 9, pp. 133–148, Jan. 1981.

[22] T. Akutsu, "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots,," *Discret. Appl. Math.*, vol. 104, pp. 45–62, Aug. 2000.

[23] R. Lyngsø and C. Pedersen, "Pseudoknots in RNA secondary structures," in *Proceedings of the fourth annual international conference on Computational molecular biology*, RECOMB '00, (New York, NY, USA), pp. 201–209, ACM, 2000.

[24] S. Sheikh, R. Backofen, and Y. Ponty, "Impact of the energy model on the complexity of RNA folding with pseudoknots," in *Combinatorial Pattern Matching* (J. Kärkkäinen and J. Stoye, eds.), vol. 7354 of *Lecture Notes in Computer Science*, pp. 321–333, Springer Berlin Heidelberg, 2012.

[25] E. Rivas and S. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots.," *J. Mol. Biol.*, vol. 285, pp. 2053–2068, Feb. 1999.

[26] R. Dirks and N. Pierce, "A partition function algorithm for nucleic acid secondary structure including pseudoknots," *J. Comput. Chem.*, vol. 24, pp. 1664–1677, Oct. 2003.

[27] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinform.*, vol. 5, pp. 104+, Aug. 2004.

[28] H. Jabbari, A. Condon, and S. Zhao, "Novel and efficient RNA secondary structure prediction using hierarchical folding.," *J Comput Biol*, vol. 15, pp. 139–163, Mar. 2008.

[29] H. Jabbari, I. Wark, C. Montemagno, and S. Will, "Knotty: efficient and accurate prediction of complex RNA pseudoknot structures," *Bioinformatics*, vol. 34, pp. 3849–3856, 06 2018.

[30] S. Washietl, I. L. Hofacker, and P. F. Stadler, "Fast and reliable prediction of noncoding rnas," *PNAS*, vol. 102, no. 7, pp. 2454–2459, 2005.

[31] M. Andronescu, Z. C., and A. Condon, "Secondary structure prediction of interacting RNA molecules.," *J. Mol. Biol.*, vol. 345, pp. 987–1001, Feb. 2005.

[32] R. Dirks, J. Bois, J. Schaeffer, E. Winfree, and N. Pierce, "Thermodynamic analysis of interacting nucleic acid strands," *SIAM Rev.*, vol. 49, pp. 65–88, Jan. 2007.

[33] C. Alkan, E. Karako, J. Nadeau, S. Sahinalp, and K. Zhang, "RNA-RNA interaction prediction and antisense RNA target search," *J. Comput. Biol.*, vol. 13, no. 2, pp. 267–282, 2006. PMID: 16597239.

[34] Y. Kato, T. Akutsu, and H. Seki, "A grammatical approach to RNA-RNA interaction prediction," *Pattern Recognit.*, vol. 42, pp. 531–538, Apr. 2009.

[35] H. Chitsaz, R. Salari, S. Sahinalp, and R. Backofen, "A partition function algorithm for interacting nucleic acid strands," *Bioinformatics*, vol. 25, pp. i365–i373, June 2009.

[36] H. Tafer and I. L. Hofacker, "RNAplex: a fast tool for RNA–RNA interaction search," *Bioinformatics*, vol. 24, no. 22, pp. 2657–2663, 2008.

[37] L. Trinity, L. Lansing, H. Jabbari, and U. Stege, "SARS-CoV-2 ribosomal frameshifting pseudoknot: Detection of inter-viral structural similarity," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pp. 451–460, 2021.

[38] A. T. Belew, A. Meskauskas, S. Musalgaonkar, V. M. Advani, S. O. Sulima, W. K. Kasprzak, B. A. Shapiro, and J. D. Dinman, "Ribosomal frameshifting in the *CCR5* mRNA is regulated by miRNAs and the nmd pathway," *Nature*, vol. 512, no. 7514, pp. 265–269, 2014.

[39] Y. A. Khan, G. Loughran, A.-L. Steckelberg, K. Brown, S. J. Kiniry, H. Stewart, P. V. Baranov, J. S. Kieft, A. E. Firth, and J. F. Atkins, "Evaluating ribosomal frameshifting in *CCR5* mRNA decoding," *Nature*, vol. 604, no. 7906, pp. E16–E23, 2022.

[40] D. W. Staple and S. E. Butcher, "Pseudoknots: RNA structures with diverse functions," *PLoS Biol.*, vol. 3, no. 6, p. e213, 2005.

[41] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[42] J. Wang, L. Peck, and K. Becherer, "DNA supercoiling and its effects on DNA structure and function," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 47, pp. 85–91, Cold Spring Harbor Laboratory Press, 1983.

[43] National Library of Medicine - NCBI, "Polymerase chain reaction (PCR)," 2017. Accessed May 31, 2022.

[44] J. O'Brien, H. Hayder, Y. Zayed, and C. Peng, "Overview of microRNA biogenesis, mechanisms of actions, and circulation," *Front. Endocrinol.*, vol. 9, p. 402, 2018.

[45] R. Mirzaei, F. Mahdavi, F. Badrzadeh, S. R. Hosseini-Fard, M. Heidary, A. S. Jeda, T. Mohammadi, M. Roshani, R. Yousefimashouf, H. Keyvani, *et al.*, "The emerging role of microRNAs in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection," *Int. Immunopharmacol.*, vol. 90, p. 107204, 2021.

[46] R. Marchi, B. Sugita, A. Centa, A. S. Fonseca, S. Bortoletto, K. Fiorentin, S. Ferreira, and L. R. Cavalli, "The role of microRNAs in modulating SARS-CoV-2 infection in human cells: a systematic review," *Infect. Genet. Evol.*, vol. 91, p. 104832, 2021.

[47] G. Riolo, S. Cantara, C. Marzocchi, and C. Ricci, "miRNA targets: from prediction tools to experimental validation," *Methods Protoc.*, vol. 4, no. 1, p. 1, 2020.

[48] J. Krüger and M. Rehmsmeier, "RNAhybrid: microRNA target prediction easy, fast and flexible," *Nucleic Acids Res.*, vol. 34, no. suppl_2, pp. W451–W454, 2006.

[49] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, "The role of site accessibility in microRNA target recognition," *Nat. Genet.*, vol. 39, no. 10, pp. 1278–1284, 2007.

[50] N. N. Shaw and D. P. Arya, "Recognition of the unique structure of DNA:RNA hybrids," *Biochimie*, vol. 90, no. 7, pp. 1026–1039, 2008.

[51] C. Dieterich and P. F. Stadler, "Computational biology of RNA interactions," *Wiley Interdiscip. Rev. RNA*, vol. 4, no. 1, pp. 107–120, 2013.

[52] M. Nakama, K. Kawakami, T. Kajitani, T. Urano, and Y. Murakami, "DNA–RNA hybrid formation mediates RNAi-directed heterochromatin formation," *Genes Cells*, vol. 17, no. 3, pp. 218–233, 2012.

[53] J. A. Howard, S. Delmas, I. Ivančić-Baće, and E. L. Bolt, "Helicase dissociation and annealing of RNA-DNA hybrids by escherichia coli Cas3 protein," *Biochem. J.*, vol. 439, no. 1, pp. 85–95, 2011.

[54] J. E. Garcia-Robledo, M. C. Barrera, and G. J. Tobón, "Crispr/Cas: from adaptive immune system in prokaryotes to therapeutic weapon against immune-related diseases: Crispr/Cas9 offers a simple and inexpensive method for disease modeling, genetic screening, and potentially for disease therapy," *Int. Rev. Immunol.*, vol. 39, no. 1, pp. 11–20, 2020.

[55] O. A. for Health Protection and P. P. H. Ontario)., "Focus on: an overview of cycle threshold values and their role in SARS-CoV-2 real-time PCR test interpretation," *Queen's Printer for Ontario*, 2020.

[56] J. A. Kulkarni, D. Witzigmann, S. B. Thomson, S. Chen, B. R. Leavitt, P. R. Cullis, and R. van der Meel, "The current landscape of nucleic acid therapeutics," *Nat. Nanotechnol.*, vol. 16, no. 6, pp. 630–643, 2021.

[57] R. Dimitrov and M. Zuker, "Prediction of hybridization and melting for double-stranded nucleic acids," *Biophys. J.*, vol. 87, pp. 215–226, July 2004.

[58] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich, "Fast and effective prediction of microRNA/target duplexes.," *RNA*, vol. 10, pp. 1507–1517, Oct. 2004.

[59] J. Reuter and D. Mathews, "Rnastructure: software for RNA secondary structure prediction and analysis.," *BMC Bioinform.*, vol. 11, no. 1, pp. 129+, 2010.

[60] M. Kery, M. Feldman, J. Livny, and B. Tjaden, "TargetRNA2: identifying targets of small regulatory RNAs in bacteria," *Nucleic Acids Res.*, vol. 42, no. Web Server issue, pp. W124–W129, 2014.

[61] I. Tinoco and C. Bustamante, "How RNA folds," *J. Mol. Biol.*, vol. 293, pp. 271–281, Oct. 1999.

[62] M. Andronescu, A. Condon, H. Hoos, D. Mathews, and K. Murphy, "Computational approaches for RNA energy parameter estimation," *RNA*, vol. 16, pp. 2304–2318, Dec. 2010.

[63] N. Sugimoto, S. Nakano, M. Yoneyama, and K. Honda, "Improved Thermodynamic Parameters and Helix Initiation Factor to Predict Stability of DNA Duplexes," *Nucleic Acids Res.*, vol. 24, pp. 4501–4505, 11 1996.

[64] F. Martin and I. Tinoco, "DNA-RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription," *Nucleic Acids Res.*, vol. 8, pp. 2295–2300, 05 1980.

[65] N. Sugimoto, S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, T. Ohmichi, M. Yoneyama, and M. Sasaki, "Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes," *Biochemistry*, vol. 34, pp. 11211–11216, 09 1995.

[66] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res.*, vol. 44, pp. (D1):D7–D19, 2016.

[67] World Health Organization, "WHO - coronavirus disease (COVID-19) technical guidance: Laboratory testing for 2019-nCoV in humans.," 2020 (Retrieved Sept 6, 2020). Accessed: 2020-08-27.

[68] J. Iserte, B. Stephan, S. Goñi, C. Borio, P. Ghiringhelli, and M. Lozano, "Family-specific degenerate primer design: A tool to design consensus degenerated oligonucleotides," *Biotechnol. Res. Int.*, vol. 2013, p. 9, 2013.

[69] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health.," *Glob Chall.*, vol. 1, pp. 33–46, 2017.

[70] P. Laine, H. Nihtilä, E. Mustanoja, A. Lyyski, A. Ylinen, J. Hurme, L. Paulin, S. Jokiranta, P. Auvinen, and T. Meri, "SARS-CoV-2 variant with mutations in n gene affecting detection by widely used PCR primers," *J Med Virol*, vol. 94, no. 3, pp. 1227–1231, 2022.

[71] D. Svec, A. Tichopad, V. Novosadova, M. W. Pfaffl, and M. Kubista, "How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments," *Biomol. Detect. Quantif.*, vol. 3, pp. 9–16, 2015.

[72] M. Oppermann, "Chemokine receptor *CCR5*: insights into structure, function, and regulation," *Cell. Signal.*, vol. 16, no. 11, pp. 1201–1210, 2004.

[73] B. Chen, A. P. Longhini, F. Nußbaumer, C. Kreutz, J. D. Dinman, and T. K. Dayie, "*CCR5* RNA pseudoknots: Residue and site-specific labeling correlate internal motions with microRNA binding," *Chem. Eur. J.*, vol. 24, no. 21, pp. 5462–5468, 2018.

[74] X. Zou, J. Wu, J. Gu, L. Shen, and L. Mao, "Application of aptamers in virus detection and antiviral therapy," *Front. Microbiol.*, vol. 10, p. 1462, 2019.

[75] D. Ahn, I. Jeon, J. Kim, M. Song, S. Han, S. Lee, H. Jung, and J. Oh, "RNA aptamer-based sensitive detection of SARS coronavirus nucleocapsid protein," *Analyst*, vol. 134, pp. 1896–1901, 2009.

[76] P. Weerathunge, R. Ramanathan, V. Torok, K. Hodgson, Y. Xu, R. Goodacre, B. Behera, and V. Bansal, "Ultrasensitive colorimetric detection of murine norovirus using nanozyme aptasensor," *Anal. Chem.*, vol. 91, p. 3270–3276, 2019.