Phylogenetic Prediction of Nucleation Complexes and their Roles in Capsid Assembly

Eva Ursula Weiß

Doctor of Philosophy

University of York

Department of Biology

November 2019

Abstract

Infectious diseases are among the leading causes of death, especially in developing countries. Novel insights into viral life cycles can be exploited for antiviral therapies. One such opportunity is packaging signal mediated assembly of viral capsids, in which multiple dispersed sequence/structure motifs in the genome, called packaging signals (PSs), regulate capsid formation. I have developed a novel phylogenetic method to group viruses by their PS distributions. This method is specifically designed to identify PSs that are in similar positions in extended families of viruses, thus identifying candidates within the PS distribution that may have specific functions. We exemplify this for two viral families: *Hepadnaviridae* and *Leviviridae*. After identification of the PS motif in hepatitis B virus (HBV), the method was applied to different sets of HBV sequences. The distribution pattern of PSs highlighted small groups of highly conserved PSs. We investigated their roles in formation of the nucleation complex, and identified a pair that formed into PSs on the same short fragment and may have a double functional role. Using this focus on function, PSs were predicted in related avian and mammalian HBV strains. Application to *Leviviridae* required an extension to include a secondary structure context. This identified six PSs conserved across MS2, BZ13, and Qβ. I developed a computational model of virus assembly in order to test their roles in nucleation and demonstrated that three of these PSs play crucial roles in nucleation of assembly. Studying PS-mediated assembly through phylogeny has thus led to an increased understanding of the essential nucleation of this process in two unrelated viruses.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

In the years since embarking on this PhD journey I was faced with many challenges and obstacles to overcome. From two pregnancies to broken hard drives and stolen laptops, it was not a smooth ride. There are many people who have supported me through these years in one way or another, without whom I would not have made it to the end or even started.

Firstly, I would like to express my gratitude towards the *Wellcome Trust* for funding the CIDCATS programme and making this interdisciplinary project possible. It allowed me to make the jump from my laboratory biology background to a fully computational PhD.

My supervisors *Eric Dykeman* and *Reidun Twarock* have been encouraging, supportive, and challenged me as needed to ensure I achieved the best outcome in this project. Reidun gave me the freedom to explore my own ideas but also provided input and honest feedback. Working together with Eric in the lab and on the modelling taught me a lot about the subject as well as closely collaborating on a project. He has been great teacher and mentor. Both their hard work and passion for good science have inspired me and I feel honoured to have had the chance to work with and learn from two great scientists. Thank you to both as well for fighting my corner over the years and guiding me throughout my PhD.

At the University of York I would also like to thank my TAP members *Nathalie Signoret*, *Leo Caves*, and *Anje-Margriet Neutel* for the discussions and constructive criticism over the years. They always made sure that I kept the biology in mind and that my project stayed on track.

A profound thank you to our experimental collaborators at the University of Leeds in the *Stockley group*. *Peter Stockley* and his group worked closely with us on many parts of the HBV work. Without their hard and good work this part of my project would have lacked experimental data and validation. Thank you to Peter as well for allowing Eric and me to visit your lab and to his lab members, especially *Nikesh Patel*, for taking the time to teaching us some of your experimental techniques.

Thank you to *Vincent Moulton* at the University of East Anglia for the collaboration. His expertise in phylogenetics and bioinformatics has been invaluable in the development of the packaging signal phylogeny algorithm.

The many long days in the office have been made so much more enjoyable by the many group members I was lucky enough to engage with over the years *Emilio Zappa*, *James Geraets*, *German Leonov*, and *Richard Bingham*. Thank you for the talks about science and the world over coffee. Special thanks to German and Richard for allowing me to bounce ideas, for the advice, and for the occasional help with programming questions.

I am grateful to have been lucky enough to be a part of *YCCSA*, which provided a interdisciplinary environment full of interesting people and their exciting research. Being surrounded by people from different backgrounds and disciplines made for many interesting insights. Thank you to all the people of YCCSA (and IGGI) also for the social aspects of everyday office life.

Thanks to *Lisa Sha Li* and *Farzad Fatehi Chenar* for impromptu baby sitting allowing me some extra time in the last stretch of writing this thesis.

To *Anita Bradley-Gilbride* and *Marianne Drabek* a big thank you for giving me feedback on my thesis and being there for me when things were rough. I also owe great gratitude to *Johannes Schauer* for introducing me to computational work and for helping me fix my computer on more than one occasion.

I also thank my parents *Peter* and *Maria Weiß* for always pushing me and believing in my abilities. I would not be where I am today without your support.

Thank you to my sisters *Madgalena* and *Janina Weiß* for the emotional support over the years.

Finally, a thank you to my little family. To my sons *Erik Luis Thor* and *Linus Mario Loki Weiß*, through whom I learned the value of time management, multitasking, and being functional on a couple hours of sleep. My choice to have you two at the beginning and end of my PhD certainly added more challenges to this journey and sometimes made me feel spread too thin but your smiles and cuddles brightened the darkest days. I am grateful to have had these to come home to after many long days in the office. When nothing seemed to work you, reminded me that there is more to life than my buggy code. A special thanks to their amazing childminder *Caroline Stevens* for keeping my boys happy and entertained while I was busy with this work. Last but not least, to my husband *Jimmie Weiß*, for always being there and supporting me with love and patience. I could always count on you, whether I needed to rant about things not working, bounce ideas, practice conference talks or just switch my science brain off and recharge for a bit. The years and especially last months have been stressful but you never failed to calm me down and assure me that I will get to this point. Thank you for being my rock in these stormy times.

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author except where stated otherwise in the text. For all figures that were not created by me I obtained a copy right license, which is stated in the figure legend together with the respective citation. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged by explicit references.

Most of Chapter 2 with the exception of the aptamer analyses has been published in Patel et al. (2017).

The original MS2 capsid assembly model described in Chapter 7 was developed by Dr Eric Dykeman and modified by both of us as detailed in Chapter 7 for use in Chapter 8.

Chapter 1

Introduction

We live in evolutionary competition with microbes – bacteria and viruses. There is no guarantee that we will be the survivors.

Joshua Lederberg (Cullington BJ. 1990. Emerging viruses, emerging threat. Science 247:279-80.)

In humanity's constant battle against viruses research into viral evolution and different parts of their life cycles has been essential. These basic understandings form the foundation for any modern antiviral treatment or vaccine available today. Despite all these achievements, the fight is far from won. For many known viruses there are still no cures, others evolve resistance against such treatments, whilst the threat of a new human-infecting virus is constantly looming. Therefore, we continue to develop new methods to study viruses, identify essential features, and discover ways to interrupt their infectious cycles.

Features of a virus that are particularly important for its fitness are usually conserved, i.e. they occur in all strains of the same species or even in different species. More closely related strains or species tend to share more such features. Phylogeny studies the evolutionary history and relatedness of organisms by using this concept. It can, thus, provide insights into the origin and spread of a virus. Any feature or set of features can be the basis of a phylogeny and it usually

reflects the level of relatedness to be studied. On the most fine-grained level are genomic sequence alignments, i.e. identifying corresponding regions in different sequences (Zvelebil and Baum, 2008, Chapter 4), which allow the resolution of even the closest sequence relatives. One use of this high resolution is in epidemiology in the reconstruction of transmission trees, i.e. tree representations of how one person infected the next, in an outbreak (Kenah et al., 2016). The more distant, however, the strains are, especially when different species are considered, the more difficult it becomes to accurately represent their evolutionary history. Therefore, more slowly changing features are used in such cases starting with amino acid sequences of proteins or only comparing more conserved domains. More recently Bamford and Stuart have used the fold of the capsid protein to group viruses with no recognisable sequence similarity (reviewed in Bamford et al. (2005)). Their phylogenetic clusters include viruses that infect hosts from different domains of life indicating a potentially ancient evolutionary relationship. In between single transmission events and ancient links are many more levels of relatedness, which can be explored by utilising the right features for comparison to gain an appropriate resolution. Seeing how a certain feature evolves, on what time scale, and which stage of infection exerts the most evolutionary pressure, i.e. forces adaptive mutations to increase fitness, can prove useful when considering it as a target for antiviral therapy and how easy it would be for a virus to become resistant. Conversely, finding which subset of a feature is conserved can provide insights into the function it performs. Similarly to conserved parts of a protein illustrating an important functional domain, also other features are likely to include essential parts as well as more variable ones that are more likely to evolve more quickly and thus differ between strains or species. Studying the conservation of a feature can thus provide two-fold information: where it differs between viruses can inform phylogeny and provide insights into how quickly it evolves and the type of evolutionary pressure, where it is the same, on the other hand, highlights the functionally most important parts and may help to uncover

novel roles. Both of these aspects were used in this project for studying packaging signal-mediated assembly in two viral families, *Hepadnaviridae* and *Leviviridae*.

1.1 Background

1.1.1 RNA Secondary Structure

In most life forms the genetic sequence is present in the form of DNA, i.e. a molecule consisting of two strands of nucleotides, which contain a sugar (deoxyribose) with a phosphate group forming the backbone and four bases: adenine (A), guanine (G), cytosine (C), and thymine (T) (Alberts et al., 2002c). The genetic code lies in the sequence of these bases, i.e. the primary structure, as sets of threes form codons, which encode amino acids, the building blocks of proteins (Alberts et al., 2002c). Despite sometimes occurring single-stranded in the genomes of single-stranded DNA viruses, DNA is usually double-stranded. RNA differs from DNA by some of its components and its functions. RNA contains ribose as sugar units and uracil (U) instead of thymine (T), which affect its stability and structure (Alberts et al., 2002a). It is mostly known for its roles in the process of gene expression, i.e. going from a gene in the DNA to a protein; however, in

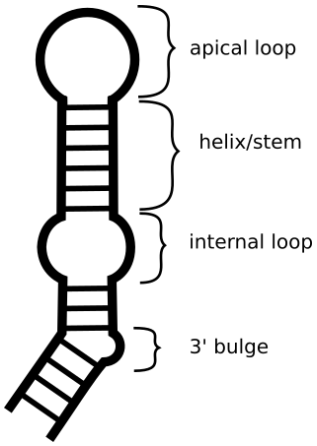

FIGURE 1.1: **Anatomy of a stem-loop.** An RNA stem-loop consists of at least an unpaired region on the top - the apical loop - and a base-paired helix/stem below. Further unpaired regions are either internal loops or one-sided bulges. Horizontal parallel lines represent base-pairs while thicker outer lines represent the sugar backbone.

some viruses it also takes the role of the genome. Unlike DNA, RNA, on the other hand, is commonly found single-stranded in nature with the exception of double-stranded RNA viruses. Instead of interacting with another copy these molecules interact with themselves forming a number of different secondary or tertiary structures. When in double-stranded form, either through interaction with another RNA molecule or with itself, RNA has a higher persistence length than double-stranded DNA, i.e. it is less bendable (Kebbekus et al., 1995). Therefore, small structures can be considered rigid (Gary et al., 2007). The simplest secondary structure is a stem-loop (SL) (Figure 1.1). It occurs when a stretch of RNA folds back onto itself forming a base-paired helix or stem and a single-stranded loop on top called the apical loop. Mismatches within the stem result in either bulges if one-sided or internal loops. Smaller and larger structures often serve important biological functions such as translation initiation or inhibition (Malys and Nivinskas, 2009; Deiorio-Haggar et al., 2013). One example of a more complex tertiary structure is the clover-leaf structure of the tRNA (Figure 1.2). In the following section packaging signals (PSs) will be introduced, which function as simple SLs. Note that a list of abbreviations can be found on page 387.

1.1.2 Packaging Signal-Mediated Assembly

At the end of its life cycle, after a virus has sufficiently replicated its components within a host cell, it needs to form viral particles, i.e. particles containing all components necessary for continued infection, to start a new infection cycle (Neuman and Buchmeier, 2016). Assembly of capsid protein (CP) subunits into full capsid and specific packing of the genome inside this protein container are essential parts at that stage. These two processes are often interlinked, especially in single-stranded RNA (ssRNA) viruses, and optimised for efficiency. For a long time this processes was thought to be governed by electrostatics, the interaction of positively and negatively charged molecules (Belyi and Muthukumar, 2006; Forrey and Muthukumar, 2009; Devkota et al., 2009; Hagan, 2009; Cadena-Nava

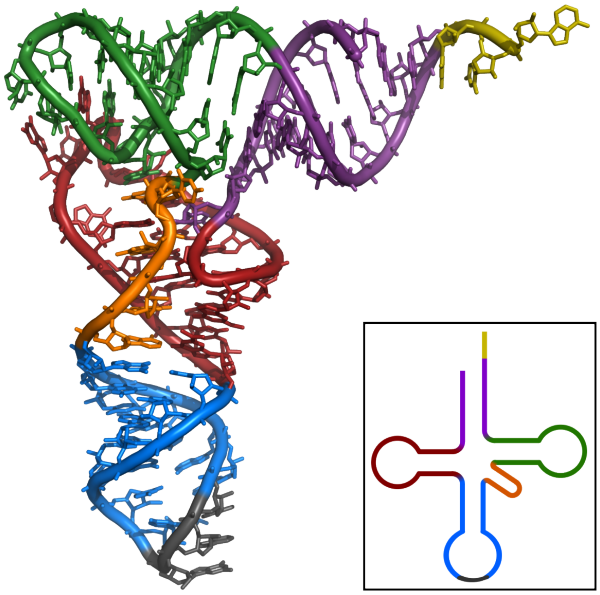

FIGURE 1.2: **X-ray structure of the tRNAPhe from yeast.** It shows the classic cloverleaf structure of tRNA as a schematic (small box) and as 3D structure. Data was obtained by PDB: 1ehz and rendered with PyMOL. Source: https://en.wikipedia.org/wiki/File:TRNA-Phe_yeast_1ehz.png licenced under CC-BY-SA by user Yikrazuul, retrieved on 2018-11-12.

et al., 2012). Positive charge on the capsid inside was supposed to be responsible for non-specifically binding the negatively charged genomic RNA, which was a passive entity in the process. This would mean that a virus would package any RNA of the right length such as host messenger RNA (mRNA) and not just its genomic RNA. While this is possible and does happen in wild-type viruses, viruses have evolved strategies to minimise this. In one experiment only 1% of packaged RNA was host-derived indicating a strong specificity for viral genomic RNA (Routh et al., 2012). Not being able to package its own genome over other host RNA would cause a massive reduction in fitness as fewer functional viral particles would be produced. The current model of ssRNA virus assembly and packaging, therefore, is based on an active role for the RNA in the process. Over time more and more examples of specific RNA-CP interactions have been discovered (Twarock and Stockley, 2019). Small structure/sequence elements in the genome called PSs are responsible for the observed effect. In addition to being important ensuring specific genome packaging, these PSs also make the assembly process more efficient (Ford et al., 2013; Dykeman et al., 2014). In *in vitro*

experiments CPs of some viruses can spontaneously assemble given the right conditions. However, addition of genomic RNA accelerates this process and makes it more efficient (Stockley et al., 2013b).

PSs are usually SLs whose apical loop portion presents a conserved motif and interacts with CPs (Figure 1.3). This interaction has a different effect in different viruses. For instance, in satellite tobacco necrosis virus PSs' main function is to overcome electrostatic repulsion between the positively charged amino-terminal arms of three CPs, which results in a more ordered conformation of these arms. In bacteriophage MS2, on the other hand, PSs are responsible for triggering a major conformational change in some of the CPs, which gives rise to heterodimers necessary to form a capsid (Stockley et al., 2013b). In viruses with many, dispersed PSs not all have the same roles requiring them to have distinct affinities to CP. One of these crucial roles is error correction in the form of dissociation and re-association. Therefore, PSs are not uniform but contain inherent variability to allow for different affinities.

In an *in silico* model replacing all native PSs with high affinity ones reduces packaging efficiency due to trapping in stable intermediates (Stockley et al., 2013b). In addition, low affinity PSs may be important for disassembly of the capsid upon entry into a new host cell. They may serve as first detachment points from which CPs start to "peel off" the RNA. High affinity PSs, on the other hand, are few and serve as nucleation sites of assembly. They form strong initial contacts with CPs and trigger the process (Stockley et al., 2013b) (Figure 1.3 B and C). Another important factor for fast and specific assembly is the gradual increase in CP concentration (Dykeman et al., 2014). Usually, in *in vitro* and *in silico* experiments, CP is provided in one large dose, sufficient to package all RNA. However, *in vivo* protein is being synthesized and increases gradually. This was found to be important for efficiency and specificity of assembly in the presence of competitor RNA in *in silico* experiments. Adding all CP at the beginning results in more competitor RNA being incorporated and fewer functional

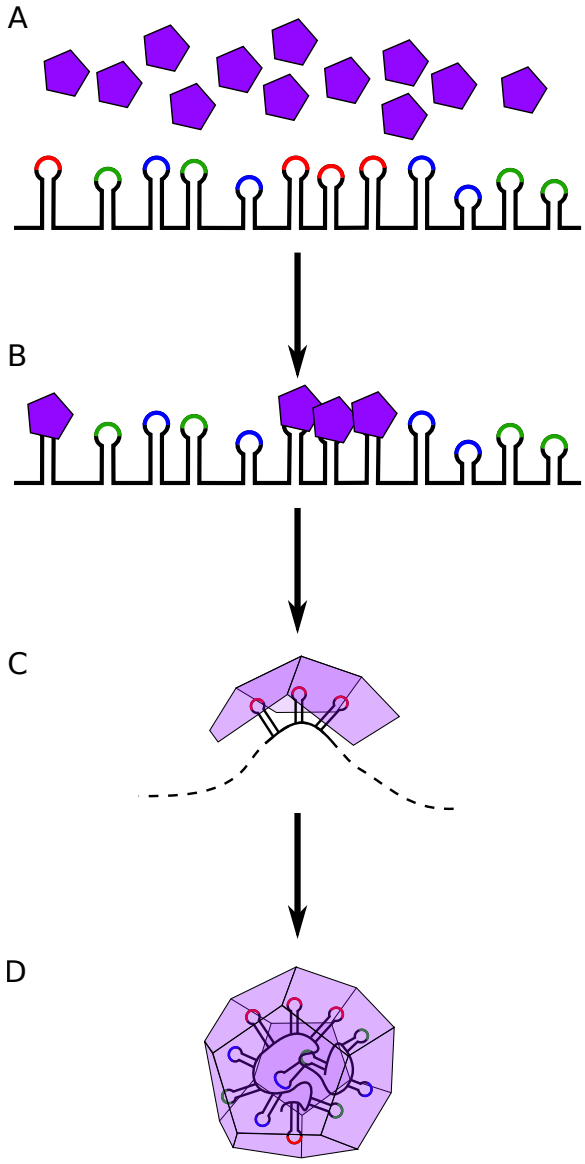

FIGURE 1.3: **Model of packaging signal-mediated assembly.** (A) Stem-loops displaying a high (red), medium (blue) and low (green) affinity packaging signal motif form on the genomic RNA. (B) When a high enough concentration of capsid protein is reached, the protein binds stably to high affinity packaging signals (red). This effect can only be observed when the capsid protein concentration is not too high, as otherwise capsid protein condenses on the RNA and the specific effect is masked. (C) Packaging signals aid in capsid protein interaction and the capsid begins forming. (D) As medium (blue) and low (green) affinity packaging signals bind to capsid protein assembly can complete.

viral particles being produced (Dykeman et al., 2014).

1.1.3 Packaging Signals in DNA Viruses?

There have been numerous studies on PS-mediated assembly in mostly ssRNA viruses; however, whether these structures also occur in DNA viruses that package an RNA pregenome is poorly understood. Such viruses can be subdivided into retroviruses such as human immunodeficiency virus (HIV) and pararetroviruses such as hepatitis B virus (HBV) (Temin, 1985). While both groups reverse transcribe an RNA intermediate into DNA for replication, when this step occurs in the life cycle differs. Retrovirus particles contain RNA and it is only within the new host cell that it is reverse transcribed into DNA, which then serves as a template to generate mRNA and genomic RNA for progeny viruses. Pararetroviruses on the other hand package RNA but already generate DNA within the viral particle before leaving the host cell (Temin, 1985). New research shows that *lentiviruses* such as HIV also reverse transcribe within their fullerene core but they do so in the nucleus of the next host cell (Jacques et al., 2016). Unlike most retroviruses, pararetroviruses do not integrate their DNA into the host genome as part of their normal life cycle but only as defective forms (Temin, 1985; Dejean et al., 1984).

There is evidence about one PS in HIV called ψ. Recent studies have revealed the SL structure of the 5'-leader region of the HIV RNA genome and showed that it is responsible for specific packaging of the correct form of the genome as well as inhibition of translation (Keane et al., 2015). The latter is a function often displayed by the highest affinity PSs in ssRNA viruses such as MS2. Its highest affinity PS is even named TR after translation repressor as binding of CP to this SL inhibits further translation of replicase, i.e. the viral RNA-dependent RNA polymerase, and maturation protein, a protein crucial for binding to new host cells (Valentine and Strand, 1965; Haruna and Spiegelman, 1965; Lodish and Zinder, 1966; Viñuela et al., 1967; Nathans et al., 1969; Carey et al., 1983a,b; Beckett and Uhlenbeck, 1988; Rolfsson et al., 2008).

Less is known about PSs in pararetroviruses. In HBV and related viruses it is believed that the stem-loop ε performs a similar function. However, as opposed to ψ in HIV, packaging is mediated through interaction with polymerase, which also ensures encapsidation of this protein (Bartenschlager and Schaller, 1992). For a long time no other PSs were thought to exist in HBV (Junker-Niepmann et al., 1990). The regions of CP with high affinity for RNA and DNA are assumed to aid in reverse transcription and bind these molecules unspecifically (Hatton et al., 1992). Whilst the evidence for ε being the only PS is strong and convincing, the studies ignored some alternative explanations for the results obtained. This leaves the option that there are indeed PSs in HBV other than ε as will be shown in this thesis.

1.1.4 Different Functions of Packaging Signals

Some PSs perform a double role in that binding of CP ensures not only packaging of the correct RNA but also switches off translation and frees the RNA from ribosomes, which are the enzymes performing the translation of mRNA into animo acid sequence (Alberts et al., 2002b). A mechanism like this is seen in MS2 with TR (Lodish and Zinder, 1966; Viñuela et al., 1967; Nathans et al., 1969; Carey et al., 1983a,b; Beckett and Uhlenbeck, 1988; Rolfsson et al., 2008), in HIV with ψ (Keane et al., 2015), and in HBV with ε (Nassal et al., 1990). In reverse-transcribing viruses, however, there are additional challenges to be considered. Once inside the capsid the RNA has to serve as a template for DNA synthesis. It would, therefore, be impractical for these viruses to have many high affinity PSs bound to the inside of the capsid. Moreover, they may serve an additional role in regulating reverse transcription as well, similar to their involvement in knocking down translation prior to packaging so that viral DNA is not exposed to host immune sensors. These additional functions would usually coincide with the triggering of assembly and encapsidation, i.e. the nucleation of assembly. The respective PSs together with the proteins that bind them would form a nucleation

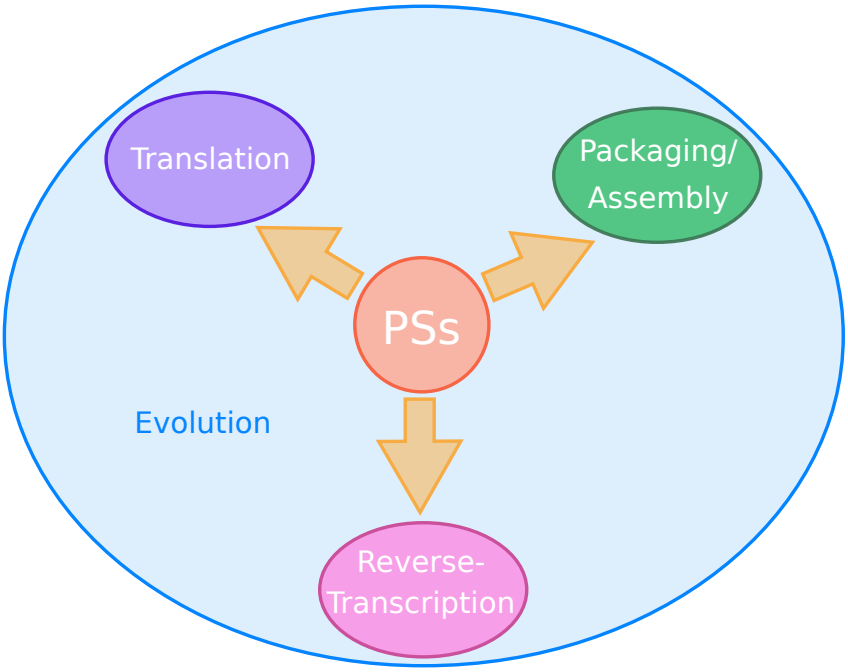

FIGURE 1.4: **Packaging signals can perform multiple functions.** Packaging signals are mainly studied for their involvement in packaging and assembly (green). However, they are known to also affect other parts of the viral life cycle such as translation of proteins where the packaged RNA also acts as mRNA (purple). In (para)retroviruses reverse-transcription of the pregenomic RNA should also be considered (pink). The constant evolution of the viral sequence affects PSs, whilst at the same time all of these crucial functions limit the possibility of change in the respective areas leading to conserved regions.

complex from which assembly commences (Beckett and Uhlenbeck, 1988) and which may set capsid geometry (Selzer et al., 2014). It is usually assumed that only one PS performs the additional function resulting in a higher evolutionary pressure on it. This PS is, therefore, more likely to be highly conserved between strains or even species (Figure 1.4).

1.2 Aims and Objectives

The aim of this project was to elucidate conservation and evolution of PSs and use these insights to expand our understanding of the nucleation of PS-mediated assembly and what could be learned from it about other, broader PS functions.

To this end, I have

1. developed a novel phylogenetic algorithm that visualises PS distributions and reconstructs phylogenetic trees based on these;
2. analysed the PS distributions and identified areas of high conservation;
3. mapped conserved PSs to available sequence as well as secondary and tertiary structure information;
4. deduced nucleation complex PSs from these mapped positions;
5. and tested the hypothesised nucleation complex through a computational assembly model.

1.3 Scope

The focus of this work was on packaging signal-mediated viral capsid assembly of RNA-packaging viruses and what could be learned about this crucial step in

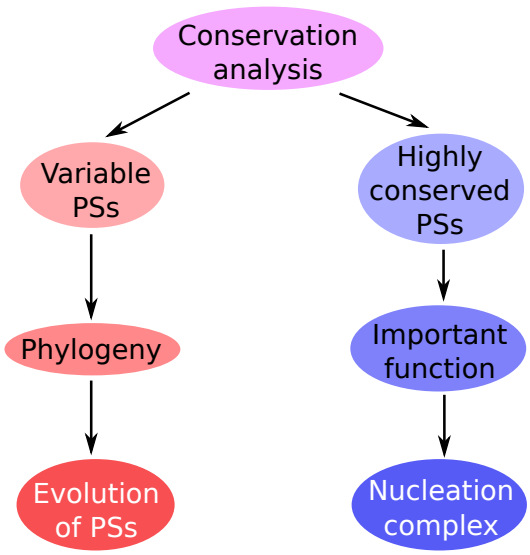

FIGURE 1.5: **Uses of conservation analysis.** Two paths for conservation analysis were explored in this project. Features, in this case packaging signals (PSs) can either be found variable (red) or highly conserved (blue). The variable ones can form the basis for phylogenetic analysis, which provides insights into how PSs evolve. Highly conserved PSs, on the other hand, indicate a more crucial functional role and are hypothesised to be involved in a nucleation complex.

the life cycle of a virus through the means of conservation or the lack thereof. As described above and illustrated in Figure 1.5, features are interesting to study as either variable or conserved. The variable parts can inform a phylogeny and thus be utilised to learn about how the feature evolves and how viral strains are related on that level. To this end, a new method was developed that allows the reconstruction of phylogenetic trees based on the presence and absence of packaging signals at aligned parts of viral genomes. In the process it includes a novel approach to global secondary structure prediction of a viral genomic RNA by taking packaging signal affinities into account and can be used to identify highly conserved features (Chapter 2). This method was applied to hepatitis B virus (Chapter 4) as well as *leviviruses* MS2 and BZ13 to gain insight into viral evolution on the packaging signal level, which was predicted to proceed at a different pace than on the level of nucleotide sequence (Chapter 6). Moreover, their genomes were examined for sites of high packaging signal conservation. A feature that is highly conserved is likely to have an especially important function or may even perform a double function. Packaging signals in conserved regions are thought to play crucial roles in the nucleation of assembly, i.e. they are involved in making the first contacts with capsid protein and trigger the assembly process, or may perform a double function as repressors of translation/reverse transcription. For hepatitis B virus, the packaging signal motif was first identified from experimental data using my bespoke algorithms and later confirmed experimentally (Chapter 3). Prediction of nucleation complex packaging signals in hepatitis B virus resulted in a novel hypothesis of reverse transcription regulation and the prediction of packaging signals in other viruses of this family without experimental data (Chapter 5). Broadening the scope, packaging signal conservation analysis in *leviviruses* was expanded to include Qβ, an *allolevivirus*, of the same family, *Leviviridae*. This identified a small set of packaging signals that are in close proximity to each other and maturation protein in the three-dimensional structure and were thus predicted to form a nucleation complex together (Chap-

ter 6). Incorporating these insights into a computational model of capsid assembly in MS2 led to a significant improvement of the capsid yield, demonstrating the crucial role of a large nucleus in MS2 and related virus assembly (Chapters 7 and 8).

Chapter 2

Phylogenetic Algorithms

In this chapter I will describe the history of phylogenetics, what we understand under the term today, and how phylogenetic trees are commonly built. Most of the basic information about phylogenetic trees and tree building algorithms provided is based on Chapters 7 and 8 in Zvelebil and Baum (2008). I will then give an example of phylogenetics based on an entire protein fold rather than primary sequence, and finally describe my own approach based on packaging signals. The basis of this algorithm will be tested on bacteriophage MS2.

2.1 History of Phylogeny

The study of phylogeny goes back hundreds of years as we have attempted to understand evolutionary history. Over time there were conflicting theories of ontogeny and phylogeny as our understanding of biology and our means to observe it have developed. Even the term "evolution" has come to carry a very different meaning to its first use in biology in the 18th century by preformationists. They believed that the whole organism is preformed within the egg/seed of the parent and while the embryo develops its features are "unrolled" (Latin: *evolvere* to unroll). It stood in contrast to epigenesis, the now accepted idea that embryonic development goes through stages of development and differentiation (Gould,

1977).

In the 19th century the most influential theories of phylogeny were developed by Ernst Haeckel and Karl Ernst von Baer. Haeckel believed that the ontogeny of an organism mimics its phylogeny, i.e. that the stages of embryonic development for a species recapitulate its evolutionary history: "recapitulation" theory (Hall, 2003). For instance, a frog's early stages appear like adult fish indicating the evolution of fish to frogs. Von Baer, on the other hand, postulated that species with similar embryonic stages have not evolved from each other but rather share a common ancestor. Differentiation of species is due to changes in development (Brauckmann, 2012). He famously coined what is known as Baer's law of embryology that, amongst others, postulates that while embryos of related species go through similar early stages, specific characteristics begin to emerge from the more general ones. More complex organisms do not go through stages of simpler adult organisms (discussed in detail in Abzhanov (2013)). A frog embryo does not resemble an adult fish.

Some of the first known phylogenetic trees were drawn during the 19th century. A sketch can be found in Charles Darwin's first notebook (Figure 2.1) as part of his theory of evolution. Later scientists also used trees to represent their views of phylogeny. A known example of an actual phylogeny based tree is by recapitulationist Ernst Haeckel (Figure 2.2). In his more artistic representation he visualised his postulated relatedness of members of the animal, plant, and protozoa kingdoms. This illustration is one of the first known representations of a tree of life. While von Baer also drew a phylogenetic tree, his sketch is less known and not widely available (Brauckmann, 2012). These trees and phylogeny in general was traditionally based on morphological features of organisms. In these cases much stemmed from Haeckel's and von Baer's views on ontogeny and how it relates to phylogeny. Other features used to classify species were measures such as their height or the length of a certain bone. Only on the mid-20th century did molecular features start to take over and molecular phylogenetics, as we know

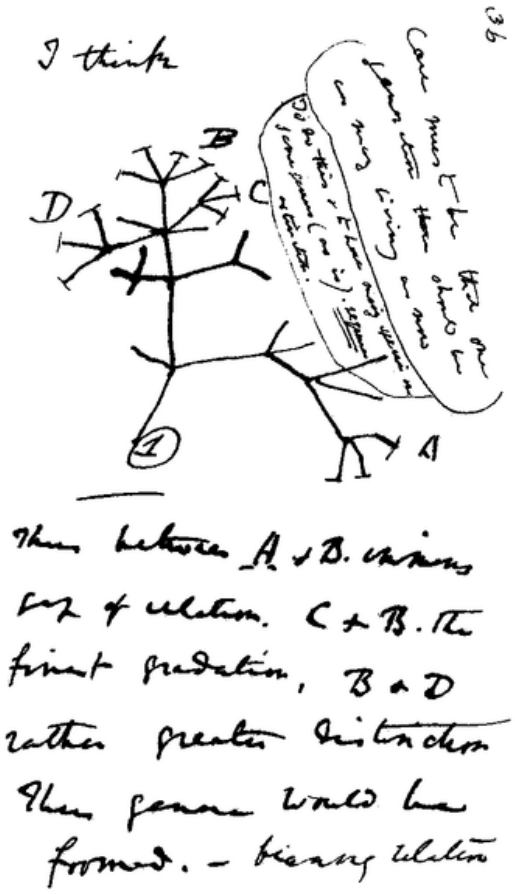

FIGURE 2.1: **Sketch of a phylogenetic tree by Charles Darwin.** Adapted from his first notebook on transmutation of species from 1837.

it today, started to develop (reviewed in Suárez-Díaz and Anaya-Muñoz (2008)).

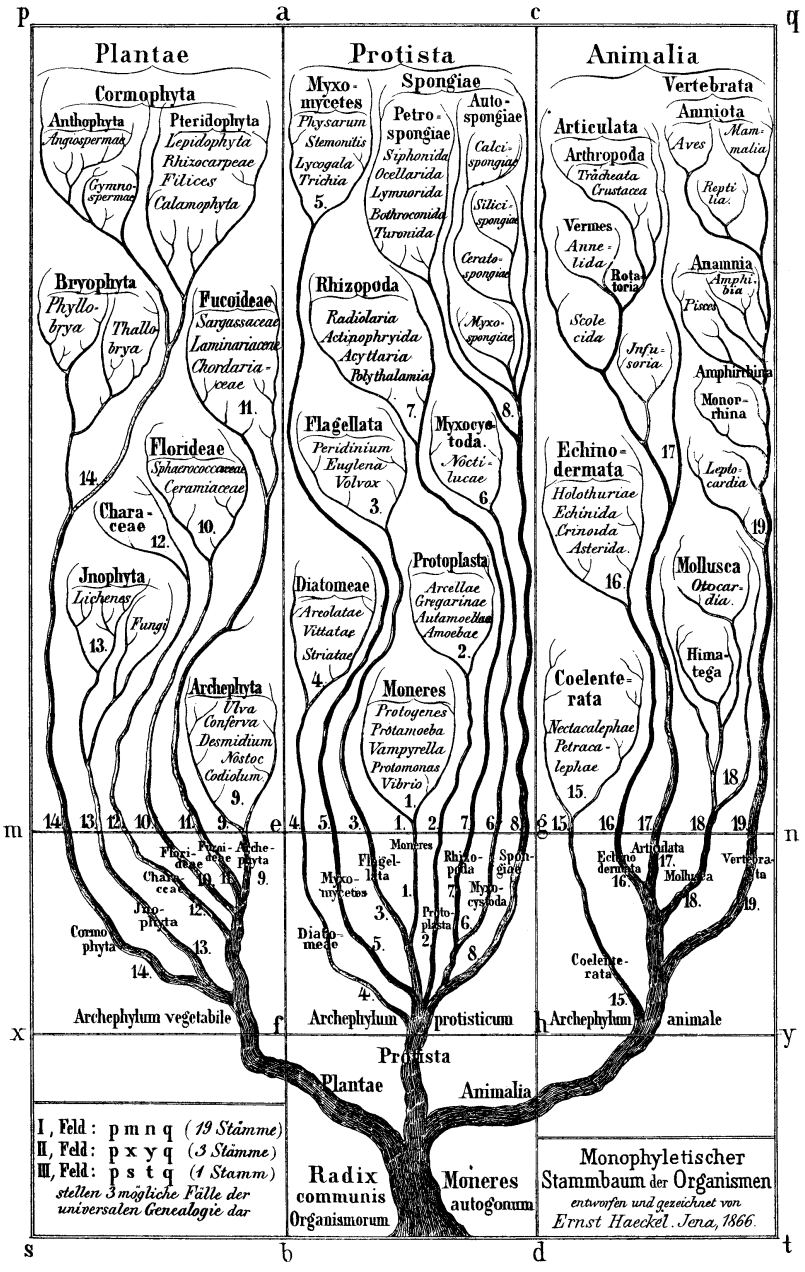

FIGURE 2.2: **A phylogenetic tree by Ernst Haeckel.** This picture can be found in "Generelle Morphologie der Organismen" from 1866 and is one of the first trees of life.

2.2 Modern Phylogenetics

Modern molecular phylogenetics relies mostly on multiple sequence alignments (MSAs) of amino acid or nucleotide sequences. However, in principle any molecular feature comparison can be used such as presence/absence or order of specific sites (Zvelebil and Baum, 2008, Chapter 7). Differences between the input sequences are used to calculate a genetic or evolutionary distance between them. There are a number of different evolutionary models that are used as the basis for the calculation of this distance matrix. They provide a measure for how much change in sequence is expected in a certain amount of time. The distance matrices or, alternatively, the alignments directly are used for tree building.

2.2.1 Characters

The basis of classic as well as modern phylogenetics are characters. While there is an intuitive understanding of what constitutes a "character" among phylogeneticists, a straight forward definition is difficult to find (discussed in Wiley (1981, Chapter 5)). For the purpose of this work, a character in the context of phylogeny can be understood as an attribute of a taxon by which it is compared to other taxa. These can take on many forms. Commonly they are nucleotide positions in a MSA but any feature can be used as a character for tree building such as the presence or absence of legs or their number. In order to be able to construct phylogenetic trees, some characters have to be in different states (Warnow, 2017, Chapter 4). These are called "informative characters", because by their difference they provide information about the evolutionary history and relatedness of the taxa. Characters that are the same across taxa are neither helpful nor informative.

2.2.2 Types of Phylogenetic Trees

A phylogenetic tree is a type of graph consisting of edges (branches), nodes (leaves), and splits. It serves as a representation of the evolutionary relationship, shown through edges, between the taxa it is based on, shown at the external nodes. Splits represent the ways in which the taxa can be "split" into subsets and often correspond to the edges in a tree (Huson and Bryant, 2006). These taxa can be single genes/proteins, whole genomes or more complex features. The connection between the taxa is via internal nodes, which represent a hypothetical common ancestor between them. Depending on the taxa this can mean different things, e.g. a speciation event, i.e. one species evolved into two, or mutation in sequence, which gave rise to two different versions, or a duplication/deletion event. The topology of the tree describes the way it branches. For a given set of taxa there are a number of different possible tree topologies, i.e. ways and orders in which evolutionary events could have occurred. The aim of constructing a phylogenetic tree is to find the one that describes actual evolutionary events as accurately as possible (Zvelebil and Baum, 2008, Chapter 7).

A phylogenetic tree can be either rooted or unrooted. While both provide insights into the relative relatedness of the taxa, only rooted trees also include information about the evolutionary direction. They specifically imply an order in which the species have split and identify a most recent common ancestor (MRCA). Unrooted trees do not make inferences about the history of the relationship and simply show the relative relatedness of current species. Rooted trees are usually achieved by use of an outgroup, i.e. a taxon, which is assumed to be distantly related to the remaining ones, the ingroup (Zvelebil and Baum, 2008, Chapter 7).

The simplest phylogenetic tree is a cladogram. Cladograms show the branching of taxa, but the branch length itself has no meaning (Figure 2.3A). They do not provide information about how recent a split is or how much the taxa have diverged from each other. Additive trees, on the other hand, represent evolutionary distance through branch length. It can be calculated by adding the lengths

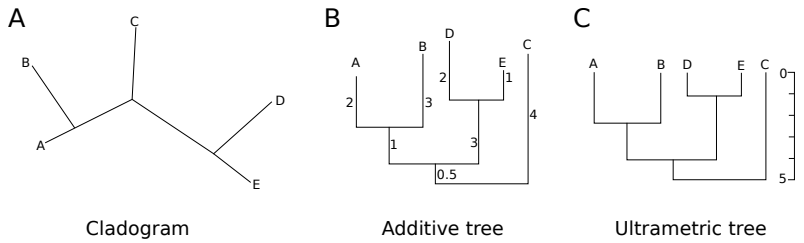

FIGURE 2.3: **Types of phylogenetic tree.** (A) A cladogram, which shows only connectedness. (B) A rooted additive tree shows relatedness as well though the numbering on the edges. (C) An ultrametric tree is always rooted. All nodes are assumed to have diverged from the MRCA at the same point in the past. Evolutionary time can be read on the side.

of the branches that connect the two taxa in question. It is thus possible to gain insight into how much the taxa have diverged; however, no information about time scale can be deduced (Figure 2.3B). This is possible in ultrametric trees, which are based on a constant rate of mutation so that evolutionary time can be calculated from divergence. As opposed to the two previous types, ultrametric trees are always rooted and all leaves have the same distance from the root, the MRCA. The time when taxa split can be read from the position of the internal node in the tree (Figure 2.3C) (Zvelebil and Baum, 2008, Chapter 7).

Trees assume evolution can only occur one way: one species split from another. However, there are also events in biology of mixing of species to give rise to new ones such as recombination or horizontal gene transfers. These types of events give rise to reticulations and thus cannot be visualised in a phylogentic tree. Instead, there is a need for a phylogenetic network (Warnow, 2017, Chapter 10). Constructing such networks is not trivial and there are a number of different methods to do so, which will not be discussed here. In this work, phylogenetic networks were not used.

2.2.3 Tree Building Algorithms

Over the decades a plethora of algorithms have been developed to construct phylogenetic trees from MSA data. Generally there are two types of methods: phenetic

and cladistic. Phenetic methods compare the sequences pair-wise and use those dissimilarities to build up a tree step-by-step. They require the calculation of a distance matrix and always produce a single tree, which is not evaluated. Cladistic methods on the other hand take the entire MSA into account directly. They are based on characters, i.e. single sites in the alignment, instead and attempt to evaluate all possible trees and identify the optimal one. As such there is no distance matrix involved and several trees with different topologies are produced in the process, which are tested against each other (Zvelebil and Baum, 2008, Chapter 7).

The most used phenetic methods are unweighted pair-group method using arithmetic averages (UPGMA) and neighbor-joining (NJ).

UPGMA starts out by assuming that each taxon is its own cluster. It then joins the two closest clusters and re-calculates the distance of the joint pair by taking the average. This process is repeated until all taxa are connected into a single cluster (Sokal and Michener, 1958). This method works under the molecular clock hypothesis, i.e. all sequences are assumed to have evolved from the MRCA at the same rate. It thus results in rooted ultrametic trees, whereas all other methods discussed below produce unrooted additive trees (Zvelebil and Baum, 2008, Chapter 7).

NJ begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branch lengths of the resulting tree is calculated. The pair with the smallest sum is considered the closest neighbours and joined (Saitou and Nei, 1987). The goal is to produce a tree that implies minimum evolutionary steps (Gascuel and Steel, 2006). It is a rigid method and consistent if only a small level of noise is present in the data. However, it is not robust against the presence of distant taxa, i.e. the topology of a NJ tree for a set of closely related taxa changes when a very distant taxa is added (Bruno et al., 2000).

A refined form of NJ is Weighbor, i.e. weighted neighbor-joining. It combines the additivity of external branches with a positivity term of internal branches

to quantify the implications of joining a pair. This makes the method less sensitive to specific biases than NJ and relatively immune to long branch attraction (LBA)/distraction drawbacks (Bruno et al., 2000). LBA describes the phenomenon when two or more taxa that are distantly related to the rest, i.e. appear on long branches in the tree, are clustered together: the long branches attract each other. They are usually fast evolving sequences, which, as a result of only four possibilities in nucleotides, have similar mutations to each other by chance and thus show convergent evolution (Felsenstein, 1978; Philippe et al., 2005).

Fitch-Margoliash (FM) tries to fit trees to a distance matrix. This matrix is constructed by counting the minimum number of nucleotide mutations required to change any differing amino acids in one taxon to the other. Starting out with each taxon as its own subset, the closest by mutation distance are joined and an average of their distances is used in the next step. More trees are generated by allowing alternative joints by a set threshold value. These are compared with each other and the tree with the least-squares fit of the pair-wise mutational distances between the tree and the matrix values (Fitch and Margoliash, 1967; Suárez-Díaz and Anaya-Muñoz, 2008).

The most used cladistic methods are maximum parsimony (MP) and maximum likelihood (ML).

MP prefers the simplest explanation of data, i.e. the tree with the fewest substitutions/evolutionary changes for all sequences to derive from a MRCA. For each site in the alignment all possible trees are evaluated and scored for the number of evolutionary changes needed. The best tree minimises the overall number of mutations at all sites. However, this method provides little information about branch length and suffers most prominently from LBA (Philippe et al., 2005).

ML uses each position and evaluates all possible trees. The likelihood for each tree is calculated and the tree with the maximum likelihood is determined by evaluating the probability that a certain evolutionary model generated the

observed data. It is computationally intensive and slow but tends to yield the best results (Warnow, 2017, Chapter 8).

Bayesian methods are similar to ML in that statistics is involved in choosing a tree. However, while ML tries to identify the tree that is most likely to give rise to the data used the Bayesian method samples from the set of trees based on their probability. It thus results in a set of trees rather than a single "optimal" one.

The final cladistic method introduced here is Quartet Puzzling. While there is only one unrooted phylogenetic tree of two or three taxa, there are several options for sets of four taxa. These smallest informative phylogenetic trees are called quartets. They can be pieced together to build up a larger tree. The original algorithm describes a three-step process. First, ML is utilised to construct all quartet trees. These are joined into complete trees using one arbitrary quartet as seed. At each step when another taxon is added, a majority vote from all quartets decides its position. Using different seeds gives rise to a set of independent trees. Out of these optimised trees, the final tree is picked by looking at which topologies occur in most of the trees in the set (Strimmer and von Haeseler, 1996).

2.2.4 Mutation Models for DNA

Calculating evolutionary distance from DNA sequence alignments requires the use of a mutation model, i.e. an estimate of how likely a substitution is. Phenetic tree building methods utilise these models to convert differences in the alignment to adjusted evolutionary distances, whereas cladistic methods utilise the models in the entire tree-building process. The choice of mutation model therefore has a high impact on the outcome.

The simplest method is to calculate the Hamming distance (Hamming, 1950), which in this application means to count the number of differences between the sequences. This method risks severely underestimating the evolutionary distance between taxa and would only be suitable for very recently diverged taxa. The

reason is that as more time passes a given site is likely to have mutated more than once. What may appear to be a single mutation event may have actually been several. It also ignores the fact that mutations do not occur at random in an organism's genome. For one a mutation can be synonymous, i.e. the codon is changed but it still encodes the same amino acid, or non-synonymous, i.e. a codon changes to another amino acid. Synonymous mutations rarely have an effect on protein expression and can be considered neutral. They are therefore found to a larger extent than non-synonymous mutations. Some sites are under a higher selective pressure than others due to coding for protein or performing regulatory functions. Changes to important functional parts of the DNA are more likely to not be viable and therefore not be found in today's taxa. To avoid this trap, several distance correction methods have been developed over the years. The first widely used substitution model is Jukes-Cantor (JK69), which assumes all mutations to be equally likely (Jukes and Cantor, 1969). The corrective formula to calculate genetic distance K from the percent difference p is $K = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$. The next updated method was the Kimura 2-parameter (K2P) method. It assumes that transitions, i.e. purine to purine or pyrimidine to pyrimidine, are more likely than transversions, i.e. purine to pyrimidine and vice versa (Kimura, 1980). It has in fact been shown that this is the case and transitions occur at a much higher rate than transversions (Janecek et al., 1996); however, the exact rate of transition to transversion can vary greatly between sequences. This was further improved upon in the Felsenstein 81 (F81) method, which adds considerations for unequal base frequencies (Felsenstein, 1981). The F84 method built on this and also included different transversion and transition rates (Felsenstein and Churchill, 1996). It is thus very similar to the Hasegawa, Kishino and Yano 1985 model (HKY85), which considers the same extensions to JK69 (Hasegawa et al., 1985). The Tamura 1982 (T82) model is also based on K2P but considers specifically G-C content (Tamura, 1992). Kimura himself expanded on his model by including a third parameter giving rise to the Kimura

3-paramter (K3P) model. Like F84 and HKY85 it also takes into account differing base frequencies but adds two different rates for transversions (Kimura, 1981). It is in that way similar to the Tamura and Nei 1993 model (TN93), but this model assigns different frequencies to the two types of transitions (Tamura and Nei, 1993). Finally, the probably most complex model is the Generalised time-reversible model (GTR), which considers any nucleotide change to be reversible and different rates for each type of nucleotide substitution are possible (Lanave et al., 1984; Tavaré, 1986; Waddell and Steel, 1997). It was adapted to account for different rates across sites by Waddell and Steel (1997). This is often the model of choice and produces the best results (Rodríguez et al., 1990). A recent review on the most commonly used substitution matrix methods can be found in Arenas (2015).

2.2.5 Bootstrapping

A common measure of phylogenetic tree robustness is the so-called bootstrap analysis or bootstrapping. The method was first developed by Felsenstein (1985) and improved by Efron et al. (1996) and later Holmes (2003). Once a tree has been generated from the complete dataset, the process is repeated a high number of times with a random subset of the data. This subset is generated by sampling the characters with replacement from the original data set (Warnow, 2017, Chapter 8). The splits within the full tree are then compared with those in the bootstrapped trees. The more often a given split occurred in the subset trees, the more confidence can be had that it is real. Each split is given a bootstrap value, which is the percentage of sampled trees it was found in (Zvelebil and Baum, 2008, Chapter 7). While it is a useful measure for split support, its interpretation is not simple and sometimes, depending on the tree-building methods used, a high bootstrap value may not necessarily indicate a high accuracy. Nevertheless it is often applied to provide some information on the robustness of the phylogenetic tree, and generally values below 50% are considered unreliable, while values

above 95% are desirable (Warnow, 2017, Chapter 8).

2.3 Bamford-Stuart's Protein Structure Phylogeny

Up until now much of the discussion on modern phylogenetics had focused on phylogenies based on MSAs of amino acid or nucleotide sequences. However, as mentioned in the beginning of this chapter, phylogenies can be reconstructed from any (molecular) feature. Depending on the type of feature that is being compared a different window of evolutionary history can be resolved. The fastest changing is nucleotide sequence. Due to it changing so quickly, it is adequate for looking at fairly recent evolution but becomes less reliable the further back a split between taxa lies. The next level is amino acid sequence. Since changes in nucleotides can occur without affecting coding, i.e. silent mutations or mutations in non-coding regions, it evolves more slowly. There is also the added complication that changes to the amino acid sequence may alter the protein's ability to function. This may result in an individual that is not viable and cannot pass on this mutation. Therefore, many mutations, which might have occurred over time, would not be visible in today's species. Most observable changes would have been neutral or positive in nature. While single nucleotide mutations can also be detrimental to the organism as can be seen in many Mendelian diseases, i.e. disorders caused by dysfunction of one gene, such as Duchenne muscular dystrophy (dystrophin) (Aartsma-Rus et al., 2006) or sickle cell anaemia (haemoglobin) (Piel et al., 2017), these are relatively rare. Going a step further from amino acid sequence is protein structure and finally function. These may require several changes on the nucleotide level translated to the amino acid level, which change the way a protein folds and behaves. Except for detrimental diseases such as the ones mentioned above, neutral or positive change in a protein would happen over a long period of evolutionary time. Conversely, protein structure/function tends

to be robust against many amino acid changes as different amino acids are similar enough to each other to preserve it. Two coding sequences can thus be markedly different, while the proteins they code for have many striking similarities in their structure and/or function. If a common ancestral relationship can be established between them, they are considered homologues, the product of divergent evolution. However, it is also possible that they are in fact not related, but are the product of convergent evolution and would thus be called analogues. So, while comparing structure and especially function of a protein can provide insights into ancient evolutionary events, care needs to be taken to not mistake analogues for homologues.

The above evolutionary time scales (nucleotide sequence, amino acid sequence, and protein structure) are different between species. This is due to differences in the time for reproduction. Changes in structure depend on changes in amino acid sequence, which in turn are caused by changes in nucleotide sequence. These will occur and accumulate over generations to which each is passed on, but length of time for a generation strongly depends on the species. While a new human generation takes 20–30 years, in cats or goldfish it is close to one year, and in many bacteria can be a matter of an hour. The faster a new generation is made, the faster nucleotide changes can occur and accumulate so they would be faster to evolve. Thus, comparing based on nucleotide sequence alone would become more difficult faster. Therefore, amino acid sequence of one protein or even of only certain domains of proteins are often the basis of comparisons between more distant species.

Viruses belong to the very fast evolving species. The generation time, i.e. the time from infection of a host cell to release of progeny virus, varies between different viruses and ranges from an hour to days or even a few weeks. For example, in bacteriophage MS2 it takes just 60–90 minutes, in human immunodeficiency virus (HIV) two days, and in hepatitis B virus (HBV) it can take 10–100 days (Propst Ricciuti, 1976; Nowak et al., 1996; Perelson et al., 1996). Evolution is

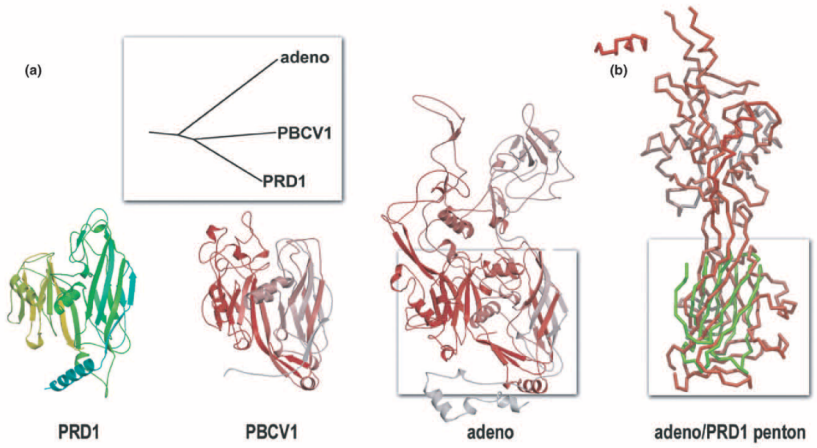

FIGURE 2.4: **Comparison of X-ray structures of PRD1-like lineage.** (a) Individual x-ray structures of viruses within the PRD1-like lineage (in orange in Figure 2.5). (b) Overlay of structures showing striking similarities. Figure 2 as published in Bamford et al. (2005). Copy right cleared with Elsevier through Copyright Clearance Center, license number 4491930039368.

further accelerated by error-prone replication enzymes, and the sheer number of offspring virus allowing for a wide array of potential mutations. Whilst nucleotide sequence MSAs are nevertheless suitable especially to look at viral strains of one species or very closely related species, this method quickly falls short the further up the taxonomic rank one ventures. Even aligning amino acid sequence can become difficult on the family level (see also Chapter 5).

Classifying viruses and understanding their evolutionary relationships on a broader level therefore benefits from a different approach. One such complementary approach has been introduced by Bamford and Stuart, which suffers less from these problems and offers a different angle: the grouping of viruses by the folds of their capsid proteins (reviewed in Bamford et al. (2005)). While other viral proteins, such as reverse transcriptase or integrase, can be present or absent depending on species, all particle-forming viruses require a capsid to package their genomic material and potentially other proteins. It is consequently an ideal candidate to compare and classify different viral species by, and lends itself for even the most distant comparisons.

Bamford *et al.* were most interested in uncovering the evolutionary ties

between viruses that infect hosts from different domains of life, i.e. eukarya, prokarya, and archaea. To that end they grouped viruses into lineages based on a small number of different capsid protein folds. The idea was first introduced when the structure of bacteriophage PRD1 capsid protein was found to exhibit striking similarities to the respective human adenovirus protein (Benson et al., 1999). The theory was brought forward that viruses have in fact been around since before the split into the three domains of life (Bamford et al., 2002; Bamford, 2003). Moreover, they concluded that the different viral lineages do not share a common ancestor but are polyphyletic (Bamford, 2003). The PRD1-like lineage was further expanded to also include Bam35 (prokaryote host), STIV (archaea host) (Benson et al., 2004), and PBCV1 (eukaryotic host) (Bamford et al., 2005). Figure 2.4 further illustrates the striking similarity in the capsid protein folds of three members of this lineage: PRD1 (bacteria), PBCV1 (algae), and adenovirus (human). While the adenovirus protein includes additional domains, the core fold is highly similar.

Having uncovered these similarities, Bamford *et al.* have applied the idea to other viruses and reconstructed a phylogenetic tree based on the similarities in the folds (Bamford et al., 2005). The PRD1-like lineage described above and two additional ones, as well as a control group (dark blue), are shown in Figure 2.5. The protein folds are characterised as PRD1-like, HK97-like, resembling bacteriophage HK97, and BTV-like, resembling eukaryotic virus BTV. Interestingly, also the HK97-like lineage includes viruses that infect prokaryotic, archaean, and eukaryotic hosts (Bamford et al., 2005). The basis for the phylogenetic tree is a gap-penality-weighted superposition of the structures, which means all the structures were superposed on each other and it was then scored how closely residues were to each other in 3D. This resulted in a set of probabilities, which were converted into evolutionary distances. From there a tree could be reconstructed using standard methods (Bamford et al., 2005). This demonstrates that the PRD1-like lineage is not the only one crossing domains. Note, that the relationship between

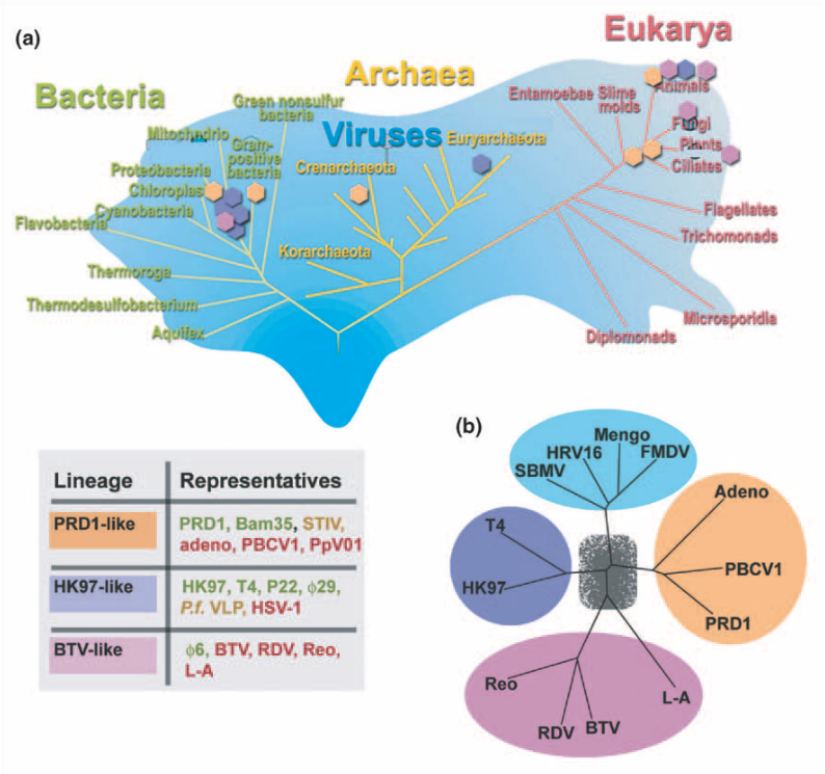

FIGURE 2.5: **Viruses can be grouped into four lineages by their capsid protein fold.** (a) Viruses infect cells from all domains of life: bacteria, archaea, and eukarya. Hosts of viruses from each lineage are marked by hexagons in the respective colours. (b) Phylogenetic tree of viral lineages based on capsid protein structure. Figure 1 as published in Bamford et al. (2005). Copy right cleared with Elsevier through Copyright Clearance Center, license number 4491930039368.

these distant viruses could not have been uncovered based on nucleotide or amino acid sequence. These differ so much for these proteins between the species that an evolutionary relationship is not recoverable. Looking beyond simple sequence comparison can thus provide new insights into the evolution of viruses. However, it is also possible that the driver is convergent evolution due to the limited geometric repertoire of capsid architectures (Twarock and Stockley, 2019).

2.4 Phylogeny Based on PS Profiles

As introduced above, phylogenetic relationships between species and strains are mostly understood based on the difference between their genomes. More mutational differences mean greater distance in the phylogenetic tree. Bamford et al.

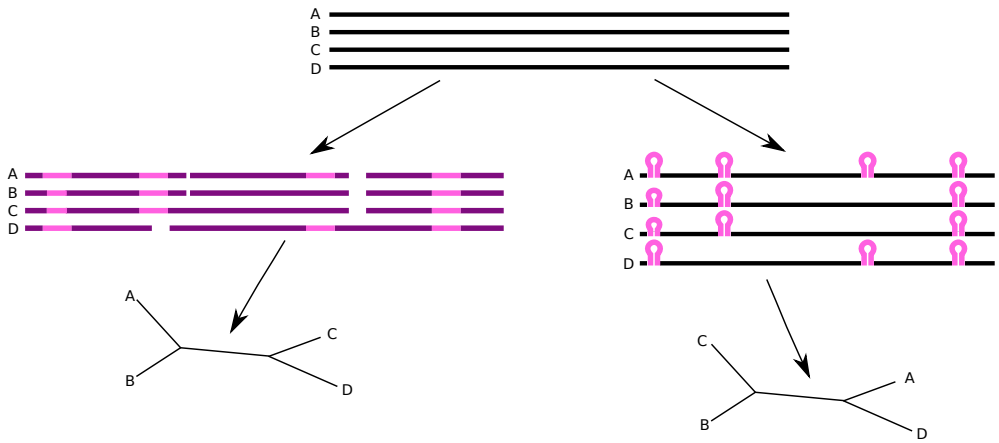

FIGURE 2.6: **The concept of PS-based phylogenetic trees.** Given a set of genomic RNA sequences A,B,C, and D the currently common way to generate a phylogenetic tree is to align the sequences. Then, based on similarities and differences in the aligned nucleotides, a phylogenetic tree is built (left). Utilising PSs for the process requires the sequences to first be folded into a set of non-overlapping stem-loops (SLs) (pink). The basis for tree building is then the presence or absence of a PS in a given position. This may result in a different tree topology (right).

(2005) revolutionised phylogeny of viral species by building a tree based on the similarity of viral capsid structure and uncovering links between viruses that infect hosts of different domains of life. Interestingly, this tree topology was distinct from the sequence based ones indicating that evolutionary relationships of viruses may be more complex than was previously assumed.

Picking up on this idea that phylogeny can be based on structural elements related to function, my aim was to produce phylogenetic trees of virus strains based on their packaging signal (PS) profiles, i.e. which types and distribution of PSs they use. This would allow us to make inferences about evolutionary relationships of viral strains from the point of view of PS-mediated assembly. These may show different topologies compared to trees based on genomic sequence itself (Figure 2.6). To simplify the process, some of the methods of molecular phylogenetics based on sequence alignments as introduced above were utilised. To that extent PS profiles were represented as pseudo-DNA sequence, so that available tree-building tools could be used.

2.4.1 Algorithms and Methods

Most stretches of RNA can potentially form into several secondary structures - some overlapping and some not. In order to generate PS profiles a set of non-overlapping SLs was required. Essentially, a global picture of secondary structures for the sequence needed to be found. However, simply running common RNA folding programs such as Mfold on the complete RNA sequence would not necessarily provide a global structure that is an accurate depiction of the real situation *in vivo*. The problem with these folding programs is that the energy parameters used to calculate the minimum free energy (MFE) structure are based on experimental measures of small hairpins and helices (Mathews et al., 1999). How well these translate to larger structures and long-range interactions is questionable. Additionally, such an MFE global structure represents the state of the RNA at a thermodynamic equilibrium, i.e. given enough time in solution the molecule would adopt this structure, but tells us little about the kinetics. Even at thermodynamic equilibrium there are many RNA folds with similar energy, i.e. a wide and shallow folding funnel. Ignoring alternative structures with only slightly worse energies becomes especially problematic considering that the MFE also does not take into account that a viral RNA *in vivo* is rarely found "naked", i.e. not bound by viral or host proteins. In the case of single-stranded RNA (ssRNA) viruses the RNA also functions as messenger RNA (mRNA). This means that after transcription the RNA would be bound by a number of translation factors and subsequently the large and small subunits of the ribosome. The eukaryotic initiation factor (eIF)4A, which is involved in eukaryotic translation, is a helicase - an enzyme that unravels nucleic acid double helices (Rogers et al., 1999, 2001). Any secondary and tertiary structures on the RNA would be removed. The prokaryotic ribosome was also found to have helicase activity (Takyar et al., 2005). Only after translation has been inhibited by a virus-specific mechanism can the RNA start to fold again presenting PSs. As translation inhibition represses the binding of new ribosomes the ones that were already attached would

finish protein synthesis. This leaves a trail of ssRNA behind the last ribosome, which can begin to form into SLs and larger structures as more RNA becomes available. This may trap the RNA in a stable alternative structure that is not the MFE structure. Moreover, SLs would be picked entirely based on their inherent stability whereas a PS would be further stabilised through its interaction with capsid protein (CP). Thus, I assumed a more accurate prediction is to fold the RNA locally into SLs, add their CP binding energies to their folding energies, and stitch these local folds together into a global picture minimising overall energy. While this method ignores longer range interactions and does not resolve tertiary structure, it is adequate for studying which SLs can be formed simultaneously. This also allowed the use of other measures of stability to be taken into account as described below.

2.4.1.1 RNA Fragmentation

The aim was to stitch together local folds into a global picture. To this end, the whole RNA genome had to be cut into smaller fragments. In order to avoid biasing the folds all possible fragments of a given size were generated. This was achieved using a sliding window approach. For a size X, the first fragment was 1 to X, the second 2 to $X + 1$, the third 3 to $X + 3$, etc until $N - X$ to N. The fragments were thus highly overlapping and effects of fragment position on possible and favourable folds could be avoided. To evaluate the effect of fragment size, three sizes were tested: 30, 60, and 90 nucleotides. The functional parts of SLs, especially PSs, are assumed to be quite small and thus fit into the 30 nucleotide window. However, it is possible that *in vivo* a larger SL forms with more or longer helix portions, which make it more stable. Some larger SLs in MS2 are predicted to be between 70 and 80 nucleotides long (Dai et al., 2017). Since it is not practicable to test every window size, 90 nucleotides were used for the largest, which should cover most full SLs, and 60 nucleotides for a medium length.

2.4.1.2 RNA Secondary Structure Prediction Using the Partition Function

The partition function Z gives the sum of the Boltzmann factors $e^{-\beta E_i}$ over all possible N states (see Equation (2.1)). The Boltzmann factor consists of Euler's number e, the thermodynamic β (see Equation (2.2), where k_B is the Boltzmann constant and T is the temperature in Kelvin), and the energy E_i of the respective state i. In this case, the states are different folds of an RNA fragment. The probability P_i for each state i at thermodynamic equilibrium can be calculated from the Boltzmann factor and Z (see Equation (2.3)). When sampling folds from the partition function, P_i gives the probability of sampling any fold. Due to its dependence on the fold's energy E_i, more stable folds are more likely to be sampled. How often a fold has been sampled out of an ensemble is thus a proxy for relative stability of that structure. At the same time, using the partition function allows a broader view of folds, especially when several structures are similar in energy. However, given a large enough ensemble even less stable, rarer structures will be sampled.

The partition function Z as calculated for all N states:

$$Z = \sum_{i=1}^{N} e^{-\beta E_i} \tag{2.1}$$

The thermodynamic β:

$$\beta = \frac{1}{k_B T} \tag{2.2}$$

Calculation of the probability P_i for a given state i:

$$P_i = \frac{e^{-\beta E_i}}{Z} \tag{2.3}$$

To exemplify the above calculations, all possible structures of the sequence `AAACCCAAAAGGGAAA` with folding energies are given in Figure 2.7. Every time a

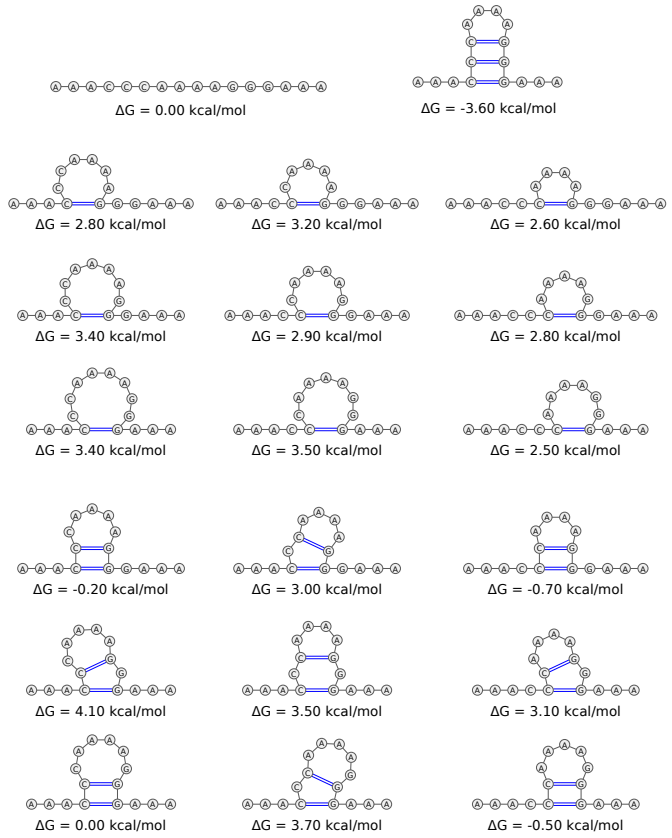

FIGURE 2.7: **Partition function example.** All possible structures with folding energies are given for the example sequence `AAACCCAAAAGGGAAA`. The structures were generated in VARNA (Darty et al., 2009) and energies calculated with the RNAeval web server (Lorenz et al., 2011; Hofacker et al., 1994; Lorenz et al., 2016).

structure for this sequence is sampled from the partition function one of these structures would be picked. The frequency at which any of them would be sampled depends to an extent on its energy. The partition function would take on the following value:

$$\begin{aligned}
\beta &= \frac{1}{0.0019872041 \text{ kcal/molK} \times 273.15\text{K}} \\
&= 1.842283 \text{ mol/kcal} \\
Z &= \mathrm{e}^{-1.842283 \times 0.0} + \mathrm{e}^{-1.842283 \times -3.6} + \mathrm{e}^{-1.842283 \times 2.8} + \ldots + \mathrm{e}^{-1.842283 \times -0.5} \\
&= 768.8067
\end{aligned} \tag{2.4}$$

The fully folded SL would have a probability of

$$P = \frac{\mathrm{e}^{-1.842283 \text{ mol/kcal} \times -3.6 \text{ kcal/mol}}}{768.8067} = 0.9874582 \tag{2.5}$$

The next lowest energy in this example is -0.7 kcal/mol. The probability for that structure would thus be:

$$P = \frac{\mathrm{e}^{-1.842283 \text{ mol/kcal} \times -0.7 \text{ kcal/mol}}}{768.8067} = 0.004723329 \ . \tag{2.6}$$

This means that with a probability of almost 99%, the fully folded structure would be sampled from the partition function in this example and the next most stable one would only be sampled 0.47% of the time.

2.4.1.3 Structure Processing

In order to be able to add affinities to PSs, these have to be identified among all the SLs in the complete set of structures. This required storing the structural and sequence information for all SLs in a format that was easily searchable using regular expressions. An in-house Fortran 90 implementation of the Mfold algorithm (Zuker, 2003) by Eric Dykeman, Tfold, in partition function mode returns a file with the fragment sequence followed by the sampled folds, one per line. The next step was to extract individual hairpins, truncated at bifurcations, and count the number of times they occur in the ensemble. Additionally, for each SL the start, end, and apical loop positions were determined, as well as the apical loop length, and the positions and lengths of all helices. This step was parallelised using GNU parallel (Tange, 2011). The pseudocode for this algorithm is shown in Appendix A (Algorithms A.1 and A.2).

After extraction of the SLs, the folds for all fragments were concatenated into one file and merged across overlapping fragment windows as follows: The process took advantage of the positional and structural information saved about each SL during the extraction step. If two SLs were identical in all these properties,

they were merged, i.e. only added once to the list of SLs, and their respective occurrences added together. When several genomes were to be analysed at the same time, this step was parallelised using GNU parallel (Tange, 2011). For pseudocode see Algorithm A.3 in Appendix A.

In addition to the structure in Vienna format and the structural properties, the output of this algorithm also includes a list of apical loop sequences and of searchable sequence/structure combined. The latter takes the form [5' helix sequence]_[5' bulge/internal loop sequence]_[5' helix sequence]__[apical loop sequence]__[3' helix sequence]_[3' bulge/internal loop sequence]_[3' helix sequence], whereby the bulge and helix parts were repeated as needed to show the entire structure. Basically, the apical loop sequence is in the middle separated by two underscores on each side. On either side are the respective helix sequences 5' and 3'. Bulge or internal loop sequences are separated from helix sequences by single underscores. The principle is illustrated on an example of an SL in Figure 2.8. This allowed searching for folds with specific sequence elements in apical loops, including internal loops and bulges and lengths of these, as well as the helix.

The files generated in the merge were used as input for selecting the SLs as above. To determine if a SL was a PS and, if applicable, to which affinity tier it belonged, the sequence/structure file was searched using regular expressions for the PS consensus motif. In regular expressions brackets represent a group of characters where there is more than one alternative separated by "|" so that "(a|b)" matches either "a" or "b". Unless a "^ " is given at the start or a "$" at the end, anything can be present before and after this matching piece, respectively. Each character represents one in the match unless otherwise specified. A single number in curly brackets defines the exact number of matches e.g. "^ a{2}$" matches "aa" but not "a" or "aaa". The number of matches in curly brackets can also be given as a range, e.g. {,2}, {2,} or {2,4} for up to two, at least two or between two and four matches, respectively. The number of matches can otherwise also be defined by "*" for zero or more, "+" for one or more, and "?"

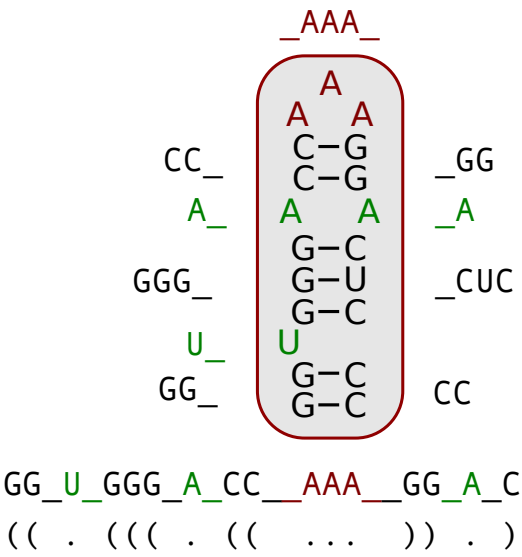

FIGURE 2.8: **Example of SL sequence/structure encoding.** The SL highlighted in grey is shown in the encoded form below. The single elements and their encoding are visualised next to the structure. They are separated by single or, in the case of the apical loop, double underscores from each other. The same colouring (black for helices, green for internal loops/bulge, and red for apical loop) is applied throughout. Below the encoding is the same structure in Vienna format.

for zero or one.

The motif and respective dissociation constants are user given. For example, the highest affinity PSs of MS2 require a single "A" in the 5' bulge and either two base-pairs followed by a four nucleotide loop or three base-pairs and a three nucleotide apical loop. The last two nucleotides in the apical loop need to be a pyrimidine ("Y") and an "A". The respective search motif as provided by the user would be: `X{2}(X_A_X{2}__X|_A_X{3}__)XYA__X{5}`. This expression matches any SL that has any nucleotide followed by one pyrimidine and one "A" in the apical loop then three basepairs with any nucleotides and a single "A" in the bulge followed by at least two basepairs (`X{2}_A_X{3}__XYA__X{5}`). Alternatively, it fits any SL with any two nucleotides followed by a pyrimidine and an "A" in the apical loop then only two basepairs before the "A" bulge (`X{2}_A_X{2}__XXYA__X{5}`). This simplified string is then converted into a proper regular expression so that "X" becomes "[GACUT]" and "Y" becomes "(C|U|T)" resulting in the following

TABLE 2.1: Conversions of ambiguous bases to regular expression.

Ambiguous base	Regular expression
X	[GACUT]
R	(G|A)
M	(A|C)
W	(A|U|T)
S	(C|G)
Y	(C|U|T)
K	(G|U|T)
V	(A|C|G)
H	(A|C|U|T)
D	(A|G|U|T)
B	(C|G|U|T)
T	(U|T)
U	(U|T)

search string:

`[GACUT]{2}([GACUT]_A_[GACUT]{2}__[GACUT]|_A_[GACUT]{3}__)[GACUT]`

`(C|U|T)A__[GACUT]{5}`.

A full list of ambiguous base conversions can be found in Table 2.1.

Finally, to decrease the complexity for the subsequent selection algorithm similar structures were grouped together. The most stable fold from the ensemble was considered the group representative and utilised for selection. Each group consists of structures that share the SL start and end position as well as the exact apical loop. This step is performed for all SLs fitting a given motif so that it does not remove structures with important parts of the motif outside of the apical loop.

2.4.1.4 Weighted-Activity Selection Algorithm

A given SL on an RNA spans the positions i to j where $i < j$. It can form simultaneously with another SL spanning i' to j', where $i' < j'$, if these two ranges, i–j and i'–j', do not overlap. This means $j < i'$ or $j' > i$. Computationally the

problem is similar to the activity selection problem, which deals with selecting a (maximal) set of activities, whose times do not overlap. However, not all SLs are equal, but instead have their own folding energies in kcal/mol: the lower the energy, the more stable the structure. The goal was to find a combination of local folds that minimises the overall energy, in order to approximate a MFE structure that was expanded as more RNA became released from ribosomes. This requires the use of weights, in this case energies, for the selection. This was computationally solved in the weighted activity selection (WAS) algorithm. It can be implemented either greedily or with dynamic programming (Kleinberg and Tardos, 2006, Chapters 4.1 and 6.1).

An algorithm is defined as "greedy" when at each point the currently best option is picked. The next step may depend on the previous one, but never on the next or the overall solution. There is no backtracking or adjusting previous choices. It therefore rarely produces a globally optimal solution, but potentially a sufficient approximation (discussed in detail in Curtis (2003)). An example of a greedy algorithm is solving the problem of how many tasks can be completed in a certain time frame given a list of how long each task takes. The greedy solution entails sorting the list of tasks in ascending order by length and then simply added them until the time limit is reached. At each iteration the shortest task is picked.

Dynamic programming was first described by Bellman (1954) and has since been applied to many programming problems (see books on dynamic programming such as Bellman (2003) or Bertsekas (2017)). It is an approach, in which a problem is divided into simpler, overlapping sub-problems and then solved from the solutions of these. It is often an alternative to recursive approaches by either storing previously calculated solutions, or solving the problem in its natural order. It is the basis of many bioinformatics algorithms. One example is pairwise alignment of sequences: the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The two protein sequences are placed against each other in a matrix and

for each combination of amino acids a score is calculated. The alignment is then completed by tracing back through the matrix for the optimal score. Solving this problem non-dynamically would entail to recursively match each amino acids pair and from there determining the following scores, which would mean calculating the same sub-scores several times.

As explained above a greedy approach has the disadvantage of only working in one way and not adjusting the solution for a better one. Dynamic programming on the other hand utilises an overlapping subset of problems to find the optimal solution. It is therefore more likely to produce the best outcome and was consequently picked for implementation of the WAS algorithm. The implementation is based on the pseudocode provided in the Dynamic Programming lecture slides by Wayne (2001) and Kleinberg and Tardos (2006, Chapter 6.1).

The algorithm requires a sorted list of compatible, i.e. non-overlapping, SLs: CompatibleSLs. This is achieved by first sorting the SLs in ascending order by their end positions in the genomic sequence and then finding the first next SL, which starts thereafter. Sorting was done using the quicksort algorithm (Hoare, 1961). Afterwards, a table of energies is calculated in WAS_COMP and then solved in SOL. A SL that was selected through the algorithm was set to 1 in the final table. The pseudocode for both recursive algorithms is shown in Algorithm A.4 in Appendix A .

2.4.1.5 Stem-loop Affinities

To improve accuracy of SL selection with WAS, an iterative process including approximate binding energies representing the affinity of PSs for CP was used. The binding energies were based on experimental and theoretical data available for bacteriophage MS2 (Lago et al., 2001; Dykeman et al., 2013b). When there is no affinity between a SL and CP, there are still electrostatic interactions at play, which are weaker. Kivenson and Hagan (2010) have modelled affinity of an unspecific polymer to capsid protein to be around $5.75 k_{\mathrm{B}} T$, which is approximately

3.5 kcal/mol at body temperature. So for non-PS SLs ΔG was considered as -3.5 kcal/mol. To convert dissociation constant to binding energy the following calculation was used:

$$\Delta G = RT \ln\left(\frac{K_{\mathrm{D}}}{c^{\ominus}}\right) , \tag{2.7}$$

where the gas constant R is $1.98588 \times 10^{-3} \frac{\text{kcal}}{\text{mol} \times \text{K}}$, T is the temperature in Kelvin (here the normal body temperature 310.15 K was used), and $c^{\ominus}$ is the standard reference concentration of 1 mol/L, thus:

$$\Delta G = 1.98588 \times 10^{-3} \frac{\text{kcal}}{\text{mol} \times \text{K}} \times 310.15\text{K} \times \ln\left(\frac{K_{\mathrm{D}}}{1 \text{ mol/L}}\right) . \tag{2.8}$$

To exemplify, for MS2 packaging signal TR with a K_{D} of 1.5nM the calculation would be:

$$\begin{aligned}
\Delta G &= 1.98588 \times 10^{-3} \frac{\text{kcal}}{\text{mol} \times \text{K}} \times 310.15\text{K} \times \ln\left(\frac{1.5 \times 10^{-9} \text{ mol/L}}{1 \text{ mol/L}}\right) \\
\Delta G &= 1.98588 \times 310.15 \times \ln\left(1.5 \times 10^{-9}\right) \times 10^{-3} \text{ kcal/mol} \\
\Delta G &= 1.98588 \times 310.15 \times -20.3178 \times 10^{-3} \text{ kcal/mol} \\
\Delta G &= -12.51415 \text{ kcal/mol}
\end{aligned} \tag{2.9}$$

In order to minimise biasing the selection for small, unstable SLs with a PS motif the selection was performed in several steps: First the selection was run on SL stability only. The folding kinetics of all SLs were tested to filter out SLs that are too kinetically unstable and would not realistically be present long enough to be bound by CP. Any chemical reaction needs to overcome a kinetic barrier called the activation energy to reach a transition state. The size of this barrier is a better indication of the short-term stability of a SL than of its thermodynamic formation energy. An in-house Fortran 90 program by Eric

Dykeman called rnarates calculates the highest energy a fold has to overcome. The energy barrier minus the formation energy gives the activation energy for the reverse reaction, i.e. unfolding of the SL. If this activation energy is too low, the structure is not kinetically stable. Therefore, structures with an activation energy of less than 2.5 kcal/mol were excluded by setting their stability score to -100. Otherwise, stability was simply the negative of the energy of the respective SL so low energy results in high stability and vice versa. Next, the selected SLs were checked for exhibiting a (high affinity) binding motif. If so, they were locked in by setting their stability scores to 10,000. This was to mimic early stages of packaging when CP concentrations are low: SLs form based on their stability and the few CPs present bind to and stabilise high affinity PSs, while the rest can re-fold. Finally, the WAS step was repeated with CP binding energies added, resulting in the final selection of SLs.

2.4.1.6 Generation of Packaging Signal Profiles

Once a selection of PSs was achieved, aligned genomic sequences were translated into PS profiles. These are pseudo-sequences containing one letter for a nucleotide that is not part of a PS and three different nucleotide letters for three affinity tiers, i.e. high, medium, and low. If affinities are not known, only one letter is used for all PSs. Non-PS nucleotide positions were given the letter "A" whereas high affinity PS positions were given "C", medium "G", and low "U". If affinities were not defined, only "C" was used (Figure 2.9A and B). Note that the choice of letters is arbitrary and does not carry a deeper meaning. The total number of selected PSs of each affinity tier in the profile was printed in the end. The pseudocode for generation of PS profiles from selected SLs and their numerical information is shown in Algorithm A.5 in Appendix A.

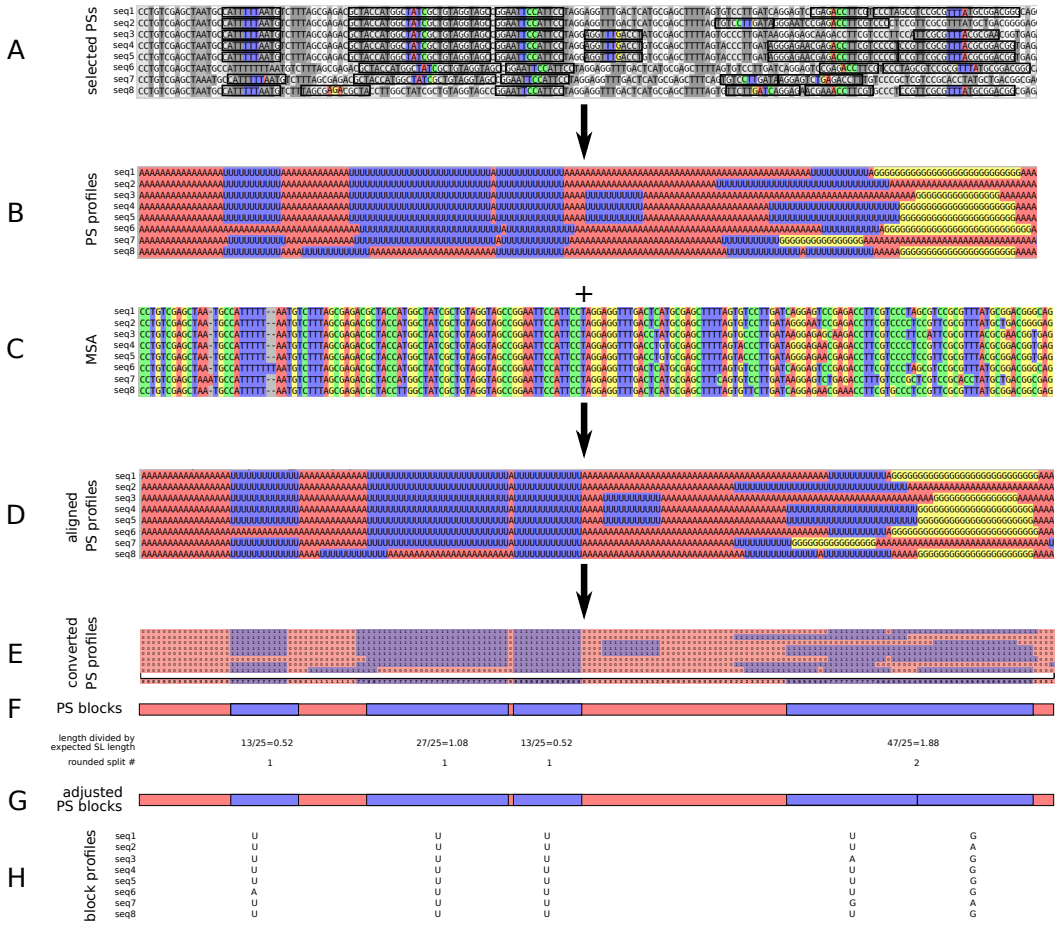

FIGURE 2.9: **PS phylogeny method example for part of MS2**. (A) SLs with a PS consensus motif (black boxes with motif in colour) are found in the sequence. (B) Every genomic nucleotide position is converted into pseudosequence with "A" (red) for no PS, "U" for low affinity PS (blue), "G" for medium (yellow), and "C" for high (not shown). The result is a set of PS profiles. (C) A MSA of the genomic sequences is used to shift the PS profiles resulting in (D) aligned PS profiles. Insertions in the MSA are translated to insertions of the respective pseudo-nucleotide at that position. In this example at first extra "A"s are inserted in sequences 1–6, and 8. Then "U"s are inserted in sequences 1–5, 7, and 8 because the insertion occurred within a PS region. (E) The aligned profiles are converted to numeric form where each PS position, regardless of affinity, becomes a "1" (purple) and all others "0" (red). Each column, corresponding to one aligned nucleotide position, can then be summed resulting in the number of strains with a PS in that position. Regions with a sum of 4 or higher, corresponding to a threshold of 50%, are marked in purple. (F) The preliminary PS blocks (purple) are the regions, where the number of strains with a PS is higher or equal to the threshold, in this example 50%. The actual blocks are adjusted using an expected SL length, here 25 nucleotides. The length of each block is divided by the expected SL length and rounded to the nearest integer. (G) Due to its length the last block was split into two. As seen in the original PS profiles this block does indeed correspond to two PSs that are flush to each other in the genome. (H) The membership for each strain and block is determined by assigning the affinity marker that occurs most in it.

2.4.1.7 Conserved Packaging Signal Blocks

In order to reconstruct a phylogenetic tree from PS profiles, they had to be converted to characters, which are the PSs. Later on, informative characters will be identified as a subset of PSs that are discriminatory between different sequences. Using the profiles as is would have introduced too much artificial variation. The exact position or structure of each PS is not important for this purpose. Rather, the comparison should be between the presence/absence of a PS in a certain region. This could be fine-tuned by comparing changes in the patterns of affinities. Therefore, the profiles were simplified using blocks of conserved PSs generated from aligned PS profiles by going through these nucleotide by nucleotide. A block is hereby defined as a region in which at least a certain number of strains have a PS regardless of affinity. If PSs are in slightly different positions with respect to the primary/secondary structure, they can still be in similar positions within the tertiary structure and fulfil the same function. Therefore, a concept is needed that tolerates variations in PS positions in strain variants to minimise noise, which needs the correct level of coarse graining. This threshold for creating a block is user defined and needs to be chosen with care depending on the sample set. For example, if a set of ten related sequences was to be compared to a set of 90 reference sequences, a threshold higher than 10% would make it difficult for these ten strains of interest to be clustered together and split from the references. This is because PSs that are only present in these compared to the reference strains would not create a block due to being fewer than the threshold. There is a trade-off between noise and resolution, which needs to be considered, when a threshold is picked. To ensure that corresponding regions were compared, complete RNA sequences were first aligned using ClustalΩ with the output format set to FASTA and the order to input order (Goujon et al., 2010; Sievers et al., 2014). This ensured that the strains were in the same order in all files used for generating the blocks.

First, the PS profiles were adjusted using the MSA. If a deletion, marked

in the alignment with "-", was present at a nucleotide position, then the same letter as in the previous position was inserted (Figure 2.9C and D). If the deletion occurred in the beginning of the sequence, an "A" for "not a PS" was used. To facilitate calculating the number of strains with a PS at any position the sequences were also converted into 0s and 1s at the same time. 0s were used for non-PS nucleotides and 1s for PSs (Figure 2.9E). This allowed summing over columns in the next step. The pseudocode for this step is shown in Algorithm A.6 in Appendix A.

Next, PS blocks were defined by summing over columns at each nucleotide position. If this sum was larger or equal to the threshold converted from percentage to number of sequences in the ensemble, then this position became part of a block. Either a new block was started or the current one extended. A block ends when the condition is no longer met (Figure 2.9E and F). The use of PS blocks instead of the PS profiles aids in abstracting the positional information to provide a more suitable framework for comparison. Whilst a full overlap between PSs in different strains was not considered necessary here, a certain degree of overlap in the aligned sequences was to ensure they would be close enough to fulfil equivalent roles. It is possible that two or more PSs are so close to each other that there is an overlap. Alternatively, they can be flush in the genomic sequence of each strain. This would create one large block instead of several shorter ones (Figure 2.10A). Conversely, a short stretch above the threshold could be the result of two sets of PSs overlapping slightly (Figure 2.10B). To correct for these instances the user could also specify an expected SL length, usually 25 nucleotides. The length of the block was divided by the expected SL length and rounded to the nearest integer (Figure 2.9F). If this was 0, the block was deemed too short and deleted. If it was 1, then one block was generated. If it was more than 1, then it was divided into several smaller blocks (Figure 2.9G). The number of blocks was saved in BLOCKN and their start and end positions in array BLOCKS_N. The pseudocode in is shown in Algorithm A.7 in Appendix A.

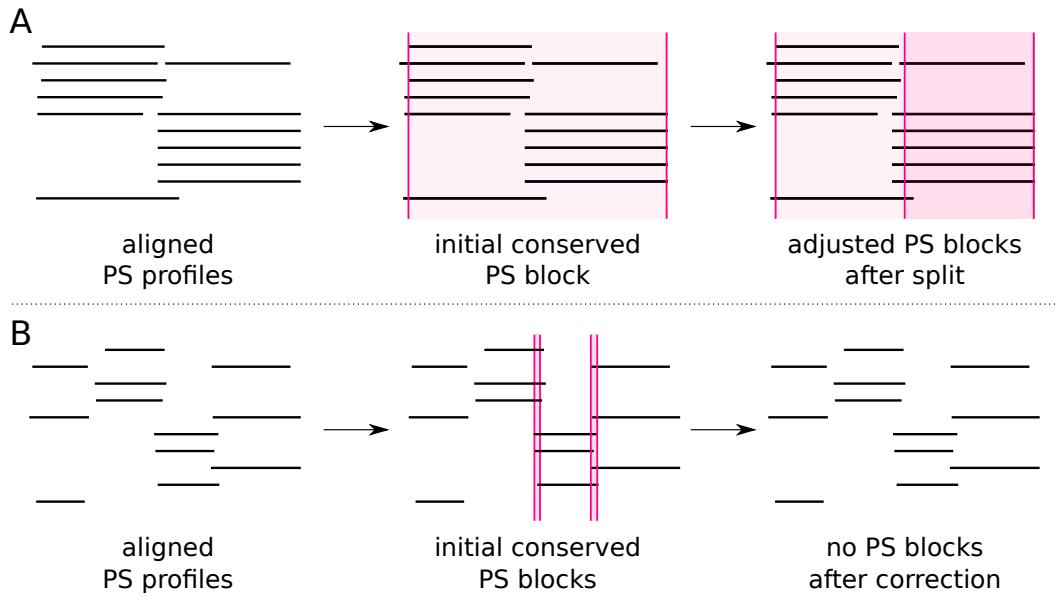

FIGURE 2.10: **PS block adjustment**. Aligned PS profiles (black lines) are the basis of conserved PS blocks (pink). A block is started when the conservation threshold is reached and continues until that condition is no longer met. In cases as shown in (A) this can result in the generation of one large block rather than two smaller ones. Without block split all these strains would be considered the same in this region despite their clear differences in the profiles. Splitting the blocks alleviates this problem. When there is only a small overlap as in (B) small PS blocks are generated despite the lack of real conservation in the area. Here, a set of three PSs overlaps with two sets of three on either side resulting in two small blocks. The adjustment using SL length deletes such too short blocks.

2.4.1.8 Assigning PS Block Membership and Conversion to Characters

After block generation, each PS in each sequence was assigned to a block if possible. Some PSs would not be in any block due to occurring in a region with too few PSs of other strains. These might occur by chance and have no functional role but are essentially noise, which is why they were excluded with the conservation threshold. Assigning block membership starts by checking if the sum in the numeric representation of the PS profiles in that strain is larger than zero between block start and end. Since non-PS positions are encoded as "0"s and PS positions as "1"s, the sum over a region would only be zero, if there was not a single nucleotide that was part of a PS in that region. The entire process,

however, is not trivial. While in essence for each block it is tested whether a PS is present in the respective sequence, some special cases have to be accounted for. Care must be given to only count each PS once. That means that if one structure spans two blocks, it can only be assigned to one (Figure 2.11). Conversely, two or more PSs may fold flush to each other appearing as one large structure in the profile. If these span several blocks, they should be assigned to as many as there are PSs. Since the exact information is not available at this step, it was approximated by using the predicted SL lengths. The subroutine AFF_ASIGN assigns the PS affinity marker to the respective block. The decision is simply made by which one occurs most often within the block for that sequence. First it is tested, whether the PS spans the entire space between the blocks. If not, then it is simply assigned to the one block (Figure 2.11H). If it does span the distance, it is tested, whether there is another PS in the next block. If there is, then the first PS is counted towards the first block and the second towards the next Figure 2.11A and B). If there is not, then the PS is assigned to the block with which is has the highest percentage overlap, e.g. in Figure 2.11C, E, and F it would belong to the left block, and D and G to the right block. To assess how useful the PS blocks are for reconstructing a phylogenetic tree, the number of informative characters, i.e. PS blocks in which membership differs between the strains, is also calculated and printed at the end. Functionally important PSs should be in most if not all strains. The number of informative characters also imply the number of PSs that are fully conserved among all strains. The pseudocode for the main program and the suboutine can be found in Algorithm A.8 and Algorithm A.9 in Appendix A.

2.4.1.9 Construction of Phylogenetic Trees

In the previous step each sequence was assigned a letter in each block depending on if it had a PS in that region or not. This assumes that the blocks have previously been divided accurately to represent the correct number of PSs that

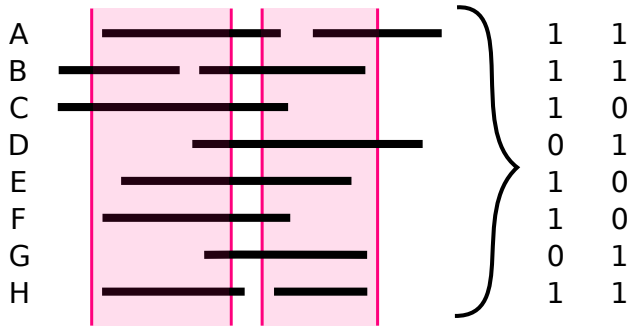

FIGURE 2.11: **PS block membership assignment**. There are different ways in which PSs can overlap between blocks. It is important to assign each PS to the correct block especially when they have different affinities. There can be two PSs for two blocks (A, B, and H) or there can be one PS that overlaps with two blocks in different ways (C–G). The assigned membership is shown on the right as 0s (no PS in that block) and 1s (PS in that block).

underlie them, because each block can only be assigned one PS from each strain. The resulting PS block profiles are strings of letters A, G, C, U, which could be used as input for tree building. Due to resembling a nucleotide sequence, standard tools could be utilised at this stage. The SplitsTree 4 program was used to reconstruct phylogenetic trees (Huson and Bryant, 2006).

2.4.2 Stem-loop Selection in MS2

To assess the performance of the SL selection algorithm, it was tested on an MS2 RNA. MS2 is a suitable model for this test because its genomic structure has been studied extensively and is well described. It was therefore possible to compare the selection made by the algorithm against the global structure published in Olsthoorn (1996); Groeneveld (1997) and Dai et al. (2017). Moreover, its PSs including affinities are known (Dykeman et al., 2013b; Dai et al., 2017).

One MS2 complete genome (EF108464.1) was randomly selected and utilised for this test. The RNA was split into 90, 60 or 30 nucleotide overlapping fragments by sliding a window in increments of 1 nucleotide. Different window sizes were tested to ensure robustness of the algorithm and check for window bias. Therefore, each window size was tested separately. Each fragment was folded and sampled

10,000 times using Tfold in partition function mode. Folding of the fragments was parallelised using GNU parallel (Tange, 2011). Next, single SLs were extracted from the folds in each fragment. The number of times a particular SL was sampled was recorded and was utilised as a proxy for stability. The extracted SLs were then merged across windows, i.e. if the exact same SL occurred in more than one window it was only kept once in the ensemble, and number of occurrences among the 10,000 samples was added together.

PSs do not only compete with each other for folding but also with other SLs. Although they needed to be considered in the selection algorithm, some additional weight needed to be given to PSs because binding to CP makes favourable energetic contributions. Therefore, exact energies of each structure were calculated and used as weights instead of occurrences in the sampling. Moreover, the kinetics of SL folding were taken into account to remove SLs that were not kinetically stable. In addition to the folding energy also the energy hurdle to unfold was calculated for each SL. If this hurdle was less than 2.5 kcal/mol, the structure was deemed unstable and assigned a very low stability to ensure it was not selected in later steps. This mostly removed small two-base-pair structures. Given PSs with variable affinities and a CP concentration that increases over time (a protein ramp) high-affinity PSs are likely to be bound first, when CP concentrations are low, while lower-affinity PSs are bound later. To mimic that, the WAS algorithm was first run without affinities added. This represents the local structures present when CP first starts binding. Each high-affinity PS selected at this stage was locked in by assigning it a very high stability. Then affinities were added to the folding energies and the WAS algorithm was run again.

To run the algorithm for MS2, a suitable set of PS motifs had to be provided. These were based on results from Dykeman et al. (2013b) and the PSs published in Dai et al. (2017). Three affinity tiers were used: For high affinity the SL had to have either a 4 nucleotides apical loop with two base-pairs and an A on the 5' side or a 3 nucleotide apical loop with three base-pairs. The last two positions

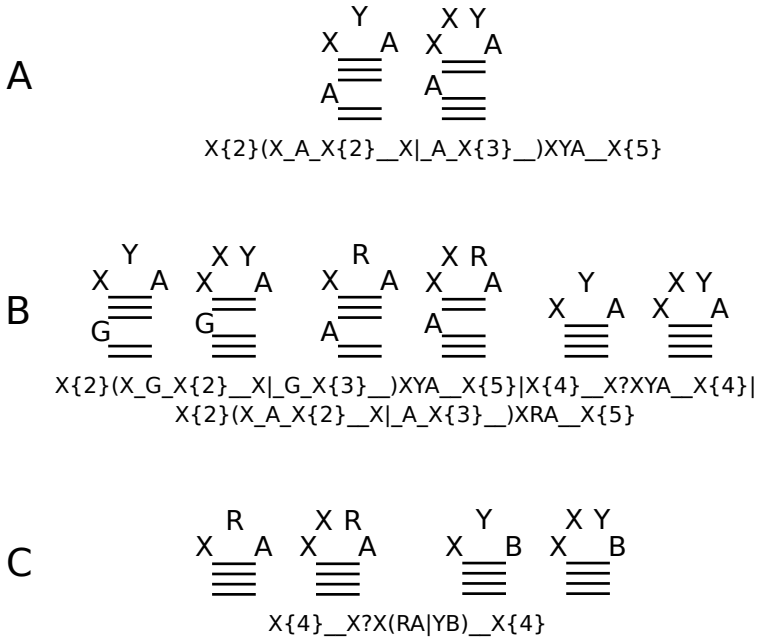

FIGURE 2.12: **MS2 packaging signal search motifs.** The motifs for the tiers of used for the SL selection are shown as structure and corresponding search phrases, shown for high (A), medium (B), and low (C) affinity tier PSs.

in the apical loop had to be Y (C or U) followed by A. No bulge was allowed on the 3' side (Figure 2.12A). This tier was given a K_{D} of 1.5 nM. The second tier included more options. Relative to the top tier, either the Y was changed to an R, or the bulge was lost, or instead of A there was a G in the bulge (Figure 2.12B). This tier was assigned a K_{D} of 150 nM. The low affinity tier, finally, had either RA or YB (G, C or U) at the end of the apical loop and no bulge (Figure 2.12C). The K_{D} was set to 1500 nM.

Going through the two-step process described above with these search motifs identified most of the published MS2 SLs, both PSs and others. Out of 74 SLs, 62 (84%) were correctly selected, 5 (7%) were not selected, and 7 (9%) were slightly modified usually through additional or less base-pairing in the apical loop. These numbers were the same for all window sizes tested indicating the robustness of the approach. For comparison the complete genomic RNA sequence was folded in Mfold on default settings (Zuker, 2003). Mfold was only able to form 49 (66%) of SLs correctly, while 18 (24%) were not present in the Mfold global structure, and

7 (9%) were modified. Which SLs were successfully predicted by each algorithm is shown in Appendix A Table A.1. This showed that using the serial local folding algorithm described above had an advantage over a simple global fold in correctly picking MS2 SLs.

Since the main purpose of this algorithm is to identify PSs for constructing phylogenies, the MS2 results were also compared with the 15 PSs published by Dai et al. (2017). Regardless of window size, the algorithm selected all 15 SLs correctly. Note that one of these represents a modified apical loop compared to the global structure. In comparison, Mfold only folded 10 (67%) of the PSs correctly, while 4 (26%) were completely absent and 1 (7%) was modified so that it would not count as PS any more. The stem-loop selection algorithm presented here can thus be considered suitable for generating global structures and selecting PSs.

2.5 Discussion

In a series of steps I have developed a set of algorithms that prepare sequences for phylogenetic analysis. Together, they take a set of RNA sequences, fragment them, fold the fragments, merge folds across fragments, and utilise SL formation and PS-CP binding energies to select an energetically optimal non-overlapping set of SLs. In the final step the genomic sequences are converted to PS profiles, i.e. pseudosequences, where each pseudo-nucleotide represents whether or not this position is involved in a PS. With these as input the phylogeny algorithm identifies regions in the genomes, where at least a certain percentage of sequences defined by the user have a packaging signal. Such PSs are more likely to be functionally important. The threshold is not hardcoded but variable, because it depends on the set of sequences the method is applied to; e.g. more closely related strains will require a greater level of coarse-graining than more distantly related ones. The threshold, therefore, needs to be adapted to the respective dataset to allow for a meaningful comparison. Once these regions, PS blocks, are defined,

every sequence is individually assessed for the blocks in which it presents a PS. The final output of the algorithm is a set of characters, where each represents one block containing one PS. A different nucleotide letter is used for absence of PSs and PSs of different affinity tiers. Since the characters look like nucleotides, the output can be processed using standard tools to generate phylogenetic trees.

The SL selection part of the method was tested on a bacteriophage MS2 genome. It was capable of correctly predicting over 80% of published SLs. For comparison, the sequence was also folded in Mfold on default parameters. This only yielded 65% of the SLs. When looking at PSs specifically, the SL selection method was capable of selecting all 15 PSs published by Dai et al. (2017), whereas the Mfold structure only contained 10 of these. It can therefore be said that this part of the method is not only adequate for folding viral genomes, but even improves upon standard tools. It is especially useful for identifying PSs, which is essential for the second part of the method, creation of characters for phylogeny.

Accurate RNA secondary structure prediction stands at the foundation of PS-based phylogeny. Over the years many different methods for RNA secondary structure prediction have been developed. Still today the most popular are based on dynamic programming approaches that aim to find the MFE structure (Mathews, 2006; Pal et al., 2017). Some examples of these are Mfold (Zuker, 2003), RNAstruct (Mathews et al., 2004; Reuter and Mathews, 2010) or the Vienna package (Lorenz et al., 2011). Whilst these programs identify the structure that is thermodynamically the most stable they fail to take a number of other factors into account such as kinetics, co-transcriptional folding or the limits in accurately determining the energy parameters that these methods are built on (Pal et al., 2017; Zuber et al., 2017), or, as is the case here, the effects of RNA-protein interactions. Moreover, Morgan and Higgs (1996) have shown that kinetics play an important role in the folding of RNA molecules. They found that whilst small domains are very stable, larger ones have a higher energy than the average of the MFE, which supports the idea that the folding process occurs in stages beginning

with small local folds and expanding to few longer range interactions (Morgan and Higgs, 1996). The folding kinetics could trap the RNA in a suboptimal energy state. A different analysis also showed a correlation between structure energy and the percentage of correctly identified base-pairs which disappears for longer sequences (Wiese et al., 2008). This may explain why secondary structure predictions decrease in specificity and sensitivity with increasing sequence length. Folding in stages and starting from smaller local folds was mimicked here by determining structures in small windows and then combining them to a global structure. Whilst this approach does cannot account for longer range interactions, they are also not relevant for the purpose here, which is to identify PS SLs.

Some of the problems of dynamic programs have been addressed in the past through continuous improvement of the energy parameters (Mathews et al., 1999; Zuber et al., 2017) and the inclusion of suboptimal structures as the MFE structures do not necessarily represent what is present *in vivo* (Zuker, 1989; Wuchty et al., 1999). A different way of obtaining an ensemble of structures is to sample the partition function (Ding and Lawrence, 2003). Sfold first applied this idea of using the Boltzman distribution to predict structures from a centroid, i.e. the structure that best represents set of structures (Ding et al., 2005). It produced better results than simply calculating the MFE structure and partition function mode has been added as a feature to many standard dynamic programs mentioned above. Also my algorithm relies of sampling the partition function to obtain an initial ensemble of structures in each window. Sampling 10,000 times ensured a larger set of structures, which improves the chances of finding an optimal set at the end of the algorithm.

To address other shortcoming alternative approaches to MFE structure calculation have been developed in the past. RnaPredict is an evolutionary algorithm, which performed similarly to Mfold for shorter sequences but both decreased in specificity and sensitivity for sequences longer than 450 nucleotides (Wiese et al.,

2008). It would, therefore, not have provided a more suitable comparison for my algorithm when folding a viral genome of over 3000 nucleotides.

Further improvement on longer sequences, especially over 1000 nucleotides, were achieved with CoFold (Proctor and Meyer, 2013). Its algorithm limits the reach of interactions in the underlying thermodynamic model to mimic co-transcriptional folding. In this way it can combine kinetics and thermodynamics but still fails to take trans-interactions into account such as protein binding to the RNA. It is the most similar approach to mine, which by combining smaller, local folds mimics co-translational folding as folding occurs behind the last ribosome. However, my algorithm also takes the impact of the trans interaction with CPs into account by adding the affinities to the SL energies. Being the best-performing program for longer sequences, it would be interesting to compare the accuracy of CoFold to my algorithm in the future.

The idea that the RNA may fold in stages has been taken up in fledFold (Liu et al., 2016). This greedy algorithm outperforms Mfold, RNAfold, Sfold, and even CoFold in sensitivity and accuracy on shorter sequences up to 400 nucleotides but all methods' sensitives decline with sequence length and on some tested sequences fledFold is still outperformed by Mfold (Liu et al., 2016) illustrating that this method is nevertheless not obsolete. In principle my serial WAS approach is similar to fledFold. Also here folding occurs in stages and is built up of smaller local structures. The main difference is that my algorithm is not greedy but still attempts to optimise the overall energy of the global structure.

Recently, a novel quantum genetic algorithm has been proposed and compared with another quantum genetic algorithm, a standard genetic algorithm, and three commonly used dynamic programs (RNAfold, RNAstructure, and Mfold) (Shi et al., 2019). It thus provides the most current performance comparison for RNA secondary structure prediction approaches. As shown before, RNAstructure and Mfold decreased markedly in accuracy with increased sequence length (between 90 and 400 nts) while RNAfold started out less accurate and became better on

longer structures. The new method performed best and lost accuracy the slowest but also became not useful beyond 500 nucleotides.

All these illustrate that the search for the best RNA secondary structure prediction method is not over, yet. Many approaches exist with their own strengths and weaknesses. Despite Mfold being a simple dynamic program, it is still relevant and commonly used today and was therefore used in comparison to my own algorithm. Comparisons with other programs, especially CoFold or fledFold, would be interesting in the future to further benchmark the performance of my folding approach. To date, I am not aware of another folding program that takes RNA-protein interactions into account, which makes the algorithm described here unique.

Utilising RNA secondary structures for phylogenetic analysis of sequences is and of itself not a new idea. In the 1990s substitution models were proposed that take the different mutation rates between helices and loops into account (Schöniger and Von Haeseler, 1994; Tillier and Collins, 1998). If all nucleotides are considered independent, the distance between sequences can be underestimated because the nucleotides in the helix portion of a SL are interdependent (Schöniger and Von Haeseler, 1994). Using different rates for base-paired or single-stranded portions of the sequence improves the phylogenetic prediction (Tillier and Collins, 1998). These approaches assume less evolutionary pressure in loop regions, where the standard DNA model is used. Whilst these studies illustrate the power of an accurate substitution model, they are not directly applicable for PS-based phylogeny. In the case of PSs it is in fact the loop regions that confer the function in most cases, as they interact with and bind to CPs. It is however, not as simple as to assume slower rates in the loop portions as only PS SL are affected and within these only some parts. Thus, a substitution model would need to assume slower evolution in nucleotides involved in CP-binding, which may be specific sequences in loops or helix lengths. Consequently, designing such a model is not a trivial task and it has not been attempted here, despite its undeniable

usefulness.

Software such as the PHASE package are specifically designed for phylogenetic analysis of RNA sequences with a conserved secondary structure (Jow et al., 2002). Since then it has become popular to use known RNA secondary structures for phylogeny. This is usually done by adjusting the DNA MSA through aligned secondary structure and then considering helix and loop portions separately (Young and Coleman, 2004; Telford et al., 2005; Biffin et al., 2007; Grajales et al., 2007; Keller et al., 2010). They often used programs that can align sequence and structure simultaneously. One popular one is 4SALE, which combines the two types of information by converting it into pseudo-amino acid sequence (Seibel et al., 2006). Since each of the four nucleotides can be present in three different states, i.e. unpaired, opening base-pair and closing base-pair, there would be twelve letters needed; e.g. A., A(, A) would be assigned three different letters. A similar approach is used in the ProfDistS program, which evaluates such alignments for reconstruction of phylogenetic trees (Wolf et al., 2008). Rather than just using adjusted sequence alignments, structures can thus still be incorporated into the phylogeny. These programs are useful for the studies above but less so for what I tried to compare in my phylogenies. The interest here was not in finding corresponding nucleotide regions utilising SLs. Programs such as 4SALE, ProfDistS or PHASE would be useful if the goal was a phylogeny of *Hepadnaviridae* based on ε or another well-defined and conserved structure but not when SLs are expected to vary between strains. The PSs are assumed to be somewhat variable not only in sequence but also in exact location. Nevertheless, there are parallels in my algorithm and 4SALE and ProfDistS in that both approaches utilise pseudosequence. The main difference is that in ProfDistS, where this encoding was utilised for calculating phylogenies, they wanted to encode more information (Wolf et al., 2008). However, for my algorithm the sequence in the structures is not important for the phylogeny in the end so my goal was to abstract the information by converting all nucleotides involved in a PS into one

pseudonucleotide. Practically, the real parallel with the pseudosequence used in 4SALE and ProfDistS lies with my encoding of sequence/structure information. Using a 12 letter system as in these two programs would have solved the problem of the SL length and adjusting block sizes or could have been used for motif search. However, my own encoding of sequence/structure is more human-readable and provides a simpler translation from search term to regular expression. On the other hand, due to need to count alternating internal loops/bulges and helix sequences from the apical loop, there is a higher risk of error when formatting my encoding.

Phylogenetic trees can be reconstructed based on any (set of) character(s). Nowadays these are often molecular features, in particular nucleotide or amino acid sequence. However, as Bamford *et al.* have demonstrated much insight can be gained by comparing the morphology of molecules (Bamford et al., 2005). Specifically, they uncovered previously unthought-of evolutionary ties between viruses that infect hosts of different domains of life, by looking at their capsid protein folds. This different approach to phylogeny based on structure and function of a molecular feature has motivated us to attempt to develop a method to reconstruct phylogenetic trees based on PS profiles. In the process I have developed an algorithm for secondary structure prediction that takes PS affinties for CP into account. Incorporating trans-interactions is unique for this method and it could theoretically also be applied to other RNA-protein interactions with known or predicted affinities. Going from an RNA secondary structure to finding certain sequence/structure motifs required an encoding that made both types of information accessible at the same time. Whilst this concept is not new, it is a novel way of encoding and allows for a human-readable way to visualise the structure and sequence of a SL as an alternative to Vienna bracket format underneath sequence. While the performance of the SL selection part has been assessed and confirmed against Mfold, further comparisons with other secondary structure prediction programs such as CoFold or fledFold would provide further information

on how well the structure prediction performs. However, determining the RNA secondary structure is only the first part of this new method. The main goal is to be able to reconstruct phylogenies based on the PSs of a virus. How well this method is suited for reconstructing phylogenetic trees will be determined in later chapters as further conclusions can only be drawn based on specific examples. In following chapters the method will be slightly adapted and applied to subtypes of HBV (see Chapter 4) and members of the *Leviviridae* viral family (see Chapter 6). At that point it will be possible to assess how this new approach compares to phylogenies based on nucleotide sequences. Of particular interest will be the pace at which PSs evolve compared to sequence, whether some PSs are more conserved than others, and if the individual sites evolve independently.

Chapter 3

Identification of a Packaging Signal Motif in HBV

The overall workflow of this chapter is summarised in Figure 3.1 below. In Chapter 1 Section 1.1.2, packaging signals (PSs) were introduced as stem-loops (SLs) in the genomic RNA of a virus that bind to the viral capsid protein (CP) to facilitate capsid assembly and genome packaging. They have mostly been discussed in the context of single-stranded RNA (ssRNA) viruses such as *Levivirus* MS2. In this chapter the identification of PSs in hepatitis B virus (HBV) will be described. While HBV is a double-stranded DNA (dsDNA) virus, it is the pre-genomic RNA (pgRNA) that gets encapsidated (see Chapter 1 Section 3.1.3). It was therefore assumed that, in addition to ε, there are other dispersed PSs in the pgRNA of HBV, that had not been discovered before.

In this chapter we present a general approach to identifying a PS motif. See Figure 3.1 for the workflow.

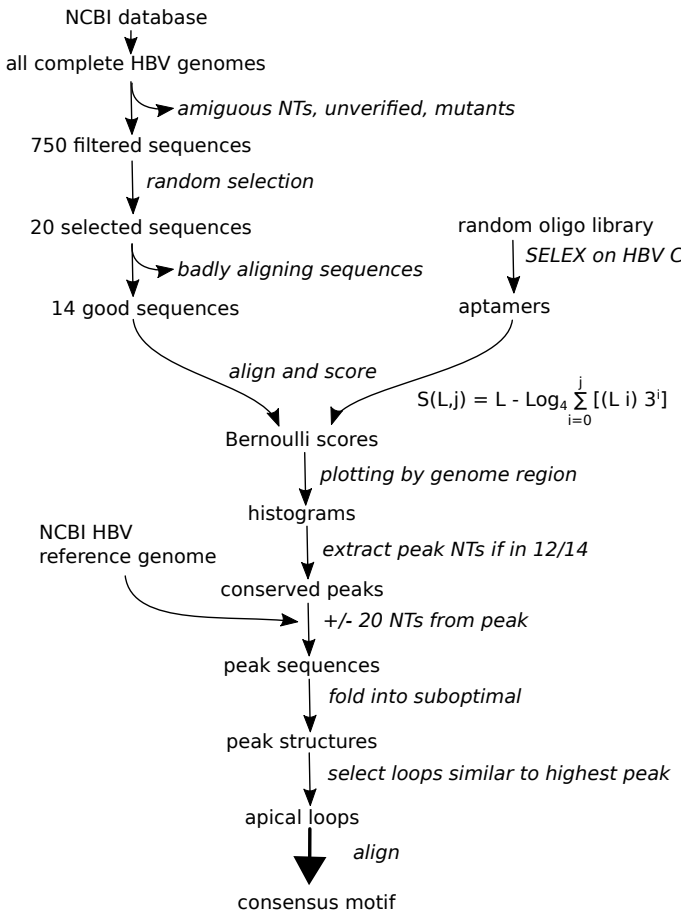

FIGURE 3.1: Workflow of packaging signal identification.

3.1 Hepatitis B Virus

3.1.1 Epidemiology

HBV infects and replicates in human liver cells, and if the infection is not cleared, it becomes chronic and can lead to liver cirrhosis or cancer in the long run. Worldwide there are 240 million chronically infected people and 780,000 die annually as a consequence (World Health Organization, 2017). The rates of chronic infection vary drastically between different regions of the world. They are the lowest in North America, Western Europe, and Australia whereas large parts of Africa, East Asia, and the Middle East have a high prevalence of chronic infection. In the majority of healthy adults acute infections are readily cleared by the im-

mune system. According to the World Health Organization (WHO) less than 5% become chronically infected. However, during infancy the chances to develop chronic infection are as high as 90% and in the first six years still 30–50% (World Health Organization, 2017). The best strategy of prevention is therefore vaccination immediately after birth or soon after. This is done routinely in the United States of America and in the United Kingdom for babies born to infected mothers (Centers for Disease Control and Prevention, 2018; NHS, 2015). Since 2017 the NHS provides the HBV vaccine to all other babies within a combined vaccine at 8 weeks (NHS, 2015). Before HBV vaccination was easily available, the vast majority of new infections were acquired perinatally or in early childhood (Centers for Disease Control and Prevention, 2008). This is still the case in endemic areas, while in Western Europe and North America the most common route of infection nowadays is through intravenous drug use and unprotected sex (Centers for Disease Control and Prevention, 2008; World Health Organization, 2017). However, in endemic areas timely access to health care is often difficult resulting in the observed high rates of chronic HBV infection. Chronic infection can be treated with a number of anti-viral drugs, which can control the infection but not cure it. This means that patients have to continue taking the medication for the rest of their lives in order to stay healthy (World Health Organization, 2017). This is particularly difficult in the above mentioned areas. Finding alternative drug targets that aim to cure the infection would improve the lives of millions of people worldwide.

3.1.2 Virology

HBV is a member of the *Hepadnaviridae* family, which also encompasses woodchuck hepatitis virus (WHV) and bat hepatitis virus (BtHV) amongst the *Orthohepadnavirus* genus and duck hepatitis B virus (DHBV) amongst the *Avihepadnavirus* genus. With a genomic length of only approximately 3200 nucleotides HBV is one of the smallest viruses known (Tiollais et al., 1985). It achieves such a

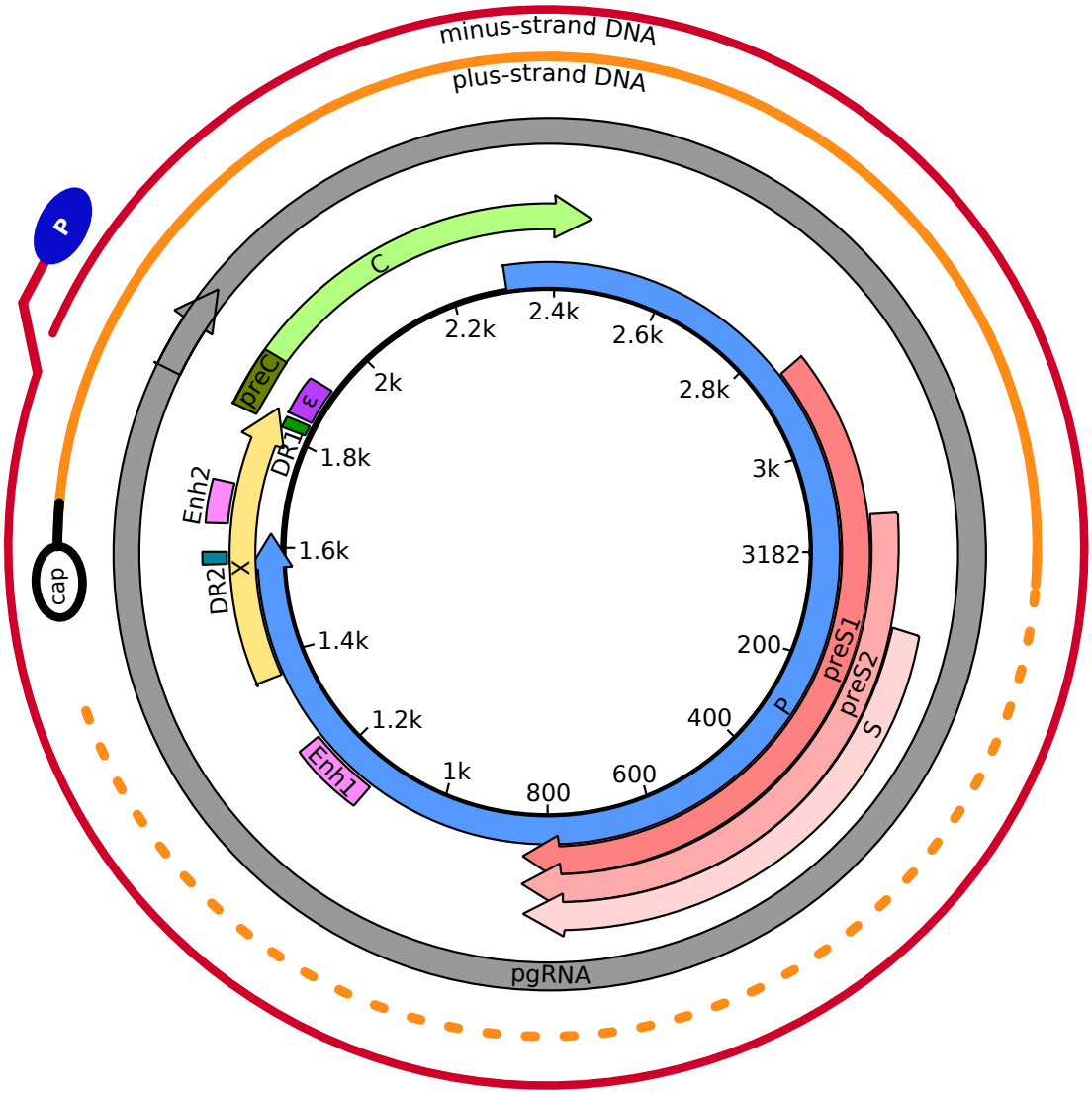

FIGURE 3.2: **HBV genome organisation.** HBV has a circular partially double stranded DNA genome. The minus-strand is complete (red) and DNA polymerase (Pol) (blue) is covalently attached at the 5' end. The plus-strand is incompletely synthesised and varies in length (orange). The pgRNA 5' cap is covalently attached at the 5' end (black). The terminally overlapping pgRNA is shown as a grey round arrow. Two enhancers (Enh1 and Enh2) are shown in pink, the direct repeats (DR1 and DR2) are displayed in dark green, and ε is purple. The overlapping genes are shown as round arrows.

small genome by having no non-coding regions and a high degree of gene overlap. The HBV genome consists of only four open reading frames (ORFs): preC/C for pre-core protein (HBeAg) and core protein (HBcAg), P for Pol, X for X protein (HBxAg), and preS1/preS2/S for long surface protein (LHBsAg), middle surface protein (MHBsAg), and small surface protein (SHBsAg), respectively (Tiollais et al., 1985). Additionally, it contains a number of *cis*-acting elements that act in the RNA: two direct repeats (DR1 and DR2), stem loop ε, ϕ, ω, and two enhancers (enh1 and enh2) (Figure 3.2). The mature viral particle consists of an outer envelop in which the surface proteins (LHBsAg, MHBsAg, and SHBsAg) are situated. The inner icosahedral T=4 capsid is made up of 240 HBcAg and contains the viral DNA and one copy of Pol (Figure 3.3) (Crowther et al., 1994; Zlotnick et al., 1996; Wynne et al., 1999). The virus is also capable of forming smaller T=3 capsids, which only contain 180 HBcAg units (Crowther et al., 1994), which are not commonly found in infection. The protein translated from the preC gene is cleaved to give rise to the small peptide HBeAg, which is secreted and whose main function is to induce T-cell tolerance against HBcAg (Milich et al., 1998). Despite being a dsDNA virus, HBV replicates via a ssRNA intermediate, the pgRNA (Summers and Mason, 1982). Within the viral capsid the pgRNA is reverse transcribed into relaxed circular DNA (rcDNA) by viral Pol. When the virus enters a host cell, its rcDNA is transported into the nucleus. There, it is repaired and becomes covalently closed circular DNA (cccDNA), which serves as template for RNA synthesis. Since viral RNA is synthesised by host RNA polymerase II, it has the same properties as eukaryotic messenger RNA (mRNA): a 5' cap and a poly-A tail (Rall et al., 1983). Five different mRNAs are produced, which all share the same poly-adenylation (poly-A) site (Tiollais et al., 1985). These encode for HBeAg, HBcAg/Pol, LHBsAg, MHBsAg/SHBsAg, and HBxAg, respectively.

The mRNA that encodes HBcAg and Pol also serves as pgRNA. Since the HBcAg ORF precedes Pol, HBcAg is synthesised in larger amounts. Once Pol

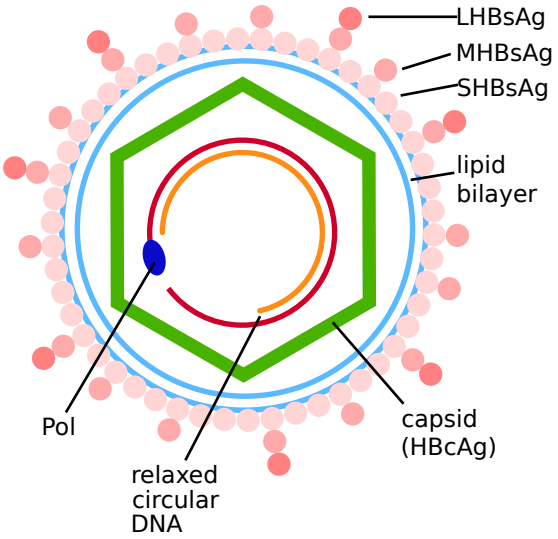

FIGURE 3.3: **HBV viral particle.** The HBV viral particle is an enveloped virus (light blue) with three types of surface proteins in the envelop (shades of pink): long (LHBsA), middle (MHBsA), and small (SHBsA). Within is the capsid (green), which is icosahedrally shaped and consists of 240 capsid proteins (HBcAg). It contains the viral DNA in its relaxed circular state (red and orange) and bound polymerase (blue).

has been made it binds to ε at the 5' end of the mRNA. This stabilises the SL and results in translational inhibition of both ORFs (Ryu et al., 2008). Further details about translational regulation of the pgRNA will be discussed in Chapter 5 "Nucleation of Assembly in HBV". The pgRNA is then encapsidated by HBcAg. Pol reverse transcribes the pgRNA after packaging within the viral capsid. Reverse transcription involves primer translocation to generate the double-stranded rcDNA present in mature viral particles. It is in this form that the virus infects a new host cell.

3.1.3 Genome Packaging and Viral Capsid

The *cis*-acting element ε at the 5' end of the pgRNA is for both packaging of pgRNA (Junker-Niepmann et al., 1990) and initiation of reverse transcription (Nassal and Rieger, 1996). It is thought to be necessary and sufficient for pgRNA packaging and the only PS in HBV (Junker-Niepmann et al., 1990). ε forms a stem-loop, which consists of two helix portions separated by a 6 nucleotide bulge,

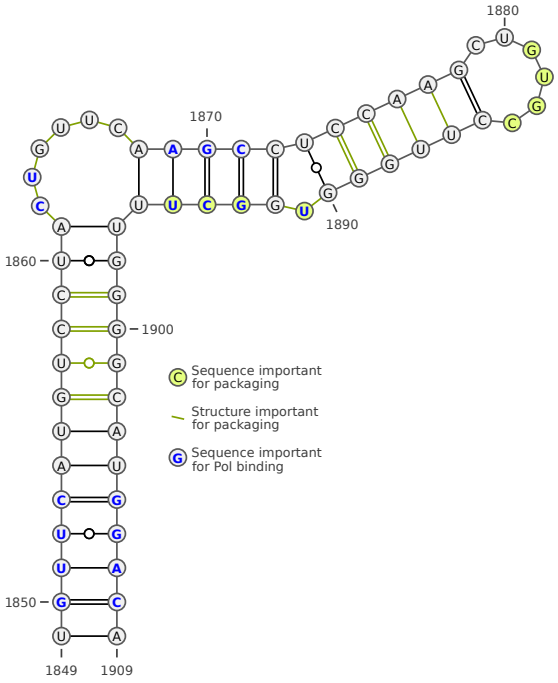

FIGURE 3.4: **Structure of ε.** The structure of ε for NCBI reference strain NC_003977.2 is depicted. ε consists of two long helices a 6 nucleotide 5' bulge and 6 nucleotide apical loop. The structural elements fulfil different roles. The sequence and structural parts important for genome packaging are marked in green and sequence important for Pol binding is highlighted in blue. The structure was generated in VARNA (Darty et al., 2009) and edited in Inkscape.

and a 6 nucleotide apical loop (Figure 3.4) (Pollack and Ganem, 1993). Different parts of its structure are important for its different functions: For pgRNA packaging the only specific sequence portions required are the apical loop and a small part of the upper helix, while the lower helix and the bulge have to be structurally present but the sequence is irrelevant (see Figure 3.4, green) (Pollack and Ganem, 1993). For binding to Pol the bulge portion, especially the first nucleotides, is most important but also the sequences of the lower and upper helices play a role (see Figure 3.4, blue) (Hu and Boyer, 2006).

Four arginine-rich repeat regions at the carboxy-terminus of HBcAg, the HBV CP, are necessary for nucleic acid binding but dispensable for assembly (Gallina et al., 1989). Only the first repeat is required for RNA encapsidation whereas the following three, which contain a known DNA-binding motif SPXX (Suzuki, 1989), bind DNA (Hatton et al., 1992). Nucleic acid binding is only possible in the folded

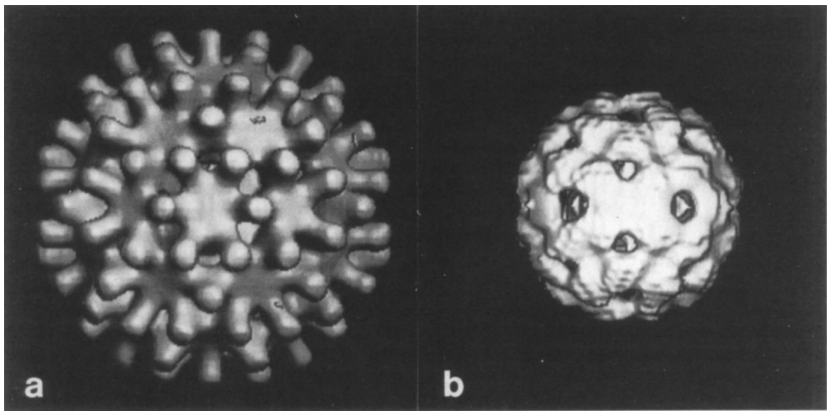

FIGURE 3.5: **Cryo-electron microscopy of HBV capsid.** (a) Reconstructed structure of HBV protein shell viewed from the 2-fold axis of symmetry. Peaks of density are situated around the 5-fold axesas pentamers and on the 2-fold axes as hexamers. (b) The density within the capsid is also icosahedrally ordered and probably represents packaged RNA. Figure 5 as published in Crowther et al. (1994). Copy right cleared with Elsevier through Copyright Clearance Center, license number 4476530588737.

state of the protein; denatured HBcAg constructs do not bind RNA (Hatton et al., 1992). In cryo-electron microscopy (cryo-EM) the structure of assembled HBV capsids was observed as either T=3 or T=4 icosahedral symmetry with 180 or 240 HBcAg protein units as dimers, respectively, whereby T=4 was present to a much larger extent (Crowther et al., 1994). Below the protein shell, wild type capsids also contain an inner shell that is icosahedrally ordered as well, which probably represents packaged RNA (Figure 3.5) (Crowther et al., 1994). Parts of the inner shell extend toward the outer capsids, which may indicate PS contacts between RNA and HBcAg proteins.

3.1.4 Reverse Transcription

During reverse-transcription Pol translocates several times along the template (Figure 3.6). First, the enzyme binds to the bulge of 5' ε and self-priming (Bartenschlager and Schaller, 1988; Wang and Seeger, 1992) synthesises a short fragment (Rieger and Nassal, 1996; Nassal and Rieger, 1996), which is complementary to that bulge as well as a part of DR1 (Figure 3.6 A). Due to self-priming involving a tyrosine residue in Pol the enzyme remains covalently attached to the 5' of the

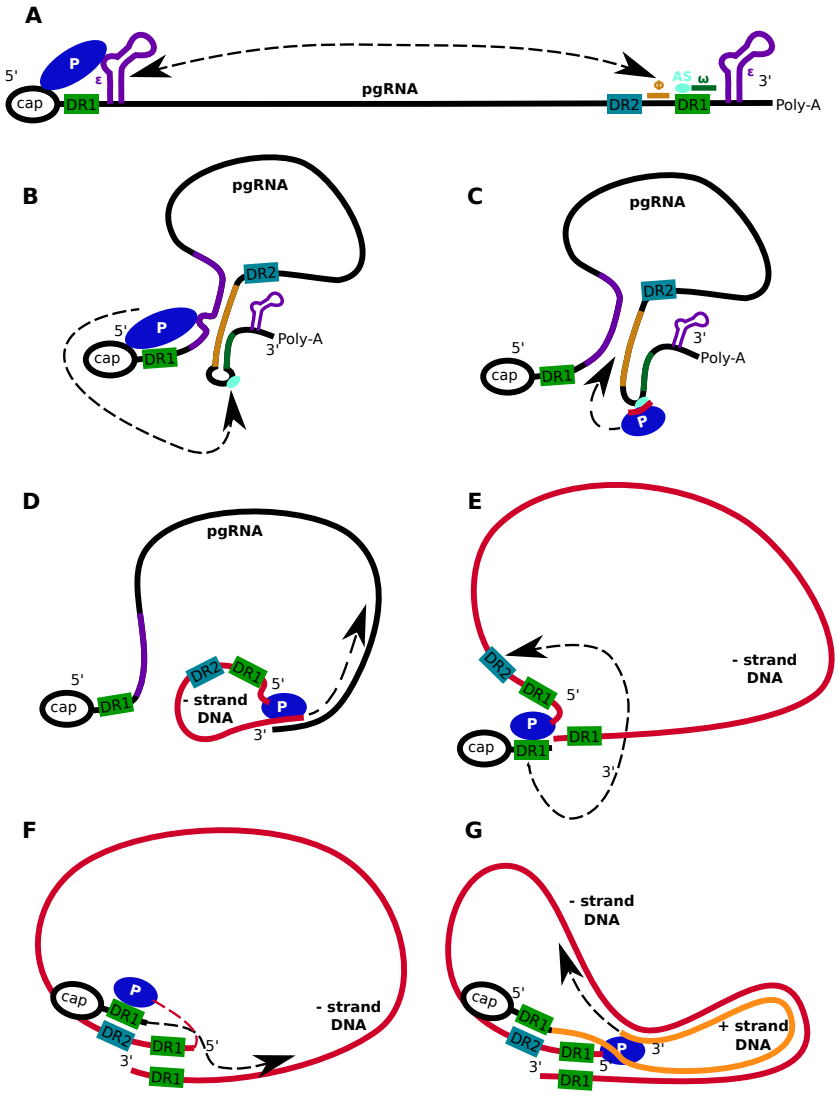

FIGURE 3.6: **Steps of reverse transcription in HBV.** (A) Packaging is triggered by Pol binding to ε and the 5' cap of the pgRNA. Once inside the capsid ε (purple) at the 5' end binds to ϕ (light brown) at the 3' end. (B) Interaction between ε and ϕ results in an effectively circularised pgRNA. Further interaction of ϕ with ω (dark green) exposes the primer acceptor site (AS, cyan) on DR1 (green). Pol synthesises a short fragment on the ε bulge, which is also complementary to the AS on DR1. Part of Pol itself serves as primer. Once the two sites are in close proximity Pol can translocate to the AS. (C) After translocation to the 3' end Pol continues DNA synthesis in 3' direction on the pgRNA. Pol stays covalently attached to the growing DNA. (D) As Pol syntheses minus-strand DNA (red) the pgRNA template is degraded. (E) Once Pol reaches DR1 at the 5' end, it stops reverse transcription of the minus-strand DNA. Instead it translocates to DR2 (blue) at the 5' end of the DNA, which is complementary to DR1. (F) The 5' end of the pgRNA serves as primer for plus-strand synthesis and stays attached to the growing DNA strand. (G) After synthesising using the 5' end until DR1, Pol translocates again to the 3' end of the minus-strand DNA. Plus-strand (orange) synthesis continues but does not complete.

growing DNA (Gerlich and Robinson, 1980). Then, Pol translocates from the 5' end to the 3' end and binds to the primer acceptor site (AS) on DR1 (nucleotides 1822–1825) (Figure 3.6 B) (Summers and Mason, 1982). From there it commences minus-strand DNA ((-)DNA) synthesis utilising the pgRNA as template, which is degraded in the process (Summers and Mason, 1982). Everything except for the terminal 15–18 nucleotides of the pgRNA including the 5' cap and DR1 are degraded (Haines and Loeb, 2007; Loeb et al., 1991). This small fragment is utilised as a primer for plus-strand DNA ((+)DNA) synthesis (Figure 3.6 E). In a second translocation step Pol moves to DR2 on the newly synthesised (-)DNA, which is complementary to DR1 on the small RNA fragment. From there (+)DNA synthesis begins using the other DNA strand as template (Figure 3.6 F). Once the 5' end of the (-)DNA is reached, Pol translocates a third time, to DR1 at the 3' end. (+)DNA synthesis continues for an unspecified amount of time resulting in a second DNA strand with variable lengths (Figure 3.6 G) (Will et al., 1987; Havert and Loeb, 1997).

A *cis*-acting element ϕ located around nucleotides 1769–1791 mediates the first Pol translocation step. It is complementary to the upper helix of ε and just upstream of 3' DR1 (nucleotides 1824–1835). Changes to this region that disrupt complementarity render the mutant severely impaired in (-)DNA synthesis (Tang and McLachlan, 2002). Restoring ϕ - ε base-pairing by compensatory mutations in ε ameliorates some of the effect on replication but does not restore it fully (Oropeza and McLachlan, 2007; Abraham and Loeb, 2006). Another *cis*-acting element ω (nucleotides 1830–1835) is necessary for the process. In an experiment Abraham and Loeb (2007) found that if it was inserted into another part of the genome together with ϕ and DR1, this was sufficient to trigger Pol translocation to that part of the genome. ω is also complementary to part of ϕ and mutations disrupting base-pairing decrease (-)DNA synthesis, whereas compensatory mutations in ω rescue all or part of replication (Abraham and Loeb, 2007). The first half of ϕ binds to the 5' portion of ε and the second half binds to ω. This brings

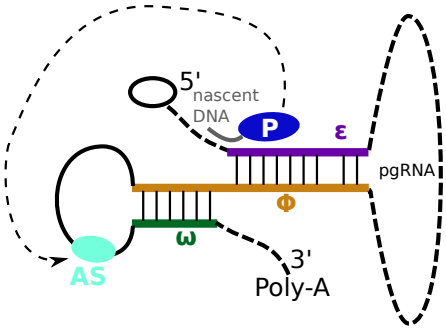

FIGURE 3.7: **Minus-strand synthesis initiation.** After polymerase (blue oval) binds to ε (purple) it synthesises a short primer (grey). ϕ (light brown) basepairs with both ε and ω (green), which circularises the pgRNA bringing 5' and 3' end in close proximity. Polymerase can then translocate from ε at the 5' end to the primer acceptor site (cyan oval) at the 3' end to commence minus-strans synthesis.

both ends of the pgRNA together to facilitate Pol translocation (Figure 3.7).

3.1.5 Genotypes

Originally, HBV had been grouped into four serotypes based on the surface antigen SHBsAg: adw, ayd, adr, and ayr. Later, a classification based on general genetic relatedness was proposed and is now the standard. Currently, there are eight recognized and two disputed genotypes: A–H (Okamoto et al., 1988; Norder et al., 1994; Miyakawa and Mizokami, 2003) and I–J (Huy et al., 2008; Tatematsu et al., 2009), respectively. Each has to vary from all others by at least 8% over the whole genomic sequence (Okamoto et al., 1988; Miyakawa and Mizokami, 2003). These genotypes show distinct geographical and ethnic distributions: While A, D, and G are found world-wide, B and C are mostly restricted to East and South-East Asia. E is found in Western Africa, F among Native populations in the Americas, H in Central and North America (reviewed in Kramvis et al. (2005)), I in Vietnam (Huy et al., 2008), and J in Japan (Tatematsu et al., 2009). Additionally, the genotypes have been further divided into subgenotypes by addition of Arabic numerals to the genotype letter. A subtype needs to differ by at least 4% from other subtypes but maximally 8%, a higher divergence indicates a distinct genotype (Norder et al., 2004). Also these show distinct distributions, e.g., A1

is common in Sub-Saharan Africa, whereas A2 is spread in North America and Northern Europe (Kramvis, 2014).

The differences in genomic sequences only translate to some extent to the actual amino acid sequence of the viral proteins. Comparing the primary sequence of HBcAg in all HBV genotypes shows some residues that vary more often. To find out whether these differences could affect PS binding, i.e. whether they are on the inside of the capsid, they were mapped onto the three-dimensional x-ray structure. Genotype differences were spread over the entire structure. Interestingly, there is even some variation in the nucleic acid-binding carboxy-terminal tails. Therefore, it would not be surprising to find variation in PS usage among the genotypes.

3.1.6 *Hepadnaviridae*

Hepadnaviridae infect a wide range of animals. They are subdivided into *Orthohepadnaviruses*, which infect mammals, and *Avihepadnaviruses*, which infect birds. Recently, a similar virus was isolated in fish and amphibian species (Hahn et al., 2015; Dill et al., 2016). Interestingly, they do not cluster together in a phylogenetic tree and thus do not represent simply a third genus (Dill et al., 2016). Since not much else is known about the newly discovered *Hepadnaviridae* species in fish and amphibians, the focus will be on the well-studied mammalian and avian viruses. Paleovirological research has found evidence of *Avihepadnaviruses* as far as 82 million years ago (Suh et al., 2013). It is therefore believed that birds were the first hosts of *Hepadnaviridae* and a transmission to mammals, which gave rise to *Orthohepadnaviruses* occurred later in evolution (Suh et al., 2013; Littlejohn et al., 2016). As opposed to rodent and bird hepadnaviruses, BtHV is capable of infecting human liver cells, which points towards HBV having originated as a zoonotic infection, i.e. acquired from animals. There is some evidence that bats were the first mammalian hosts. It was found that BtHV has almost as much intraspecies genomic variation as they differ from other *Orthohepadnaviruses* (Drexler et al., 2013; Littlejohn et al., 2016). However, this

is just one hypothesis and the origin of HBV is still debated. The geographic spread and wide array of hosts points towards a long evolutionary history with co-species evolution, divergence and host jumping (Littlejohn et al., 2016). The species closest related to human HBV are other primate infecting viruses, with whom they share a lineage (Littlejohn et al., 2016). Interestingly, non-human primate HBV show similar genomic relatedness patterns with each other based on geographic location as human viruses. If the hosts live in close proximity or even have overlapping habitat, the viruses show more genomic similarity (Starkman et al., 2003).

Over this long evolutionary time scale the *Hepadnaviridae* have maintained many similarities. First and foremost they are all DNA viruses that replicate via an RNA intermediate. Their genomes are circular and encode for a reverse transcriptase (Pol), structural proteins (SHBsAg), and capsid protein (HBcAg). The three ORFs are highly overlapping resulting in a comparatively small genome. Only mammalian viruses are thought to encode a forth gene called X; however, a similar protein was discovered in DHBV, which may perform some but not all of the same functions as the human equivalent (Chang et al., 2001). In addition to the proteins mentioned above, these two genera also have ε in common. While the exact sequence and somewhat structure differ, the relative position and function of this *cis*-acting element is nearly the same (Beck et al., 1997; Kramvis and Kew, 1998). Other cis-acting elements such as ω and ϕ have thus far not been identified in *Avihepadnaviruses* (Maguire and Loeb, 2010).

3.2 SELEX Data

Systematic evolution of ligands by exponential enrichment (SELEX) is a method to identify nucleic acid ligands that have a high affinity for a protein in question (Tuerk et al., 1990). Since PSs are meant to bind to CPs with high affinity, this method provides a suitable starting point for finding PS-like RNA sequences. Binding nucleic acid sequences are found and affinities resolved through several

rounds of exponential enrichment. Briefly, the protein is immobilised and then exposed to an excess of a random set of RNA oligonucleotides (oligos). Since the nucleic acid sequences are in excess, there is competition for binding to the protein ensuring a selection for sequences with higher affinity. After washing away unbound sequences, binders are amplified through polymerase chain reaction (PCR). The amplified sequences are then used as a starting set for the next round. These steps are repeated a few times and in each round already binding sequences are competing with each other leading to further selection for higher affinity ligands. Finally, the enriched ligands are sequenced (Tuerk et al., 1990).

Collaborators from the Stockley group (Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds) performed SELEX experiments testing a random library of 40-nucleotides-long RNA oligos for binding to HBV CP, HBcAg (Patel et al., 2017). How often certain sequences appeared in the final sequencing, i.e. the multiplicity, was used as a proxy for the affinity of that sequence. The list of aptamer sequences sorted by their multiplicity was utilised for the first part of the analysis. PCR requires primers on both ends of the sequence to be amplified. These primers were removed from the aptamers sequences before further analysis.

3.3 SELEX Aptamer Analysis

3.3.1 Multiplicities and Nucleotide Composition

To determine the level of enrichment for specific nucleotide sequences I have performed a comparison of the multiplicities before and after selection. The total number of distinct aptamers was 1,664,890. The highest number of multiplicities was 65,802 and there were 1149 aptamers with a multiplicity of 100 or higher. The sequences and multiplicities of the top ten aptamers are shown in Table 3.1. There was a rapid decline from the highest multiplicity with the fourth aptamer having less than a tenth of the multiplicty of the first. The naïve library, i.e. the RNA oligos used for the experiment before enrichment, had multiplicities of up

TABLE 3.1: Sequences of the top 10 aptamers with primer parts removed.

Name	Sequence	Multiplicity
A1	TGCGGGGTTGGTTGGGAAGGGGAGAGGATTTGAAGGACAG	65802
A2	AAGGCGGGAGGGAGGGGAAGGATGGGATGAGAAGAACGGG	14255
A3	TTGCGGGGTGGATGGGAGGGGCTTAGGGATGAATGGACGG	7912
A4	TTGCGGGGTGGATGGGAGGGGCTTAGGGATGAATGGACGG	6435
A5	AGGGGAGGCAGGGCGGGGACAGGATATTGCACACAACGGA	4954
A6	AGGGGGGAGGGAGGAGGAAAGAGAAGAACGGACGCGTGGG	4860
A7	AAGGGAGGAGTAGGAGGAAGGGAAGGCGGGATGAGGCAAG	4833
A8	GCATGGGGTGGAGGCTGGGGAACAGAGATTGGGTTGATGG	4610
A9	GGGGGGAGGTAGGGCGGCGGATAAGGGATCGGTAGCGTGG	4034
A10	ATTTGGGGAAGGAAGGGTAGGGGACGGGATCAGATTGCGG	3775

to 4, illustrating the level of enrichment that has occurred. At each PCR round a doubling of a sequence is possible and ten rounds of PCR are performed during each SELEX cycle (Bunka et al., 2011). For this experiment ten rounds of SELEX were performed (Patel et al., 2017). Thus, the maximum possible enrichment is $(2^{PCRrounds})^{SELEXrounds}$, which is $(2^{10})^{10} = 2.27 \times 10^{30}$. However, this is a very theoretical number as it assumes all PCR steps to work perfectly and no sample being lost during selection. In reality the expected maximal yields would be much lower.

The nucleotide composition differed greatly between aptamers and naïve library and is shown in Table 3.2. The naïve library had an almost equal nucleotide usage, whereas purines were highly enriched in the aptamers.

3.3.2 *k*-tuples

Along the same lines as total nucleotide compositions, we can examine sequence fragments of k letters and look at their statistical distribution. A k-tuple is

TABLE 3.2: Nucleotide composition of HBV SELEX aptamers and the naïve library in %.

	G	**C**	**A**	**T**
Aptamers	40.97	9.09	34.30	15.64
Naïve	24.64	22.03	26.10	27.22

defined as a string of k consecutive nucleotides. With four letters, ACTG, a k-tuple will have 4^k possible combinations. Analysing tuples of different lengths can give insight into the frequencies of certain motifs. This method was applied to structures that can be formed by the aptamer sequences to identify shared apical loop motifs. To avoid the assumption that the SL that performs the PS function is (part of) the minimum free energy (MFE) structure on that fragment, analysis was instead performed on a set of suboptimal structures. To do so the sequences were folded in RNAsubopt on default settings with the energy range set to 5 kcal/mol (Lorenz et al., 2011; Hofacker et al., 1994; Wuchty et al., 1999), providing a list of all secondary structures on the RNA sequence within 5 kcal/mol of the MFE structure. From these sets the apical loop sequences were extracted. Note that on any given fragment several SLs can form next to each other or as part of a forked larger structure. Since any of these SLs could be binding to CP, individual SLs were extracted from the fragment structures and considered separately. SLs with isolated base-pairs, i.e. a bulge or internal loop on both sides of the base-pair, were considered too unstable and filtered out. The percentage occurrence of each 3- and 4-tuple in the apical loop sequences was calculated. The naïve SELEX library served as a negative control.

If the distribution of nucleotides in apical loops was random, each 3-tuple and 4-tuple would occur 1.6% and 0.4% of the time, respectively. These percentages are calculated as follows: Since there are four different nucleotide options for each position, the total number of 3-tuples is $4^3 = 64$, which means that the probability of any given 3-tuple to occur is $1/64 = 0.0156$ or 1.6%. Similarly for 4-tuples there are $4^4 = 256$ combinations so the probability for each is $1/256 = 0.0039$ or

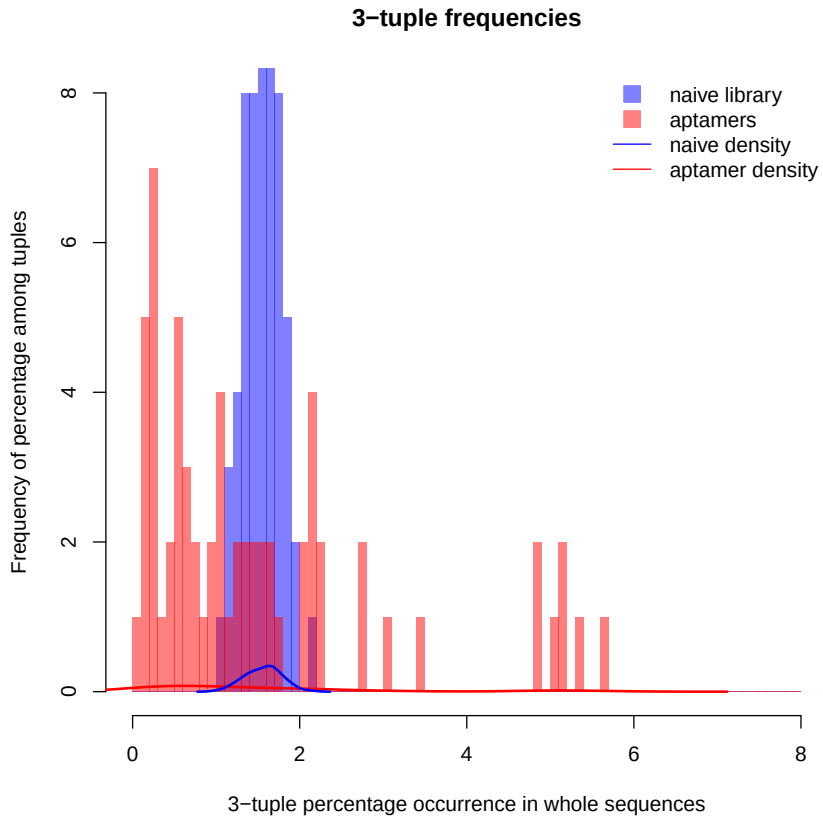

FIGURE 3.8: **3-tuple frequencies in complete sequences aptamers and naïve library.** The frequency distribution of 3-tuple percentages in aptamers (red) and in the naïve library is shown. The naïve library shows a distribution close to the average 1.6% confirming a near unbiased library. That selection took place can be seen from the aptamers having a positively skewed distributions with a few high percentage tuples.

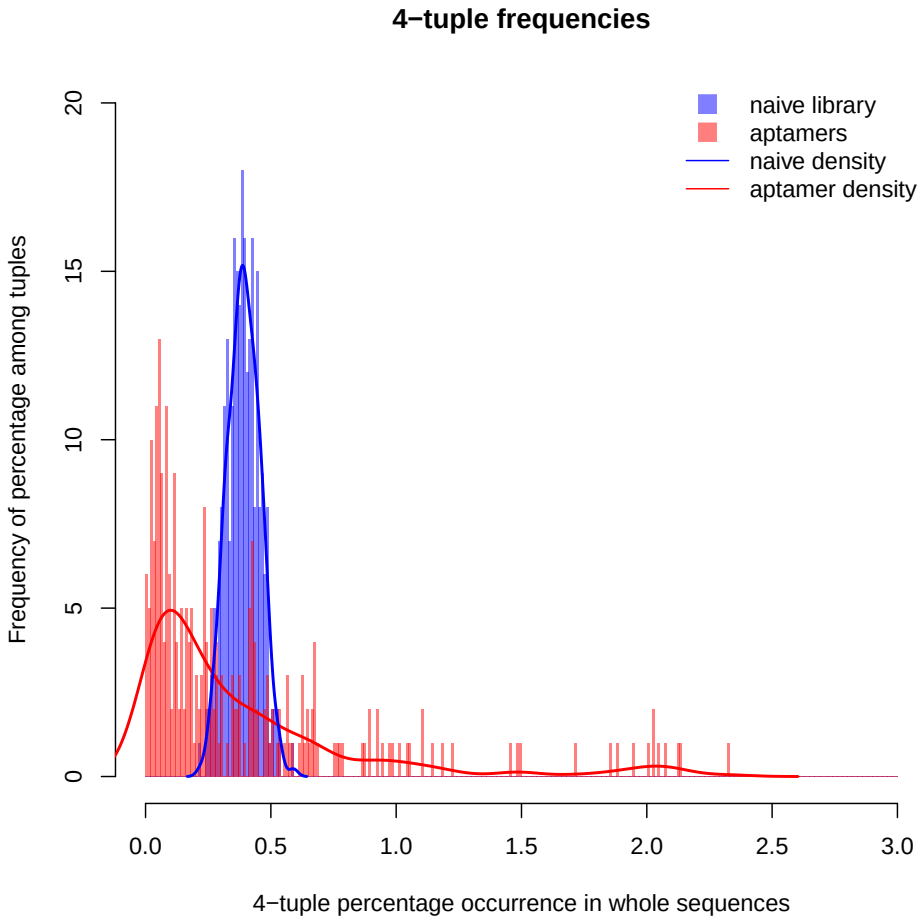

FIGURE 3.9: **4-tuple frequencies in complete sequences of aptamers and naïve library.** The frequency distribution of 4-tuple percentages in aptamers (red) and in the naïve library is shown. The naïve library has a near normal distribution centred around 0.4% as expected. The aptamers' distribution shows a pronounced positive skew with many low percentage tuples and a few high percentage tuples as seen by the tail.

0.4%. Since biases for certain combinations may have already been present at the start of the experiment, the aptamer samples were compared to the naïve library. First the tuple distributions in the whole sequences were compared. A histogram of tuple frequencies in aptamer versus naïve showed that for 3-tuples the naïve library had a narrow, normal-looking distribution of tuple frequencies with a mode close to the expected 1.6 %, whereas the aptamer samples showed a more varied distribution with a few enriched tuples (Figure 3.8). A similar trend was observed for 4-tuples. The naïve library still showed a normal-like distribution, which proves an unbiased original sample pre-selection. The aptamer sequences, on the other hand, had a pronounces positive skew indicating the enrichment of certain motifs in general (Figure 3.9). When comparing only the apical loop sequences, the effects were even stronger. For both tuple sets the naïve library showed a more skewed distribution demonstrating that some motifs are more likely to occur in the loops of SLs. Whilst in 3-tuples the distribution was almost normal, in 4-tuples the naïve library also showed a noticeable positive skew. The aptamers had even more pronounced positive skews with most tuples barely present and a long tail of a few frequent tuples (Figures 3.10 and 3.11).

To assess if certain tuples were enriched in the aptamer sample versus naïve, the mean and standard deviation (σ) of tuple occurrences were calculated for the naïve sample. If the percentage of a given tuple was more than 5σ different from the naïve mean and showed a fold change (fc) of at least 2, it was considered enriched. The fc was calculated by dividing the percentage occurrence in the aptamers by the one in the naïve sample, e.g. a tuple that occurred 2% in the aptamers and only 1% in the naïve library had a fc of $2/1 = 2$. Note that 5σ account for 99.99994% of the sample meaning that the likelihood of this difference occurring by chance is only $6 * 10^{-5}$ %. In fact, the highest values in the naïve sample were 3.1% (2.5σ) and 0.96% (3.11σ) for 3- and 4-tuples, respectively. Thus enriched 3-tuples are summarised in Table 3.3. The highest fc was observed in tuple GGG with an occurrence of 5.36% and a 6.38 fc. Table 3.4 summarises the

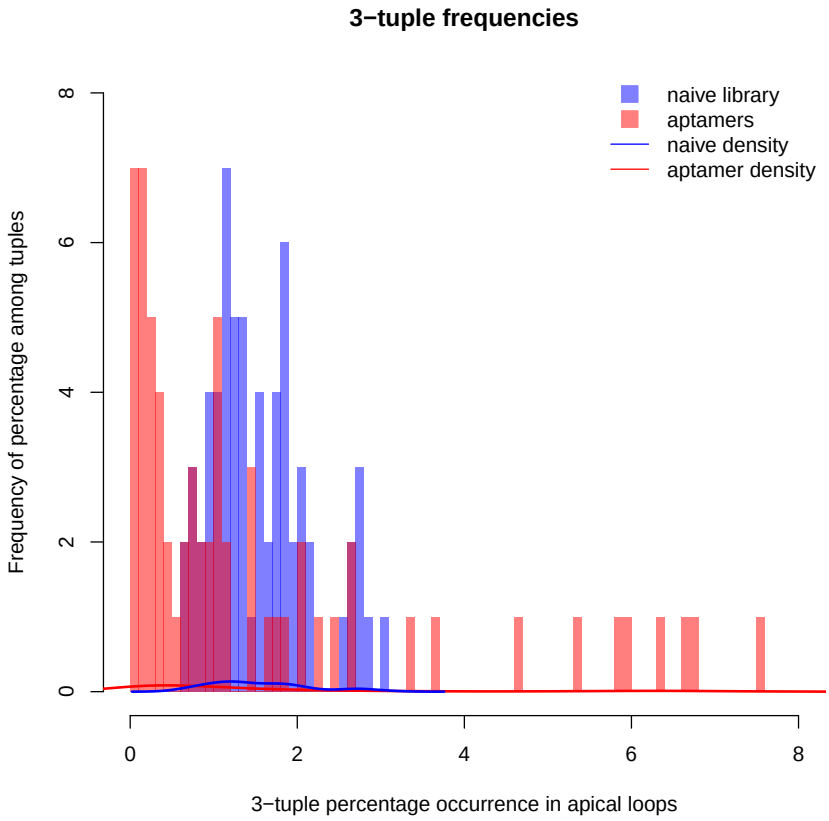

FIGURE 3.10: **3-tuple frequencies in apical loop sequences of aptamers and naïve library.** The frequency distribution of 3-tuple percentages in aptamers (red) and in the naïve library is shown. The naïve library now shows an wider distribution than for the whole sequence but is still centred around 1.6%. The aptamers have an even more positively skewed distribution with many low percentage tuples and a few high percentage tuples as seen by the tail.

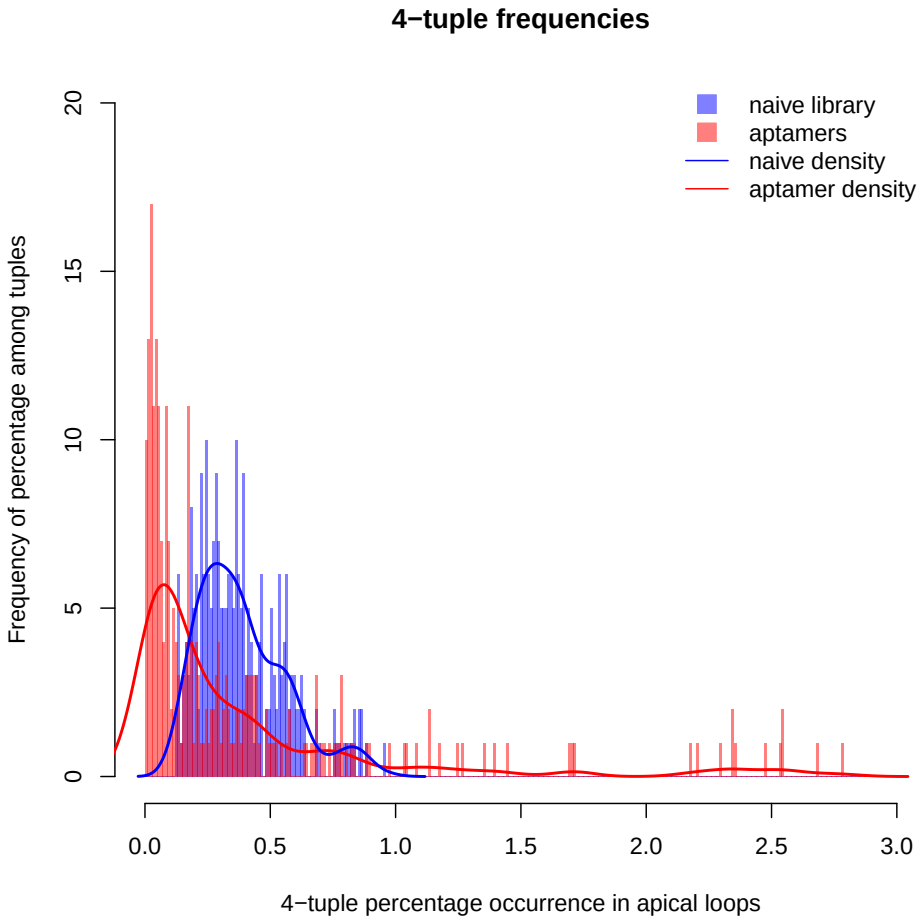

FIGURE 3.11: **4-tuple frequencies in apical loop sequences of aptamers and naïve library.** The frequency distribution of 4-tuple percentages in aptamers (red) and in the naïve library is shown. The naïve library has a slight positive skew centred around 0.4% now. The aptamers' positive skew is much more pronounced with many low percentage tuples and a few high percentage tuples as seen by the tail.

TABLE 3.3: Significantly enriched 3-tuples sorted by fold change.

Tuple	Occurrence	Fold change	σ from mean
GGG	5.36	6.38	6.27
GAG	6.71	5.69	8.52
GGA	6.64	5.19	8.40
AGG	5.88	5.11	7.13
AGA	7.50	3.97	9.83
AAG	5.98	3.20	7.30
GAA	6.40	3.02	8.00

enriched 4-tuples with the most highly enriched being GGAG (2.35%, 9.79 fc). The difference in tuple enrichment can also be seen in the histograms showing the frequencies of tuple percentages in the samples. While both the aptamers and the naïve sample have a positive skew, it is much more pronounced in the aptamers indicating an enrichment of some tuples in the apical loops of the aptamers. These data show that all possible combinations of G and A are enriched in the aptamer loop sequences as well as GAUA and GGAU. 3-tuples AUA and GAU were not significantly enriched but showed some increase in fold change. In conclusion, an apical loop that is purine-rich, i.e. contains more G and A than U and C, with a preference for G, is likely to bind to HBV capsid protein.

3.3.3 Top Aptamer Folds

To confirm the validity of the tuple analyses the five aptamers with highest multiplicity, which are assumed to be folding into high affinity PS-like SLs, were folded in Mfold (Zuker, 2003). Up until now only the differing sequences, i.e. aptamer sequences without primers, were considered. Since small, single SLs were analysed for loop tuples, this removed a source of bias as the primer portions could form SLs as well. In this case, however, larger global structures were considered, for which the flanking primer sequences may have provided stability during the SELEX experiment. Up to 500% sub-optimality, i.e. any structure differing

TABLE 3.4: Significantly enriched 4-tuples sorted by fold change.

Tuple	Occurrence	Fold change	σ from mean
GGAG	2.35	9.79	11.02
AGGG	2.35	9.40	11.02
GAGG	2.30	9.20	10.73
GGGA	2.69	8.97	12.94
GGGG	1.71	8.55	7.40
AGAG	2.55	7.29	12.15
AGGA	2.55	6.71	12.15
GAGA	2.79	6.64	13.50
GAAG	2.54	6.35	12.09
AAGG	2.21	5.97	10.23
GGAA	2.18	4.84	10.06
GGAU	1.45	4.14	5.93
AGAA	2.48	3.94	11.75
AAGA	2.36	3.75	11.07
AAAG	1.70	2.88	7.34
GAAA	1,72	2.46	7.46
GAUA	1.40	2.41	5.65

from the MFE structure by up to 500%, were allowed to ensure an ensemble of structures. For each aptamer two representative folds were selected and visualised in VARNA (Darty et al., 2009). As expected, each aptamer could form at least one stable purine-rich SL, which was often stabilised through base-pairing with parts of the primer sequences (Figure 3.12). Aptamer 1 presented one apical loop with `AAGGGGAGAGGA` or one with just `GAGAGGA`. These include 3-tuples AAG, AGG, GGG, GGA, GAG, and AGA and 4-tuples AAGG, AGGG, GGGG, GGGA, GGAG, GAGA, AGAG, GAGG, and AGGA. All of these are among the enriched ones (see Tables 3.3 and 3.4).

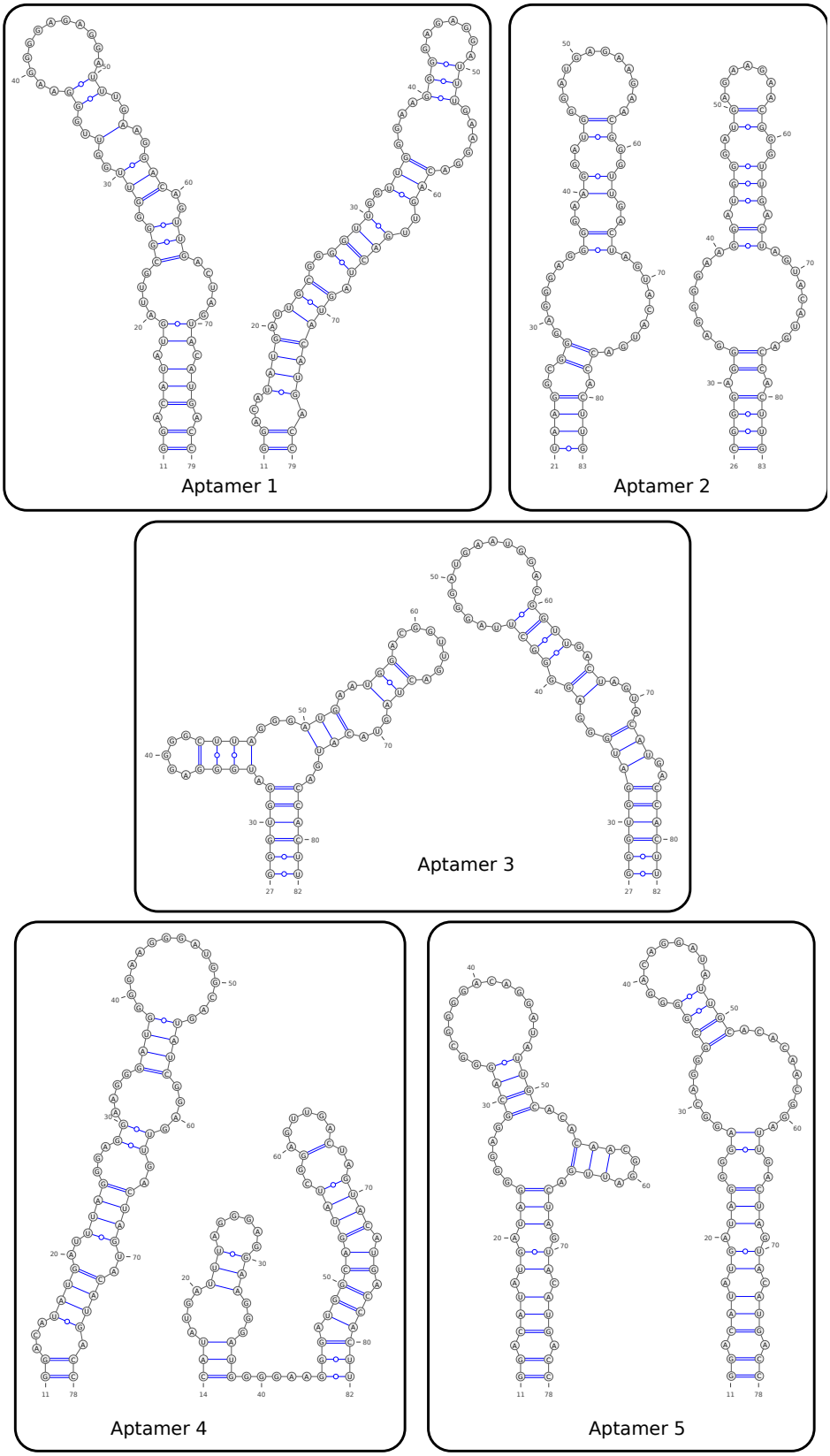

FIGURE 3.12: **Folds of the five aptamers with highest multiplicities in enriched library.** The complete aptamer sequences were folded in Mfold (Zuker, 2003) allowing up to 500% suboptimality. Two representative structures were visualised in VARNA (Darty et al., 2009) with leading and trailing single stranded portions removed.

3.4 Putative Packaging Signals in HBV Strains

3.4.1 Sequence Selection

Based on the 3- and 4-tuples, it was hypothesised that a PS motif would be enriched in purine bases (A and G). Further analysis was performed on actual HBV genomic sequences, as follows: All complete genomes were downloaded from NCBI and sequences containing mutants or ambiguous bases were filtered out leaving 750 genomes from different genotypes. Of these, 20 were randomly selected using python random number generation. 14 of the selected sequences were from genotype C, two from genotype B, two from G, one from genotype D, and

TABLE 3.5: Randomly selected HBV genomic sequences.

Alias	Accession number	Genotype
seq1	KC510648.1	B
seq2	AB206817.2	C
seq3	AF223955.1	C
seq4	AY781181.1	unknown
seq5	AB116266.1	D
seq6	AB195943.1	C
seq7	KR014086.1	C
seq8	KR014078.1	C
seq9	KR014072.1	C
seq10	KR014055.1	C
seq11	KR014014.1	C
seq12	KR013944.1	C
seq13	KR013939.1	C
seq14	KR013921.1	C
seq15	KR013816.1	C
seq16	KR013800.1	C
seq17	AB375170.1	G
seq18	AB375169.1	G
seq19	EU796069.1	C
seq20	AB540582.1	B

one unknown (Table 3.5). Of the 14 genotype C sequences, 10 were from the same publication (Hao et al., 2015). This publication supplied 339 genomes, which is about 50% of the total number of sequences; thus, explaining the large proportion of selected genomes stemming from there. To ensure reasonable similarity of the strains, they were aligned using ClustalOmega with default settings (Sievers et al., 2014; Goujon et al., 2010). The two genotype G and four genotype C sequences were removed from the set due to large differences with insertions/deletions and other mutations, resulting in a set of 14 random sequences for further analysis: KC510648.1, AF223955.1, AY781181.1, AB116266.1, AB195943.1, KR014086.1, KR014072.1, KR014055.1, KR013939.1, KR013921.1, KR013816.1, KR013800.1, EU796069.1, AB540582.1. To these, the laboratory strain (NC_003977.1), i.e. the HBV strain used by collaborators for experiments, and the current NCBI reference strain (NC_003977.2) were added.

3.4.2 Alignment Using Bernoulli Scores

To align the aptamer sequences to the HBV genomes from the database we use a probabilistic measure following GeneBee (Brodski et al., 1995). In essence, a Bernoulli score $B(L, j)$ is a measure of probability for two sequences of a length L to match non-contiguously with no more than j mismatches. B is normalised to ranges from 0 to L and is equivalent to the probability of B contiguous matches. This allows the comparison of non-contiguous matches to contiguously matching sequences as a relative probability score. It is calculated as $B(L, j) = L - \log_4 \sum_{i=0}^{j} [3^i \binom{L}{i}]$ and is based on the method by Altschul and Erickson (Altschul and Erickson, 1986). The formula is derived in Equations (3.1)–(3.3).

The probability of exactly j mismatches at length L can be calculated by imagining pulling nucleotides randomly from a bag where they are equally distributed with probabilities of $\frac{1}{4}$. So the probability P of a match is $\frac{1}{4}$ and of a mismatch is $\frac{3}{4}$. Thus, disregarding the order of match/mismatch, binomial

theorem can be applied to get:

$$P(L,j) = \left(\frac{1}{4}\right)^{L-j} \left(\frac{3}{4}\right)^{j} \frac{L!}{(L-j)!(j)!} \tag{3.1}$$

The probability of at most j mismatches at length L is calculated through a number of steps:

The sum over all combinatorial possibilities is

$$\bar{P}(L,j) = \sum_{i=0}^{j} \left[\left(\frac{1}{4}\right)^{L-i} \left(\frac{3}{4}\right)^{i} \frac{L!}{(L-i)!(i)!} \right] .$$

Let $\dfrac{L!}{(L-i)!(i)!}$ be denoted by the usual notation $\binom{L}{i}$. This simplifies to

$$\bar{P}(L,j) = \sum_{i=0}^{j} \left[\left(\frac{1}{4}\right)^{L-i} \left(\frac{3}{4}\right)^{i} \binom{L}{i} \right] .$$

Combining the power terms gives

$$\bar{P}(L,j) = \sum_{i=0}^{j} \left[\left(\frac{1}{4}\right)^{L-i+i} 3^{i} \binom{L}{i} \right] ,$$

which the $(\frac{1}{4})^{L}$ can be factored out of:

$$\bar{P}(L,j) = \left(\frac{1}{4}\right)^{L} \sum_{i=0}^{j} \left[3^{i} \binom{L}{i} \right] . \tag{3.2}$$

Taking the logarithm to base $\frac{1}{4}$ allows normalisation of the Bernoulli score to a number between 0 and L:

$$B(L,j) = \log_{1/4}[\bar{P}(L,j)]$$

Using the properties of logarithms this becomes

$$B(L,j) = \log_{1/4} \left[\left(\frac{1}{4}\right)^{L} \right] + \log_{1/4} \sum_{i=0}^{j} \left[3^{i} \binom{L}{i} \right] ,$$

which simplifies to

$$B(L,j) = L + \log_{1/4} \sum_{i=0}^{j} \left[3^i \binom{L}{i} \right] .$$

Finally, using $\log_{1/a} = -\log_a$ the final formulation is reached:

$$B(L,j) = L - \log_4 \sum_{i=0}^{j} \left[3^i \binom{L}{i} \right] . \tag{3.3}$$

If j is 0, i.e. there were no mismatches in length L, the term $\log_4 \sum_{i=0}^{j} \left[3^i \binom{L}{i} \right] = 0$ so $B(L,0) = L$. To calculate the actual probability of observing a given Bernoulli score:

$$\bar{P}(L,j) = \left(\frac{1}{4} \right)^{B(L,j)} . \tag{3.4}$$

3.4.3 Bernoulli Peaks

The processed SELEX aptamer sequences were aligned to each of the 16 selected HBV genomes generating Bernoulli scores. A sliding window with variable fragment size was used to find the maximal score at each point. It represents the likelihood of a match of certain length with thus many mismatches. Only scores higher than 12, which are equivalent to the probability of a contiguous 12 nucleotide match, were considered. Taking into account the length of a HBV genomes of approximately 3000 nucleotides, the probability of such a match is $P = \left(\frac{1}{4}\right)^{12} \times 3000 \approx 1.79 \times 10^{-4}$. To identify genomic regions of high similarity to the aptamer sequences, a score of $+1 \times$ *multiplicity* was added at that position for every match by an aptamer that had a Bernoulli score above 12 generating a weighted histogram. For instance, if an aptamer had a multiplicity of 3000 and matched at positions 100, 106, 107, and 110, then a score of 3000 would be added to each of these positions. If another aptamer had a multiplicity of 150 and matched positions 95, 100, and 106, then the scores in that area would be 150 at 95, 3150 at 100 and 106, and 3000 at 107 and 110. This resulted in a

TABLE 3.6: Aptamers producing a Bernoull score ≥ 12 and multiplicity ≥ 100. The position of the aptamer in the complete list of aptamers sorted by multiplicity, the sequence, and the multiplicity (mult.) is given.

Rank	Sequence	Mult.
95	AGAGAGGGAGGCTGGGGGAGGAGAAGGGATGCAATCGGTG	734
318	GTGGGCGGAGGGGAGGAGGATAAAGGTGAGGCGTAGATGG	283
644	GGGAAGGGAAAAAGGAAATTAAGAGTATAGATATGGCGCA	155
694	GAGGGAGATGAGAGAAAAGAAATAGGAACATATTGCGGGG	146
754	TGGGGGGGGAAGGAACGGGATGAGTAGAGGAATGTGGCGT	137
858	CAGGATGAGGAGGGCGGGGAGGAGGAAAGGATAACAGGCA	123
983	GAGGAGAAGTAGAAGAATGAAAAAAGGGATAATTGGAGGG	112
1108	GACTGCGAGGTGGATGGGTGGGGAGAGGAGATTGTGGATG	102

number of peak regions called Bernoulli peaks.

To assess how well the aptamer sequences aligned to the genomes overall and how much the results were affected by multiplicities, aptamers that produced a Bernoulli score above 12 on the laboratory strain were filtered. Surprisingly, only 12,884 out of 1,664,890 (0.8%) fulfilled that condition. Of these 12,884 the highest multiplicity was 734 and only 8 had a multiplicity above 100 (Table 3.6). This outcome illustrates the strictness of the Bernoulli score method. Achieving a score of 12 and above is not trivial. This does not mean that SELEX was not successful or that the method is not appropriate. In other viruses PS motifs are short and variable so there is room for variability even when a SL can function as PS so achieving a high Bernoulli score is still unlikely. To ensure that the method is nevertheless appropriate the outcomes will be compared to the results of the tuple analysis on the whole set of aptamers.

The naïve library sequences were also aligned to the selected genomes to determine the random noise level. The highest peak for the naïve set was 400 and was used as cut-off point for Bernoulli peak selection. Only the Bernoulli

peaks above the random noise level were considered. To ensure that the corresponding peaks in different strains were aligned, a multiple sequence alignment in ClustalOmega with default settings (Sievers et al., 2014; Goujon et al., 2010) was utilised to shift the peaks. Following the shift the exact nucleotide positions for each peak were extracted. When the positions were compared between sequences, peaks in close proximity, i.e. within 10 nucleotides of each other, were considered equivalent. Bernoulli peaks that were present in at least 13 out of the 16 genomes (81.25%) were considered conserved and further analysed. The conserved peaks were around genomic positions 750 (15/16), 990 (13/16), 1745 (16/16), 1925 (16/16), 2235 (15/16), 2620 (16/16), 2780 (16/16), 2850 (13/16), and 3025 (15/16). Figure 3.13 shows all Bernoulli peaks with conserved ones marked with a star. Notably three sequences had deletions that affected peak 2850 and two had a deletion before peak 2236. The laboratory strain contained all peaks and was used together with seq3 (AF223955.1), which also contained all peaks, and the reference strain for in-depth motif search.

3.4.4 Consensus Motif

For each selected genome, the sequences ± 20 nts from the Bernoulli peak nucleotide were extracted and folded into all possible structures with negative folding energy using a Fortran 90 implementation of the Mfold (Zuker, 2003) algorithm called Tfold (Dykeman, unpublished). Then a similarity analysis on structure and apical loop sequence was performed intergenomically for each peak and intragenomically for all peaks. The basis of this analysis were the results from the tuples so that specifically apical loop sequences that were purine-rich were selected (Figure 3.14). The final selection of loop sequences from each peak were aligned resulting in the consensus motif: RGAG, usually at the end of the loop (Figure 3.15). Interestingly, only Aptamers 1 and 4 presented apical loops with such a motif. As SL1 and SL2 also deviate from RGAG it is possible that the real PS motif is even less strict than that. Aptamers 2 and 3 also presented GGAUG

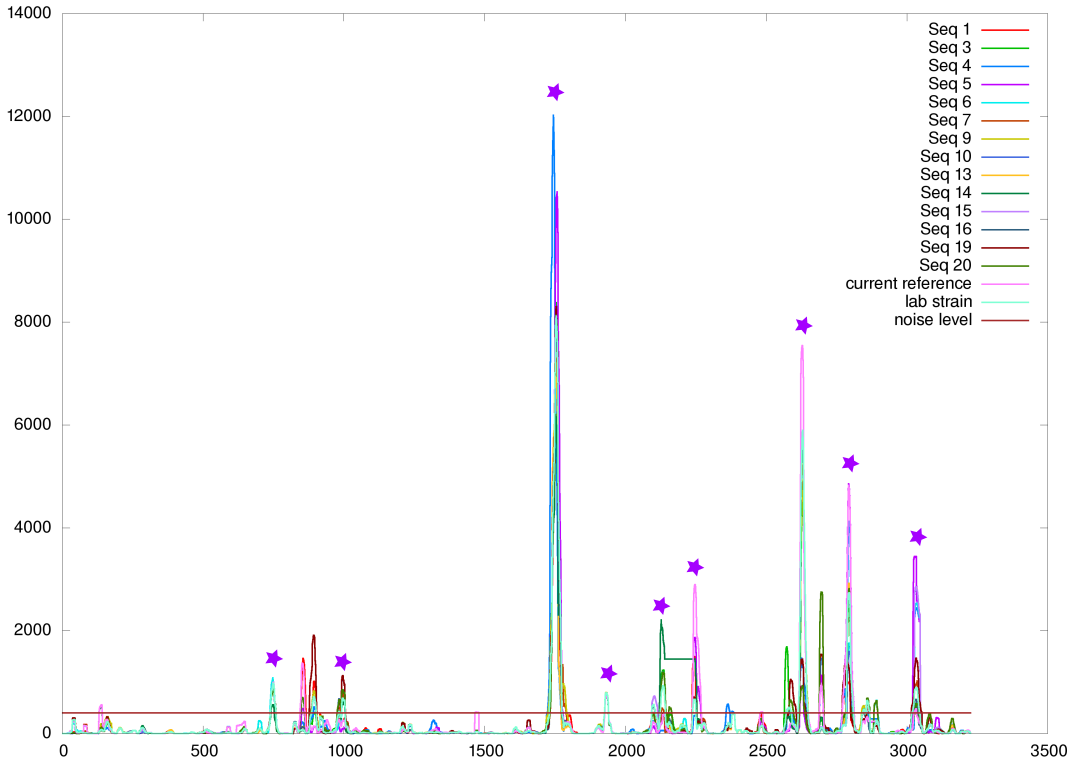

FIGURE 3.13: **Bernoulli Peaks in 16 HBV strains.** Differently coloured lines represent smoothed Bernoulli scores for one strain. The horizontal red line is the threshold determined from maximum peak height in the negative control. On the x-axis genomic positions are shown and on the y-axis are the Bernoulli scores. Peaks conserved in at least 13/16 strains are marked with a purple star.

like SL1. Further laboratory tests were required to confirm this predicted motif.

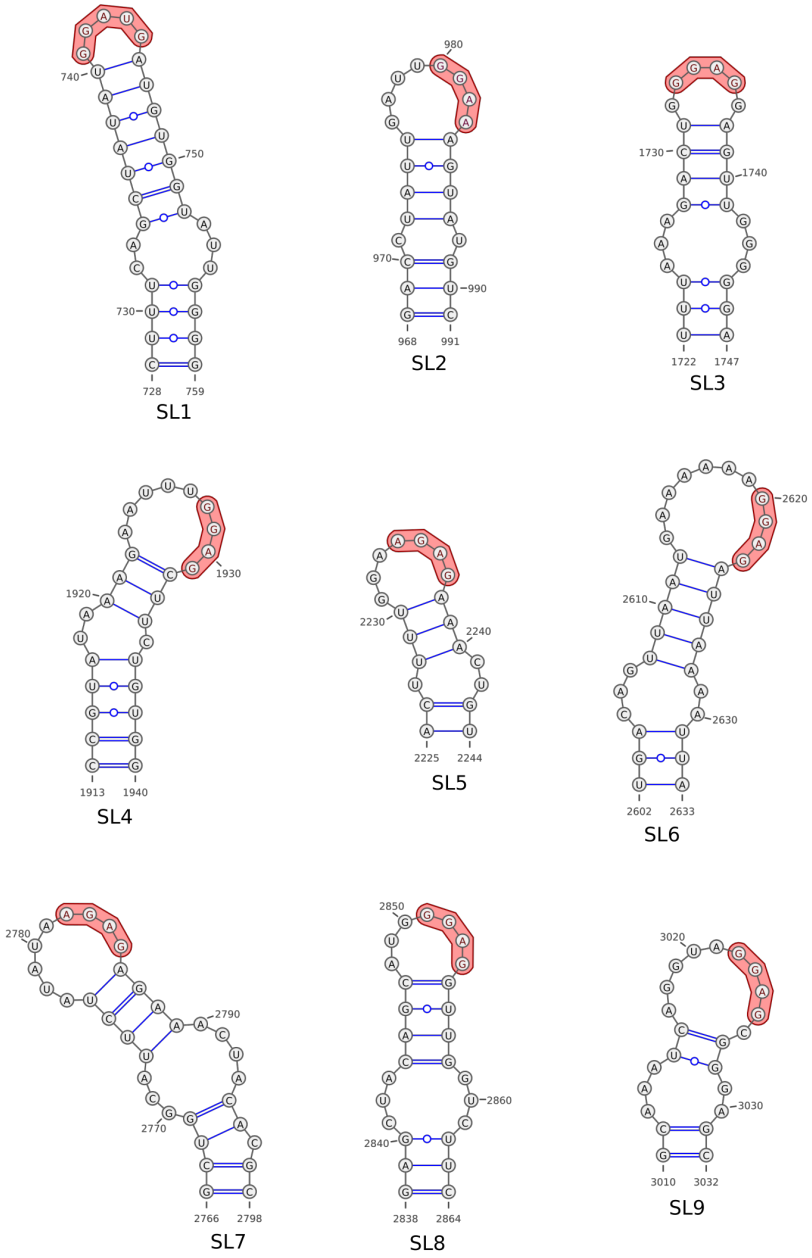

FIGURE 3.14: **Putative packaging signals at Bernoulli peaks.** Structures selected from similarity analysis shown. The consensus motif RGAG is marked in red. The structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

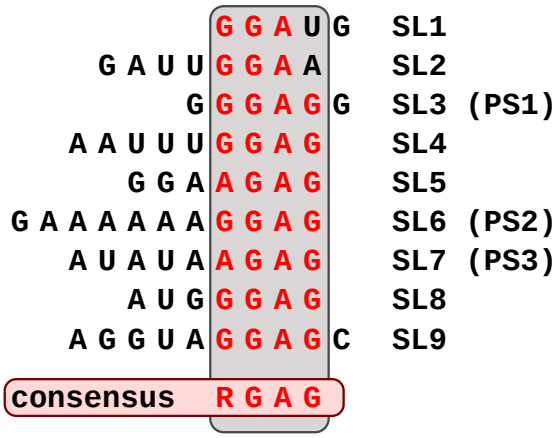

FIGURE 3.15: **Alignment of apical loop sequences.** The apical loop sequences from the structures shown in Figure 3.14 are aligned to identify the consensus motif RGAG marked in red. The SL numbers as shown in Figure 3.14 above are shown next to the respective apical loop sequences. PS1, PS2, and PS3 denote the highest peaks in the Bernoulli plot.

3.5 Experimental Validation

The work described in this section was performed by collaborators in the Stockley group (Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds). The three PSs from the highest peaks, i.e. SL3, SL6, and SL7 — called PS1, PS2, and PS3, respectively (see Figure 3.13), were tested as single stem-loops for their capability to trigger re-assembly of CP *in vitro*. The oligos were labelled with a dye to allow measurement of the hydrodynamic radius, which refers to the size of a macromolecule. In this case, it represents the agglomeration of single labelled SLs as they bind CP, which interact with each other building up a capsid. PS1, PS2, and PS3 all managed to trigger re-assembly and resulting capsids were resistant to RNase A treatment (Figure 3.16, left). This indicates that the capsids were complete and impervious to the RNA-degrading enzyme. Successful re-assembly was also confirmed visually by electron microscopy (Figure 3.16, right). To further confirm the predicted RGAG motif mutated versions of PS1 were used. Any change to the motif or internal loop resulted in loss of function (Table 3.7). Even when capsids assembled they were not RNase A resistant any more (Patel et al., 2017). While many options were tested, the results

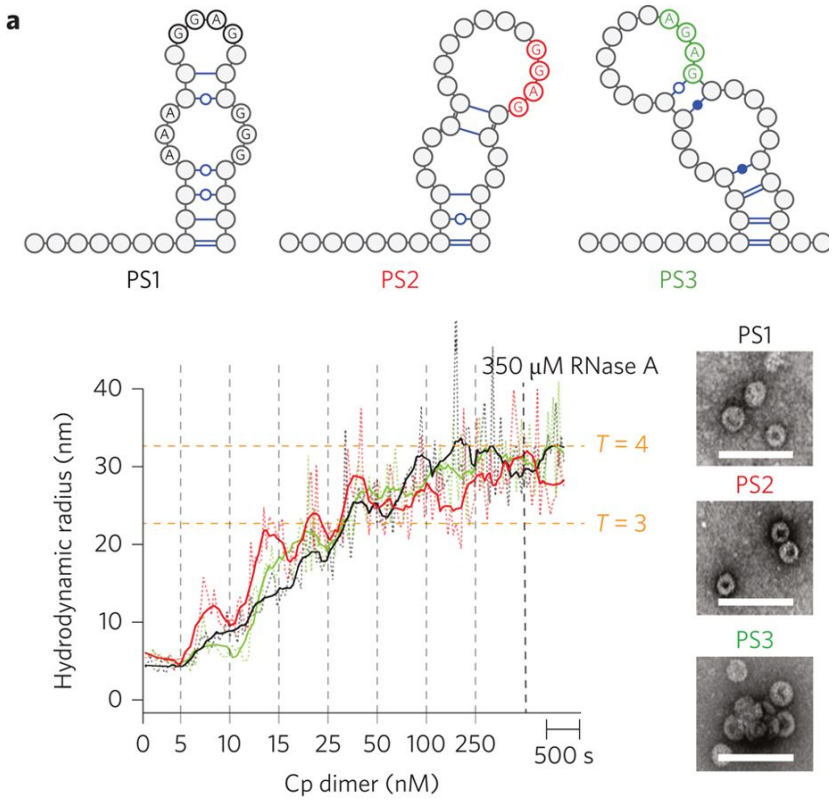

FIGURE 3.16: **Re-assembly of HBV capsid protein using single PS1, PS2, and PS3 stem-loops.** (a) Structures of stem-loops used in the experiments. The dye-labelled RNA oligos were exposed to increasing amounts of HBV capsid protein (Cp), each grey dotted line representing an addition of Cp. The black dotted line represents addition of RNAse A. The hydrodynamic radius was measured to determine capsid structure (left). Additionally, capsids were visualised by electron microscopy (right). Figure 3a from Patel et al. (2017).

of the mutation experiments technically only confirm that a combination of A and G is important in the apical loop. For instance, while the fact that neither GUUAGG nor UGGAUU produced intact capsids implies the necessity of Gs on both sides of the A, it does not exclude the possibility that UGGAUG would be functional. More extensive motif tests would be needed to properly confirm the exact variability of the PS motif.

TABLE 3.7: Effects of changes in PS1 apical loop and right-hand bulge on capsid re-assembly. Assembly behaviour in the re-assembly experiment is given as RNA-CP binding, capsid assembly, and RNAse A resistance. The table was copied from Supplementary Table 3 in Patel et al. (2017).

RNA Oligo	Loop	Bulge	Assembly behaviour	Comment
PS1	GGGAGG	GGG	+ + +	
L1	UUUAUU	GGG	+ – –	Loop Gs are important
L2	GUUAGG	GGG	+ – –	Loop Gs are important
L3	UGGAUU	GGG	+ + –	Loop Gs are important
L4	GGGUGG	GGG	+ + –	Loop A is important
L5	GGGGGG	GGG	+ + –	Loop A is important
B1	GGGAGG	AAC	+ + –	Bulge sequence / structure is important

3.6 Putative PSs in Foreign Sequences Used in Experiments

Up until now ε, which is located approximately 100 nucleotides downstream of PS1, was considered to be necessary and sufficient for packaging of HBV pgRNA and the only PS (Junker-Niepmann et al., 1990). This stands in contrast to our own findings described above. It is possible that the foreign sequences utilised in previous experiments actually formed some PS-like structures. To test this hypothesis I have analysed different sequences utilised in studies to substitute native sequence for the possibility that they express accidental PS-like SLs.

Different groups have tested different pgRNA mutants for their ability to be packaged. The mutants carried either just deletions or substitutions with foreign sequence to maintain RNA length in larger deletions. Presence of important viral proteins was ensured by supply through a helper plasmid that lacked the first 43 nucleotides of the pgRNA (no ε). Encapsidated RNA was detected with

a probe in either total cellular or capsid samples. In all experiments no or very little helper RNA was found in capsids indicating that deleting this part of the pgRNA renders it unable to package. The mutants were all packaged regardless of where the deletions were (Junker-Niepmann et al., 1990).

Junker-Niepmann et al. (1990) used one mutant where a large part of the 3' end encompassing PS1 was replaced with foreign sequence. This sequence was excised from pSV2CAT and is a part of SV40 containing the T antigen intron and poly-A signal. Another mutant had nucleotides 25 to 2778 in its sequence replaced by the *lacZ* gene while the 3' end still carried the SV40 fragment. To test whether these sequences could fold into PS-like SLs, Bernoulli scores were determined for first the *lacZ* gene and then the pSV2CAT fragment. Consistent with the HBV genome analysis, scores above 12 were plotted as before (Figure 3.17 and Figure 3.18). When aligning the naïve library to the sequences, no scores above 252 and 258 were found for *lacZ* and pSV2CAT, respectively, which is even lower than the highest noise level in the HBV genomes, which was 400. To be conservative the higher noise level was used as threshold for peaks. While the peaks were not as high as in native HBV genomic sequences, there were 14 peaks in *lacZ* and six in the pSV2CAT fragment clearly above noise level. The peak regions were analysed further as previously. In *lacZ*, six of the peak regions folded into an SL that presented an RGAG motif in the apical loop, while two more presented GGAUG and GGAA (Figure 3.19), which were also present in the HBV Bernoulli peak structures (see Figure 3.14). In the shorter pSV2CAT fragment, three of the six peak regions folded into RGAG SLs (Figure 3.20). These structures may have acted as PSs in the experiments by Junker-Niepmann et al. (1990), which would explain why encapsidation still occurred even when large parts of endogenous sequence were replaced.

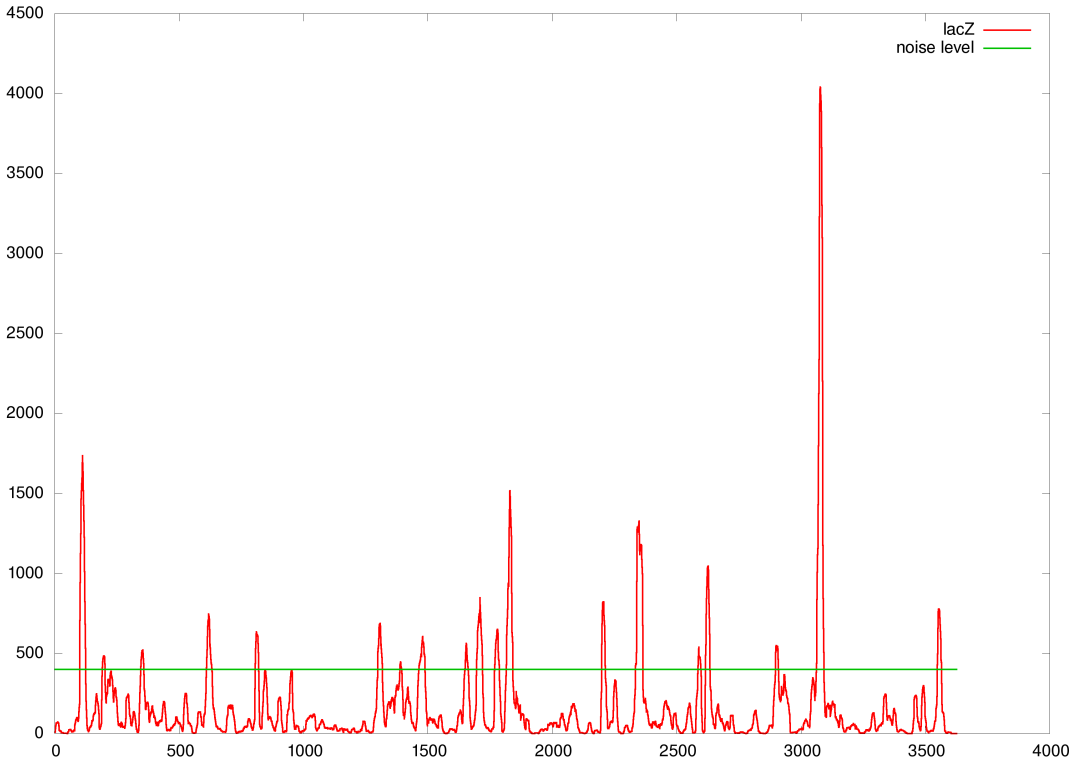

FIGURE 3.17: **Bernoulli peaks in lacZ gene.** The lacZ gene was processed as the HBV genomic sequences above to produce a plot with Bernoulli peaks. To be consistent and conservative the noise level applied here was the same as for the HBV genomes. 14 peaks above threshold were identified in the analysis.

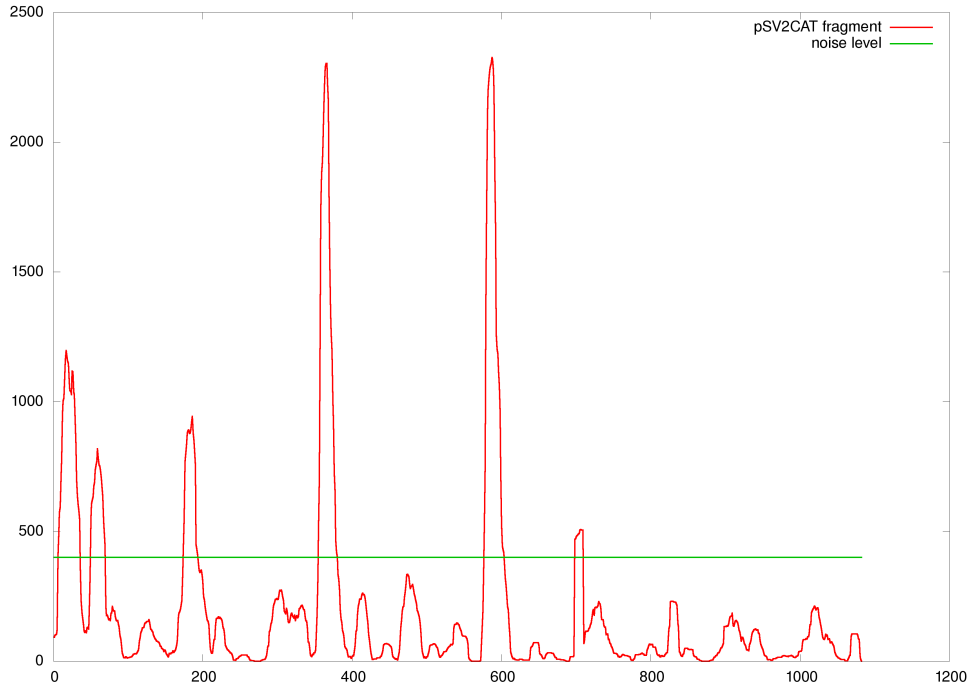

FIGURE 3.18: **Bernoulli peaks in pSV2CAT fragment.** The pSV2CAT fragment, which was used by Junker-Niepmann et al. (1990) in their experiments, was processed to produce Bernoulli peaks. To be consistent and conservative the noise level applied here was the same as for the HBV genomes. Six peaks were above the threshold.

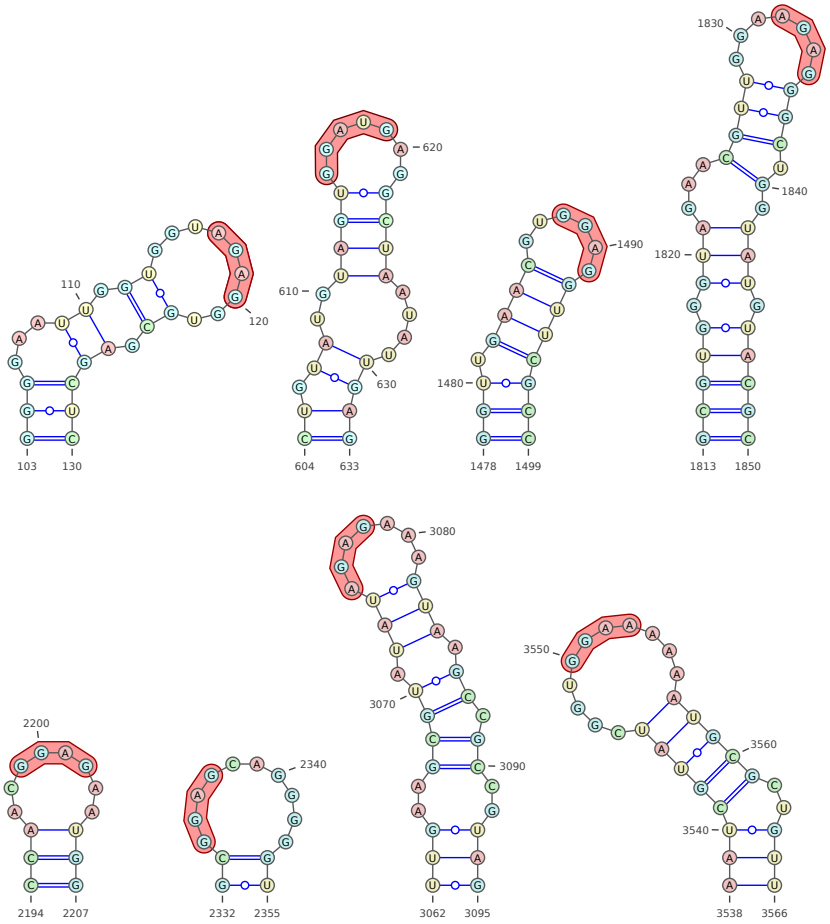

FIGURE 3.19: **PS-like structures as lacZ Bernoulli peaks.** Of the 14 Bernoulli peaks, eight regions were able to fold into a PS-like SL. Six displayed an RGAG motif. The other two showed GGAUG and GGAA, which were also seen in the HBV sequences. The structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

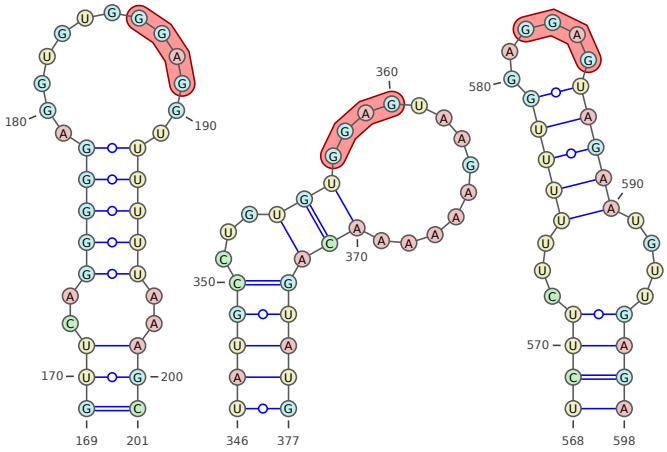

FIGURE 3.20: **PS-like structures at pSV2CAT Bernoulli peaks.** Of the six peaks, three could fold into RGAG SLs. The structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

3.7 Discussion

The question was asked whether a reverse transcribing DNA virus such as HBV would employ PSs-mediated assembly. Experimental collaborators performed experiments identifying RNA sequences that bind to the internal surface of HBV capsid protein HBcAg. Analysing these sequences by folding them into sets of suboptimal structures and identifying three and four nucleotide long enriched motifs, I found that they tended to fold into purine-rich SLs. Further, specific regions in HBV genomes with high similarity to those sequences were identified. Through folding those regions and comparing within and between genomes, a putative PS motif RGAG was found in the apical loop sequences. Comparing this with the apical loops present in top aptamer folds it could be seen that the motif may be too strict. While three of the predicted PSs were confirmed experimentally, the mutational experiments supported the motif but are not sufficient to fully confirm it. Rather than RGAG, they can only corroborate RRAR. In order to properly confirm that the PS motif is in fact as specific as RGAG, a further series of re-assembly experiments would be necessary. Suggested loop constructs and predicted assembly behaviours if RGAG is correct are summarised in Table 3.8.

Table 3.8: Proposed loop motifs to test experimentally and their predicted assembly behaviour in RNA-CP binding, capsid assembly, and RNAse A resistance.

Loop	**Predicted assembly behaviour**
UGGAGU	+ + +
UAGAGU	+ + +
UAAAGU	+ + −
UGAAGU	+ + −
UAAAAU	+ − −

When first described ε was said to be necessary and sufficient for specific packaging of the viral pgRNA (Junker-Niepmann et al., 1990). In deletion experiments large parts of the genome (nucleotides 765–2654, nucleotides 84–937) were found to be completely dispensable for packaging. Constructs with deletions or substitutions of these regions for unrelated RNA were detected in the viral capsid. This and the fact that a helper genome containing the important proteins for encapsidation but lacking the first 43 nucleotides, which include ε, was not encapsidated led them to conclude that the important signal was located at the 5' end of the pgRNA. Furthermore, it is even possible to package foreign RNA into viral cores by adding the 137 nucleotide sequence between nucleotides 3134 and 88 of the pgRNA, to a gene of similar length to the HBV genome. Note, that in this study a different genomic numbering is used than is now convention. There, the pgRNA starts at position 3100, which is at 1818 in conventional numbering. The total length of this strain was 3182. Thus, the respective regions would be $(1818 + 34 =)1852$ to $(1818 + 88 + 82 =)1988$. ε is situated at 1847 to 1907, whereas PS1 is 1722 to 1747, and thus not a part of this fragment. Whilst these positions are not expected to be fully accurate, they still show the general region where this fragment was situated. Biological importance was further illustrated by the high degree of conservation in this region between different mammalian *Hepadnaviruses* (Junker-Niepmann et al., 1990). Their results led them to conclude that there is only this one packaging signal in HBV.

Work by Bartenschlager et al. (1990) also highlighted the importance of Pol and showed that certain mutations in the protein resulted in ablation of pgRNA packaging. Similar to Junker-Niepmann et al. (1990) they also used a helper construct lacking ε to provide a functional copy of Pol in *trans*, which was not encapsidated, and a construct containing ε added upstream of *lacZ*, which was encapsidated when Pol was functional (Bartenschlager et al., 1990). They later also showed the interdependency of Pol and pgRNA for encapsidation as neither is packaged without a functional interaction (Bartenschlager and Schaller, 1992).

Further mutational experiments on ε performed by Pollack and Ganem (1993) indicated the necessity of this *cis*-acting element for pgRNA packaging. In all these experiment *lacZ* mRNA could be packaged in to capsids when wildtype ε was upstream on the same RNA and Pol present. Fallows and Goff (1995) investigated the effects of mutations in different parts of the ε SL on RNA packaging and replication using HBV pgRNA constructs. Their results are consistent with work done later by Hu and Boyer (2006) on which changes to the SL would ablate binding to Pol further supporting the idea that Pol-ε interaction is necessary for pgRNA packaging.

In addition to the argument about necessity for packaging, some work has suggested that the interaction between Pol and ε ensures that pgRNA is packaged specifically over other RNAs. *In vitro* reassembly experiments using purified capsid protein and different RNAs but lacking Pol Porterfield et al. (2010) found no specificity of HBcAg for pgRNA as it also packaged other RNAs such as LacZ, CCMV RNA1, and *Xenopus* elongation factor RNA. Even in competition assays pgRNA had no advantage LacZ RNA indicating that specificity is conferred through Pol binding to ε (Porterfield et al., 2010).

The studies mentioned above are just some examples over the years that have shown the importance of ε and Pol. A recent review about HBV packaging still explains pgRNA packaging through this interaction (Selzer and Zlotnick, 2015). This stands in contrast to our own findings in this chapter and in Patel et al. (2017). As summarised above we were able to identify a number of SLs in the HBV genome that act as PSs and trigger re-assembly of capsid protein *in vitro*. This begs the question how our findings can be consolidated with previous work.

PSs are by design highly variable. Even without taking into account affinity related variations, the motifs are so short and vague that it is possible for them to occur at random in foreign sequence. The virus would have evolved to not package RNA from its host but there is no selective pressure against *Escherichia coli (E. coli)* mRNAs such as *lacZ*, which was used as substitute in the exper-

iments by several groups (Junker-Niepmann et al., 1990; Bartenschlager et al., 1990; Bartenschlager and Schaller, 1992; Pollack and Ganem, 1993). To test that hypothesis, the substitute sequences from Junker-Niepmann et al. (1990) were analysed for PS-like SLs. Interestingly, both sequences could form at least some SLs that presented the identified RGAG motif and might thus have taken on the roll of PSs in the experiment. Also CCMV RNA1 and *Xenopus* elongation factor RNA as used in Porterfield et al. (2010) can fold into such SLs (data not shown).

Apart from our own results, there were also other studies that cast doubt on the role of ε as sole PS. In experiments by Hatton *et al.* encapsidation was achieved even when only using the core protein gene under control of a foreign promoter cloned into *E. coli*. ε, present in the precore region of the HBV genome, was not included and seemed to not be necessary here for core proteins to assemble and encapsidate RNA (Hatton et al., 1992). This indicates that RNA and capsid protein subunits interact with each other independent of the Pol-ε interaction and also that this is important for packaging. Why packaging was possible in absence of ε, despite having been deemed necessary for this role by Junker-Niepmann *et al.* is unknown.

The case for more PSs spread over the genome is further supported by the cryo-electron microscopy structure of HBV capsids by Crowther *et al.* (see Figure 3.5). Capsids are able to form with or without packaged RNA when expressed in *E. coli*. Capsid proteins that have a deletion in the carboxy-terminal region rendering them unable to bind nucleic acid resulting in empty shells were still able to assemble into the same structures as wild type proteins. The structure of capsids was observed as either T=3 or T=4 icosahedral symmetry with 180 or 240 capsid protein units, respectively. Interestingly, below the shell structure, the authors observed that some wild type capsids contained an inner shell that was also icosahedrally ordered and concluded that this is probably RNA. This inner shell seemed to have some loose contacts with the outer capsids, which may indicate PS contacts between RNA and capsid protein. The fact that the inner structure also

followed icosahedral symmetry indicated further that there are regular contacts with capsid protein at several points of the RNA as opposed to just one (Crowther et al., 1994).

Recently, an asymmetric cryo-EM structure of HBV capsids was resolved (Wang et al., 2014). The lack of icosahedral averaging allowed the visualisation of the actual organisation of the RNA and the location of proteins inside the capsid. Interestingly, the distribution of RNA is highly similar across capsids with a density corresponding to Pol in a particular position. The regular organisation of the pgRNA supports the idea on a regulated assembly process along a Hamiltonian path similar to what has been observed in some ssRNA viruses that utilise PS-mediated assembly (Dykeman et al., 2013b; Dai et al., 2017). The PSs may thus aid in regulating the path of assembly. A regular RNA organisation may in turn assist the movement of Pol along the pgRNA for DNA synthesis in HBV (Wang et al., 2014).

Nevertheless, ε appears to be necessary for packaging and assembly *in vivo*. In experiments where it was deleted, packaging did not occur. This begs the question whether ε serves another function that is necessary for the genome to be packaged *in vivo* but may be dispensable *in vitro* or in a bacterial expression system. One possibility is that ε plays an essential role in switching off translation of the viral mRNA so that it can serve as pgRNA.

Translation needs to be switched off before the RNA can be used for genome replication. This mechanism is common among viruses and found for example in coliphages of the *Leviviridae* family (Kolakofsky and Weissmann, 1971a,b). The viral polymerase competes with the ribosomes and once it is bound, prevents further binding of new ribosomes. In eukaryotic cells translation occurs through binding of the small ribosomal subunit to initiation factors associated to the 5' cap. It then scans the mRNA for a AUG start codon within a Kozak sequence (Pestova and Kolupaeva, 2002). Once a suitable AUG is encountered, the large ribosomal subunit is recruited and translation is initiated. Due to this mechanism

protein synthesis usually commences at the first AUG. However, translation from a later start codon is possible through amongst others leaky scanning, which entails the small ribosomal subunit "missing" an AUG and continuing its search (Kozak, 1999, 2002). Since the HBcAg ORF precedes Pol, HBcAg is synthesised in larger amounts. Translation of Pol occurs sporadically at a lower rate. When Pol is synthesised it binds to ε at the 5' end of the mRNA, which spans the start codon for HBcAg. Binding of Pol stabilises the SL and thereby inhibits translation of both ORFs (Ryu et al., 2008).

In Table 3.9 experimental conditions from a selection of studies and their effects on RNA encapsidation are summarised. The following alternative explanations are possible for the respective observations:

1. The SV40 sequence part used to substitute the 3' terminal genomic sequence has potential to form RGAG-presenting SLs. These could substitute for PS1 (see section 3.6 "Putative PSs in Foreign Sequences").

2. Without ε Pol cannot inhibit translation; thus, the pgRNA is not free to form PSs and get encapsidated (eIF4A is a helicase, which removes secondary structures on the mRNA).

3. Mutations in ε inhibit the binding of Pol as seen in the work by Hu and Boyer (2006).

4. Also *lacZ* has the potential to form PS-like SLs (see Chapter 3.6 "Putative PSs in Foreign Sequences").

5. Mutations in Pol that interfere with its binding to ε would also prevent it from switching off translation.

6. Translation initiation occurs at the 5' cap. Pol interacts with it via eIF4E (Kim et al., 2010). If the distance is too large between ε and 5' cap, Pol cannot interact with both to switch off translation. This shows that just

TABLE 3.9: Some experimental conditions and effects on RNA encapsidation. Whether or not pgRNA was found in assembled capsids is indicated by + or -, respectively.

	Condition	**RNA in capsid**	**Reference**
1	HBV genome with large substitution at 3' with SV40	+	Junker-Niepmann et al. (1990)
2	HBV genome without ε	-	Junker-Niepmann et al. (1990)
3	HBV genome with mutated ε	-	Fallows and Goff (1995); Hu and Boyer (2006)
4	ε with *lacZ*	+	Junker-Niepmann et al. (1990); Bartenschlager et al. (1990); Bartenschlager and Schaller (1992); Pollack and Ganem (1993)
5	Mutations in Pol	-	Bartenschlager et al. (1990)
6	Increased distance 5' cap and ε	-	Jeong et al. (2000)
7	pgRNA without 5' cap	-	Jeong et al. (2000)
8	Capsid expressed in *E. coli* (no ε)	+	Birnbaum and Nassal (1990); Crowther et al. (1994)

the presence of ε and Pol is not enough to package - interaction with 5' cap (potentially to inhibit translation first) is necessary.

7. The system used was developed as a good expression system for genes in eukaryotes and translation does happen to a sufficient degree despite the lack of a 5' cap (Fuerst et al., 1986). But without it, Pol cannot switch it off.

8. In the absence of pgRNA and other HBV proteins, capsid proteins self-assemble in an *E. coli* expression system. They form T=4 capsids with inner density corresponding to *E. coli* RNA. In these systems neither ε nor Pol are present. Nevertheless assembled capsids contain RNA if the arginine-rich carboxy-terminal tail is not truncated. Above I have given some examples of foreign (prokaryotic) mRNAs that contain PS-like SLs. It is conceivable that some mRNAs in *E. coli* do so well enough to trigger assembly. For example, the Shine-Dalgarno (SD) sequence, the ribosomal binding site in prokaryotes, is AGGAGGU in *E. coli*. Translational regulation through secondary structures is not uncommon in prokaryotes since they do not employ a helicase during translation. It is thus thinkable that in some mRNAs the SD sequence can be the apical loop of a SL presenting a RGAG to the overexpressed HBcAg. Binding of HBcAg would further stabilise the structure making the SD inaccessible for ribosomes and thus switching off translation. If more PS-like SLs are present on that mRNA, packaging may occur. Moreover, in Birnbaum and Nassal (1990) it was found that the particles preferentially packaged the core protein mRNA over cellular RNA. Note that SL4 and SL5 are within the core protein gene.

To summarise, these studies can support the hypothesis that ε is essential for packaging because it is involved in switching off translation, which frees the pgRNA from ribosomes and allows for (other) PSs to form. In prokaryotic expression systems neither ε nor Pol are necessary for the packaging of RNA if the

nucleic acid-binding carboxy-terminal tail is not truncated. Nevertheless, HBcAg displays specificity for its own mRNA, which contains SL4 and SL5. This may be due to the difference in translational mechanisms between prokaryotes and eukaryotes making ε dispensable in *E. coli*. The interaction between Pol and ε can in this way confer packaging specificity by only the pgRNA becoming free of ribosomes and available for packaging. Without this interaction there would be no direct advantage of the pgRNA over other RNAs that contain PS-like structures as seen in the work by Porterfield et al. (2010). Whether this is indeed the case would have to be investigated experimentally.

In addition to PS-HBcAg and Pol-ε interactions also HBcAg phosphorylation plays an important role in ensuring that the correct RNA is packaged. A fine balance of charges is needed between the positively charged arginine residues, the double negatively charged phosphorylated serine residues on the carboxy-terminal domain of the HBcAg protein and the negatively charged nucleic acids inside the capsid (Le Pogam et al., 2005; Lewellyn and Loeb, 2011). Mimicking (de)phosphorylation through mutation of the serine residues in the carboxy-terminal domain to uncharged or negatively charged amino acids can dramatically reduce packaging efficiency (Gazina et al., 2000) or result in mispackaging of shorter, spliced pgRNA (Köck et al., 2004). In a cryo-EM structure of capsids assembled from HBcAg in different mimicked phosphorylation states, the conformation of the carboxy-terminal domain and the organisation of the RNA inside the capsids was markedly different under different conditions (Wang et al., 2012). These studies support the idea that in addition to RNA-protein interactions that drive specific pgRNA packaging there is also an element of electrostatic interactions ensuring the correct size of RNA is packaged.

Interplay between RNA and proteins for capsid assembly, RNA packaging, and DNA synthesis in HBV is complex involving probably a finely balanced combination of HBcAg phosphorylation states, interactions between HBcAg and Pol, between Pol and the pgRNA through ε, and between HBcAg and the pgRNA

through PSs. The additional component of Pol in this balance places HBV in a different position compared to ssRNA viruses studied before for PS-mediated assembly. The precise PS motif is likely unique to this viral species; however, the more complex interactions may be applicable to similar viruses at least in the same viral genus of *Orthohepadnaviruses* if not the entire family of *Hepadnaviridae*. Whilst more research would have to be done to fully understand this interplay and how or if it is applicable to other reverse-transcribing viruses, the identification of PSs in HBV opens up the possibility of a new drug target for this particular virus. Currently, the most commonly used antiviral drugs against HBV are nucleotide analogues and interferon, which rarely achieve cure of the infection and stopping treatment often results in relapse (Wu et al., 2019). To date, several potential drugs targeting the Pol-ε interaction have been tested for their antiviral effect (Lin and Hu, 2008; Feng et al., 2011; Jo et al., 2020). The success of these to inhibit the virus gives hope that also PS-HBcAg interactions could be used successfully as a drug target. The wider the range of possible tools to fight the infection, the higher the chance to successfully cure more patients in the future.

Chapter 4

Application of Phylogeny to HBV

In Chapter 2 "Phylogenetic Algorithms" I introduced a novel method for reconstructing phylogenetic trees for viruses assembling via a PS-mediated assembly mechanism. Instead of using multiple sequence alignments (MSAs) of the genomic sequences, these phylogenies are based on the PS profiles of the viruses, i.e. the distribution of PSs in their RNA genomes. This takes into account not only primary structure, the sequence, but also secondary structure, the SLs, and their function, the PSs. Here, this method is applied to a number of different HBV data sets starting with a generic set of (sub)genotypes and then looking at four different longitudinal and regional studies. This inclusion of different levels of relatedness between the study strains as well as the number of study strains versus the number of reference strains also provided the opportunity to showcase the importance of setting the right conservation threshold for PS blocks. The hope was to find a different level of resolution in the phylogeny using this new method.

4.1 Evolution and Origin of Hepatitis B Virus

Hepadnaviridae infect a large variety of hosts. Apart from humans, the mammal infecting genus *Orthohepadnaviruses* of this family is common in many rodents

such as woodchuck or groundsquirrel and in different types of bats. Given a PS motif for each of these species, this would provide an opportunity to compare the phylogenetic approach based on PSs at different scales: human hosts and between human and other species. The bird-infecting other genus *Avihepadnaviruses* on the other hand is found in an array of different bird species such as duck, heron, or parrot. This wide array of hosts indicates a long evolutionary history and the origin of HBV infection in humans is to this day controversial. Three main hypotheses are currently considered, all with their own merits and problems. The oldest hypothesis places the origin of HBV infection in the New World (Bollyky et al., 1998). This idea is based on their phylogenetic work that concludes a substitution rate placing the HBV most recent common ancestor (MRCA) no more than 1000 years ago. According to Bollyky et al. (1998) it is the result of a zoonotic infection from rodents to humans. From the, which Americas it is hypothesised to have spread to Europe and further, following colonization in accordance with diversification between New and Old World viruses around 400 years ago. The fact that there are HBV strains specific to different species of Old World non-human primates, such as chimpanzees, contradicts this hypothesis (MacDonald et al., 2000). Instead, it points towards an infection of these apes long enough ago to allow for adaptation, spread, and evolution within these hosts. MacDonald et al. (2000) therefore hypothesised that rather than there having been one transmission event between humans and non-human primates, there were in fact several from different primate species. This would also explain why HBV from different non-human primate species do not cluster together, but cluster close to different human viral genotypes on phylogenetic trees. However, this would imply this evolutionary process to have taken place over the excessive time frame of over 10 million years. This is in conflict with predicted evolutionary rates of HBV, which have been found to be much faster (Zehender et al., 2014). The third hypothesis considers HBV to have infected humans since the emergence of anatomical humans. Rather than having acquired the virus through

different events of zoonotic infections or through fairly recent horizontal infection, the world-wide distribution of the virus and its diversification is thought to be the result of migration and evolution over 10,000 years (Zehender et al., 2014).

Having a method for phylogeny that can resolve a different evolutionary time scale such as the one presented here, may provide new insight into this question. Given a PS motif for all *Hepadnaviridae* species, a complete phylogeny could be reconstructed for this viral family. This may yield a better understanding of the relatedness between the species and how they evolved from each other. However, at the moment only the motif for HBV has been discovered so that focus will be on the evolution of PSs within this species.

4.2 Recombination in Hepatitis B Virus

4.2.1 Circulating and Sporadic Recombinants

Recombination is a common method in nature to increase diversity in a population or an individual. Many viruses employ intergenomic recombination with other strains or subgroups to gain competitive advantages. Many different recombinants of HBV have been isolated and described and can be divided into two categories: circulating and sporadic. A circulating recombinant would have some advantage over the "pure" genotypes or be at least competitive, whereas a sporadic one would form in one individual but not spread to others. All recombination breakpoints in HBV reported in the literature are visualised in Figure 4.1 (top). Figure 4.1 (bottom) focuses on circulating recombinants only. There are at least six recognized circulating types: one A/D, one A/E, one B/C, one C/B and two C/D recombinants. The A/D recombinant was isolated initially in four people in India (Simmonds and Midgley, 2005) and later confirmed with another eight isolates also from India (Yang et al., 2006; Shi et al., 2012a). The breakpoints are around nt 1808 (A $\longrightarrow$ D) and nt 2354 (D $\longrightarrow$ A). The A/E recombinant has been isolated in four people in Guinea and had its breakpoints around nt 1896–

1906 (A $\longrightarrow$ E) and nt 2419–2423 (E $\longrightarrow$ A) affecting the C gene (Garmiri et al., 2009). The B/C recombinant is so successful that it is the only form of genotype B circulating in continental Asia. Pure B is only found in Japan. The presence of several subtypes of this "BC" genotype indicates a recombination event that took place a long time ago with one recombinant that then spread and diversified over time. The breakpoints of this are around nt 1740–1838 (B $\longrightarrow$ C) and nt 2443–2485 (C $\longrightarrow$ B) encompassing the preC and C gene, which is switched from B to C (Bowyer and Sim, 2000; Simmonds and Midgley, 2005; Fares and Holmes, 2002; Sugauchi et al., 2002, 2003; Ye et al., 2010; Shi et al., 2012a). The C/B recombinant, on the other hand, was only isolated from 18 Chinese patients and is not as common. Its breakpoints are at nt 2276 (C $\longrightarrow$ B) and nt 224 (B $\longrightarrow$ C) (note the circular nature of the HBV genome) encompassing a small portion of C overlapping with P and a large part of P and S (Shi et al., 2012b). The two C/D recombinants CD1 and CD2 are both found in western China, especially Tibet. CD1 and CD2 share the 5' breakpoint at nt 10 (C $\longrightarrow$ D) but CD1 switches back at nt 799 (D $\longrightarrow$ C), while CD2 has the 3' break point at nt 1499 (D $\longrightarrow$ C) (Cui et al., 2002; Wang et al., 2007; Simmonds and Midgley, 2005; Yang et al., 2006; Zhou et al., 2011; Shi et al., 2012a). This means that for CD1, only the S gene is affected and for CD2 also a large part of P and some of X. Sporadic recombinations have been described between A/C, A/D, A/E, A/F, B/A, B/A/C, B/C/A, C/A, C/B, C/G, D/A, F/C, D/E, D/F, E/A, E/D, F/A, F/G, G/A, G/C, and G/F, whereby the major genotype is named first (Hannoun et al., 2000; Shi et al., 2012a; Bollyky et al., 1996; Simmonds and Midgley, 2005; Mizokami et al., 1997; Bowyer and Sim, 2000; Owiredu et al., 2001a; Fares and Holmes, 2002; Yang et al., 2006; Chauhan et al., 2008; Ye et al., 2010; Kurbanov et al., 2005; Garmiri et al., 2009; Lopez et al., 2015; Kato et al., 2002; Sugauchi et al., 2002, 2003; Luo et al., 2004; Mulyanto et al., 2012; Cui et al., 2002; Suwannakarn, 2005; Huy et al., 2008; Laoi and Crowley, 2008; Chekaraou et al., 2010; Fallot et al., 2012; Araujo et al., 2013). Yang et al have summarised different kinds of recombinants

and concluded that genotypes A and C have the highest propensity for recombination and most combinations are in fact A/D or B/C (Yang et al., 2006). A summary of published recombinants can be found in Appendix A Table A.2.

4.2.2 Hot Spots of Recombination

Recombination breakpoints are spread across the genome. There are, however, some "hot spots" for recombination shown and corroborated in several publications over the years. Otherwise, breakpoints tend to be at the boundaries of genes, most commonly C and S (Simmonds and Midgley, 2005). Yang et al mapped the boundaries of several types of recombinants and summarised the hot spots for break points. Of the 25 different types, 60% had a break point near DR1, often close to the 5' end of the PreCore/Core gene (nucleotides 1640–1900), making this the most common point of recombination. It is often associated with a second breakpoint at the 3' end of the Core gene (nucleotides 2330–2485), identified in 32% of recombinants. Another common point is around preS1/S2 (nucleotides 3150–10), which was identified in 28% (Yang et al., 2006).

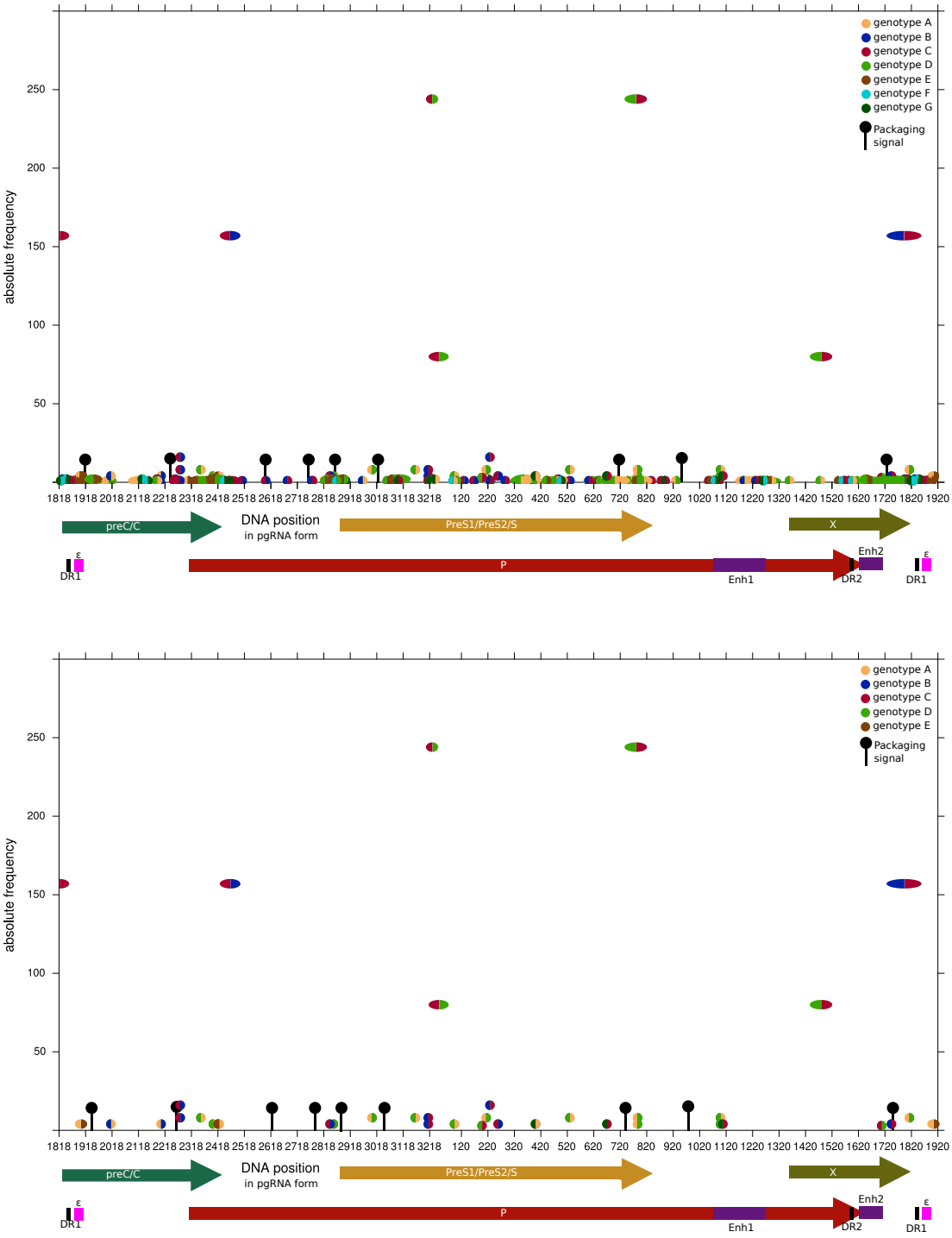

FIGURE 4.1: **Recombination events between different HBV genotypes.** Genomic positions of break points are rearranged into pgRNA format causing points between nt 1818 and 1920 to occur double. Colours of data points indicate the genotype left/right of the break point. Elongated data points indicate uncertainty in exact break point position. Absolute frequencies refer to the number of published isolates. Packaging signal positions are indicated by black lollipops. Top: all published recombinants. Bottom: only circulating recombinants (absolute frequency ≥ 2).

4.3 Methods

In Chapter 2 different approaches to phylogeny were described. All rely on some "characters", an attribute by which the taxa are differing. For viruses phylogeny is mostly based on MSA so the characters are amino acid or genomic sequence positions. Bamford and Stuart introduced a novel approach to grouping viruses based on the structure of their capsid proteins. The characters were based on the distances of the folds in 3D. Due to the slower evolving nature of a protein structure similarities could be observed between viruses formerly thought to be unrelated. This showcases the power of using different features of viruses as characters for phylogeny. It allows resolution of evolutionary relationships on vastly different scales. To add to this idea, I developed a new approach that utilised partially conserved PS positions as characters. This combines RNA structure, an SL, with function, a PS, for comparing viral strains. Here, this method is adapted for HBV and applied to different datasets.

4.3.1 Conversion to pgRNA Form

HBV has a circular DNA genome and while there is a convention for start and end when submitting a sequence to the NCBI database, sequence boundaries are often not uniform. This has implications for phylogenetics, since the ends cannot be aligned. To avoid this pitfall and ensure that all sequences used had uniform start and end sites, the sequences were all first converted into pgRNA form. Additionally, the conversion ensured that no SLs would be missed due to their occurring at the boundaries of the genomic sequence as supplied in the database.

The converted sequences started at the TATA box and ended at the poly-A signal. While these are not technically the true end points of the pgRNA, they are the start and end points given in the literature for HBV and were thus used for the conversion. The TATA box acts as promoter in eukaryotic cells recruiting

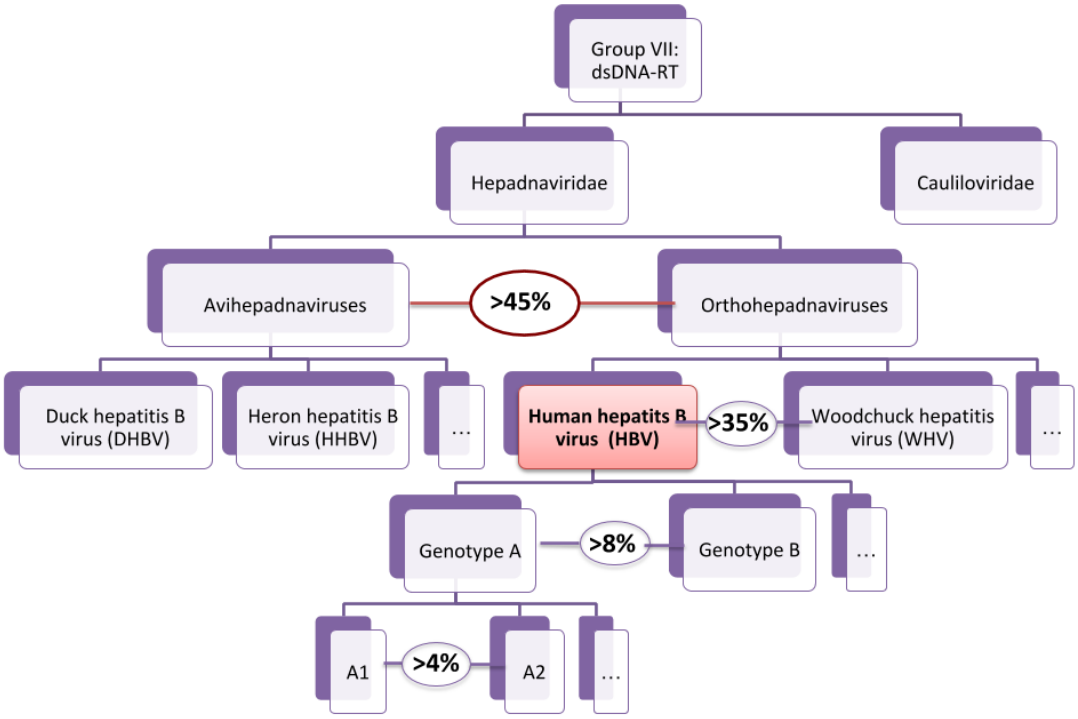

FIGURE 4.2: **HBV viral classification.** The viral classification starts at the species level and becomes more general. For HBV, strains are further grouped into genotypes and subgenotypes depending on their level of genomic sequence identity or rather the opposite, i.e. genomic difference. Subgenotypes are at least 4% but less than 8% different from each other, whereas genotypes differ by least 8%. Above species are the genera. HBV belongs to the genus *Orthohepadnavirus* together with some other mammalian viruses such as WHV, whose genome is already more than 35% different. Beyond genus is the viral family, in this case *Hepadnaviridae*. This also encompasses *Avihepadnaviruses*, which include, among others, DHBV and HHBV. The genomic differences between species of these two genera of the same family are at almost 50%.

transcription factors and thus RNA polymerase II. The published sequence for HBV is CATAAATT (Quarleri, 2014). Despite the actual RNA start site being situated approximately 30 nucleotides downstream, the TATA box was utilised as start site for the pgRNA here as it is easily identifiable. The poly-A signal used in HBV is TATAAA (Simonsen and Levinson, 1983) and thus differs slightly from the common AATAAA found in mammals (Levitt et al., 1989). TATAAA was utilised as the pgRNA end.

The conversion program takes as input a sequence in FASTA format, approximate start and end positions, the reference start and end sequences, and a prefix for the output. Given this information, the sequence is read in and searched for the reference sequences -100 to +1000 of the approximate positions. This reduced the amount of sequence that had to be searched, and thus sped up the program. Once found, start and end positions are calculated. The sequence is then printed starting from the start position until the end of the original sequence and from the beginning of the original sequence until the end position. The output is in FASTA format.

4.3.2 RNA Folding

Once converted into pgRNA form, the sequences were processed and folded as described in Chapter 2 Sections 2.4.1.1 and 2.4.1.2. Briefly, first the sequences were fragmented into highly overlapping 30 nucleotide windows sliding by 1 nucleotide. Each fragment was then folded in Tfold with use of the partition function and sampling 10,000 times for each frame. The folds were then processed as described in Section 2.4.1.3. Single SLs were extracted from each fold and the number of times they occurred among the 10,000 sample folds was recorded. The latter could be utilised as a proxy measure of stability as a more stable structure is more likely to be sampled from the partition function. Next, the structures were merged across windows, i.e. each unique SL was retained together with its summative number of folds across windows.

4.3.3 PS Phylogeny

The PS phylogeny is based on PS profiles, i.e. pseudosequences where each nucleotide position is encoded as a string indicating whether the virus has a PS at that genomic position or not. The selection of SLs and generation of profiles from these is described in detail in Sections 2.4.1.4–2.4.1.6. Briefly, a set of non-overlapping SLs was found by optimising the overall additive SL energies, i.e. finding a set of SLs that, when their respective energies are added together, result in the lowest summative energy (see Section 2.4.1.4). In addition to the respective folding energies, these also added the energies for PS affinities for CP. Since specific affinities and tiers are not known in HBV at the moment, only one tier was used, whose K_D was set to 15 nM. This corresponds to the first affinity tier in MS2 with exception of TR, which has a K_D of 1.5 nM. As such a biologically known high PS affinity was utilised, whilst being conservative in not using the highest. The SL selection results were compared using K_D of 1.5, 5, 10, and 20 nM. The same SLs were consistently selected for all conditions except for the lowest, 1.5 nM where two SLs were selected differently (data not shown). This showed that the results are robust in the vicinity of the affinity used here and using the lower TR affinity would promote some unstable SLs. The PSs were identified using the search motif `X{2}__X*RGAGX*__X{2}`, i.e. at least two base-pairs surrounding the RGAG containing apical loop. Additionally, ε was utilised as anchor structure to more easily compare sequences of variable lengths. The advantage of using ε is not only that it is a very stable SL known to occur in a particular form in every strain but also that it occurs in the repeated region; thus, in pgRNA form, it provides an anchor on both ends. The search motif for this was `AG__CTGTGC__CT`. Based on this set of SLs PS profiles were generated by converting every nucleotide position that was part of a PS into a "C", the anchor positions into "T", and every other position into "A". Next, the profiles were aligned with each other utilising the MSA of the pgRNA sequences. Going through the alignments nucleotide position by nucleotide position, when more

than a preset threshold percentage of sequences in the set had a PS, a PS block was started or continued. The threshold depended on the data set. It needs to be low enough to allow resolution of groups whilst also being high enough to reduce noise. Which level is required for resolution depends on the relatedness of the strains and the number of study versus reference sequences used. Finally, going through the blocks again, it was checked for each sequence, whether it had a PS there or not. If yes, a "C" was assigned for that block to that sequence, if not then an "A". These block profiles were then supplied to SplitsTree 4 (Huson and Bryant, 2006) and neighbor-joining (NJ) trees using Hamming distances constructed.

4.4 Phylogenetic Trees of Genotypes and Subgenotypes

Having identified the likely PS motif in HBV as RGAG in Chapter 3, the method developed in Chapter 2 to utilise these PSs for phylogeny was put to the test. The first set of genomic sequences used consisted of representative strains for most common genotypes and subgenotypes of HBV as well as some common A/D, B/C and C/D recombinants. The aliases used in the phylogenetic trees and the respective accession numbers and countries of origin are summarised in Appendix A Table A.3. For the recombinant strains the alias consists of the genotypes they are made up of, the major genotype first, followed by the genomic location of the recombination. The aliases of the other strains are simply their subgenotype. When more than one representative was used, they were identified by numbers after an underscore.

The sequences were converted to pgRNA form and folded in overlapping windows of 30 nucleotides. PS profiles were generated using `X{2}__X*RGAGX*__X{2}` as search motif, which represents any SL with at least two base-pairs and RGAG anywhere in an apical loop of any total size. All possible thresholds from 1%

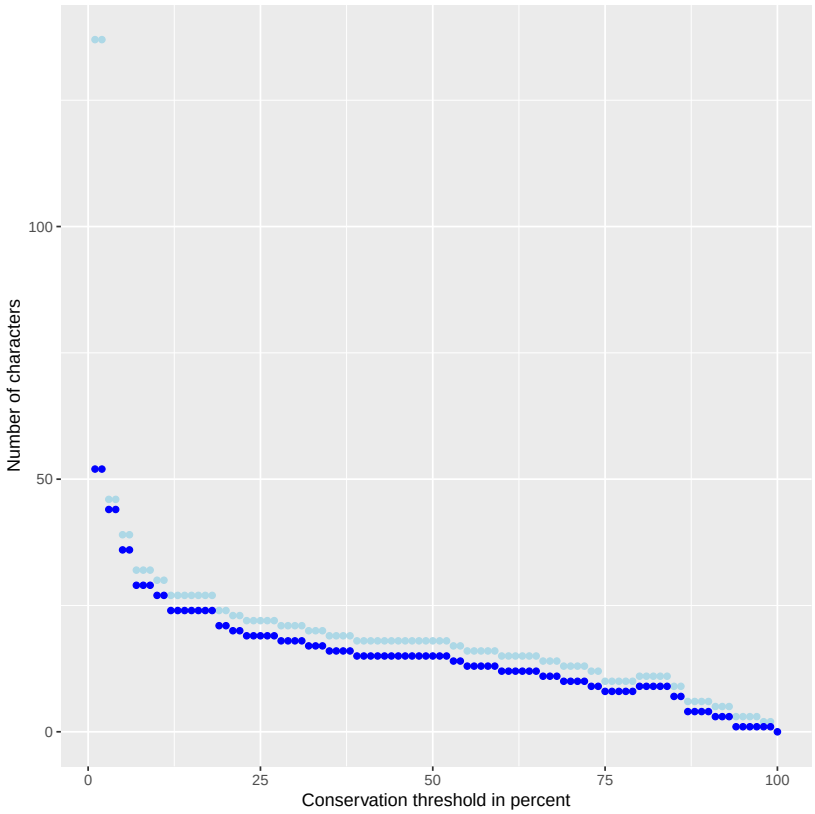

FIGURE 4.3: **Impact of conservation threshold on number of characters.** Conservation thresholds from 1% to 100% were tested for the genotype comparison. The number of total characters, i.e. blocks, (light blue) and informative characters (blue) they resulted in are shown.

to 100% were tested to find a suitable value for this data set (Figure 4.3). The default threshold of 50% was used for creating the conserved PS blocks because the number of blocks was stable around that value. A PS, thus, had to occur in at least half of the strains at a certain position for a PS block to be started or continued. This resulted in eleven blocks for this set of sequences of which nine were informative, i.e. not the same among all sequences (Figure 4.4). The blocks were situated in nucleotide positions 170–185, 263–286, 473–502, 867–880, 1028–1047, 1099–1113, 1190–1214, 1273–1296, 1319-1342, 1433-1457, 1498-1518, 1619–1637, 1707–1731, 2107–2120, 3074–3093, 3112–3139, 3198–3211, and 3212–3259. Note that these positions refer to aligned pgRNA form. For example, PS1, PS2, and PS3 are located in blocks 3198–3211, 867–880, and 1028–1047, respectively. Interestingly, only the PS1 block position is highly conserved between all

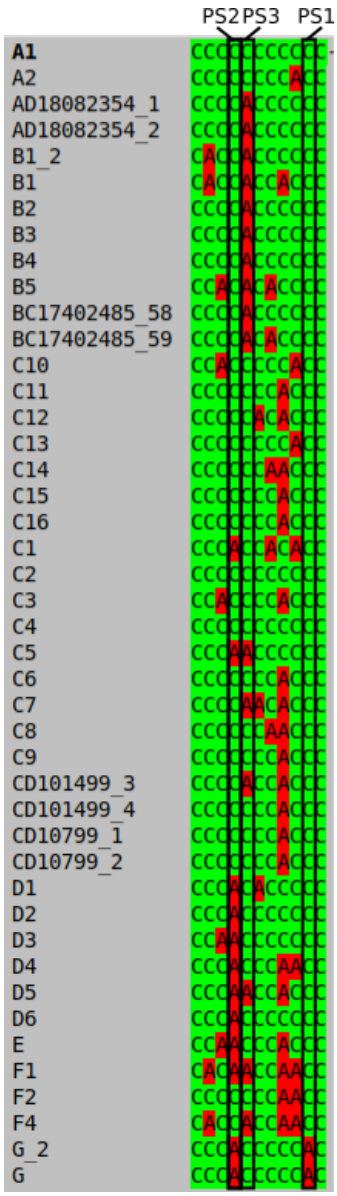

FIGURE 4.4: **Packaging signal blocks in HBV genotypes.** The blocks were created using a conservation threshold of 50%. "C" (green) represents having a PS in that block and "A" (red) means no PS. The blocks where PS1, PS2, and PS3 are located are indicated. They are visualised in SeaView (Gouy et al., 2010).

genotypes except G, whereas PS2 is absent in D, E, some F, and G strains and no PS3 is present in A/D recombinant, B, B/C recombinant, and some F strains.

The PS blocks were input into tree-building program SplitsTree 4 (Huson and Bryant, 2006) and a phylogenetic tree was reconstructed using the NJ method with Hamming distances (Figure 4.5). Generally, the relative distances between genotypes decreased when PSs were used rather than MSAs. Even genotype G,

which is the most distant, was not further removed than other genotypes. The overall clustering was not distinctly by genotype. The genotype C strains mostly clustered close to each other but other genotypes were intermixed. PS profiles did not appear to be distinct between the different genotypes given this conservation threshold. This may indicate different rates of evolution for PS profiles compared to genomic RNA sequence without being simply faster or slower. Due to few characters, evolution on the level of PS profiles can only occur in large jumps compared to smaller steps on the genomic sequence level. This is similar to protein structure as seen in Bamford and Stuart's work on phylogeny by capsid protein fold, where only a single character is used. This single character can take on a small number of values as the structure of a viral capsid protein is restricted to a limited set of functional folds. So whilst it is possible that that the threshold of 50% was too high to allow for sufficient resolution, it may also be that different subgenotypes and genotypes have diverged at different rates resulting in some C subgenotypes being more similar to other genotypes than other C subgenotypes.

To see how the sequences used in Chapter 3 sit within this phylogeny, they were added for another analysis. This time all 20 randomly selected sequences were included. This meant that two sequences, seq2 and seq11, which had large insertions and deletions, respectively, had to be aligned to the other ones. This proved difficult with ClustalΩ (Goujon et al., 2010; Sievers et al., 2014) and required manual alignment of those sequences in SeaView first (Gouy et al., 2010) giving preference to fewer, larger insertions/deletions. This adjusted MSA was used to reconstruct a NJ tree in SplitsTree 4 (Huson and Bryant, 2006). To be consistent also here Hamming distances were used (Figure 4.6). As expected the strains clustered by genotype including the randomly selected sequences. Interestingly, seq4, which is of unknown genotype, did not cluster with any of the others but was closest to Fs. The recombinants clustered with their major genotype, e.g. A/D with A.

Next, the 20 random sequences were also processed to create PS profiles as

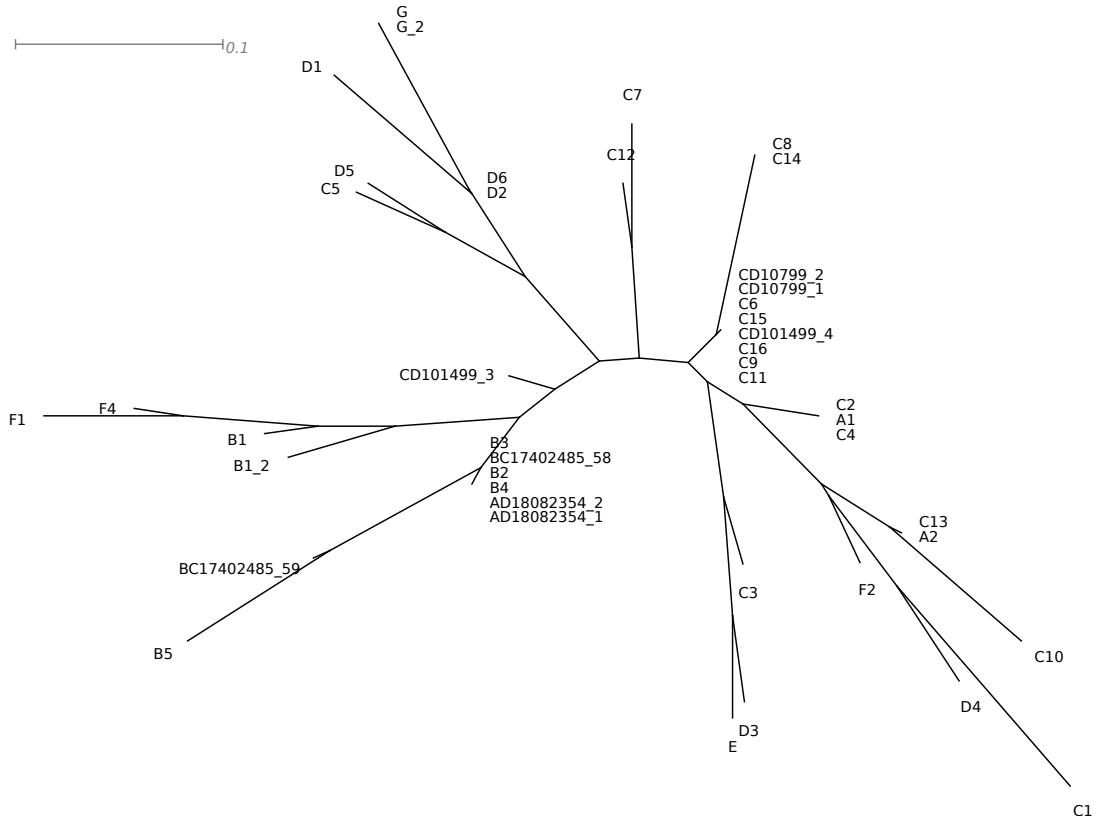

FIGURE 4.5: **Phylogenetic tree of PS profiles HBV genotypes.** NJ tree using Hamming distances including (sub)genotypes and A/D, B/C and two types of C/D recombinants was constructed in SplitsTree 4 (Huson and Bryant, 2006).

above. They were then pooled with the genotype representatives to find conserved PS blocks. Using the threshold 50%, 18 PS blocks were identified with this set of sequences of which 16 were informative (Figure 4.7). The blocks were at nucleotide positions 170–185, 264–287, 474–503, 867–884, 1029–1048, 1100–1114, 1191–1215, 1274–1297, 1323–1343, 1434–1458, 1499–1519, 1620–1638, 1709–1731, 2112–2125, 4277–4296, 4315–4342, 4503–4520, and 4521–4569 in the aligned sequences in pgRNA form. The aligned sequences are longer than before due to the large insertion in seq2 so later block positions cannot be directly compared to the ones above. Whilst PS2 and PS3 are still in blocks 867–884 and 1029–1048, respectively, PS1 is in block 4503–4520.

Based on the PS blocks, a NJ tree was reconstructed. It showed striking differences compared to the one with just the genotype representatives (Figure 4.8). Whilst before relative distances were small in general, here two sequences were far

FIGURE 4.6: **Phylogenetic trees of MSA of HBV genomes utilised in PS identification and genotypes.** An MSA in ClustalΩ (Goujon et al., 2010; Sievers et al., 2014) was used. The NJ tree was reconstructed in SplitsTree 4 (Huson and Bryant, 2006) using Hamming distances. Included are all reference (sub)genotypes as used for Figure 4.5 as well as the laboratory strain, the NCBI reference strain, and the 20 genomes randomly selected to be used for PS identification in Chapter 3 "Identification of a Packaging Signal Motif in HBV".

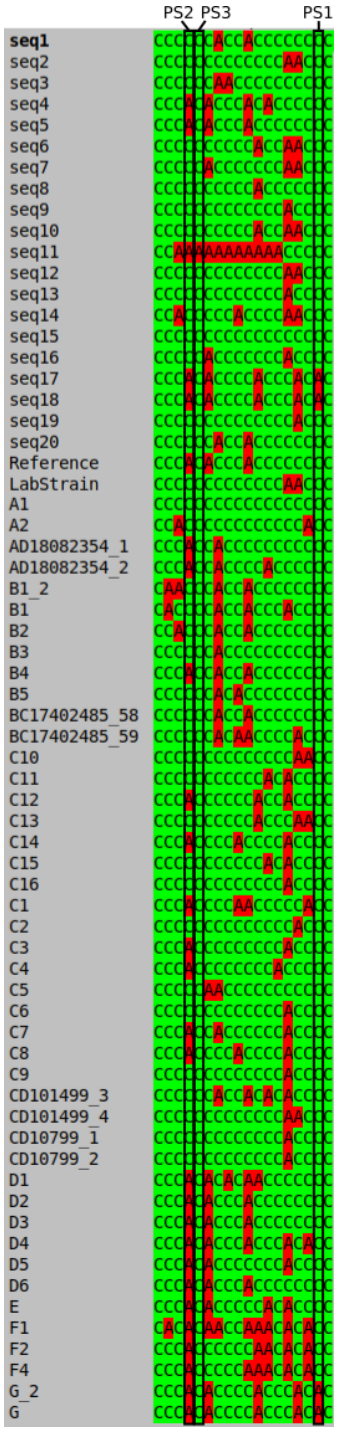

FIGURE 4.7: **Packaging signal blocks in randomly selected HBV strains.** The blocks were created using a conservation threshold of 50%. "C" (green) represents having a PS in that block and "A" (red) means no PS. The blocks where PS1, PS2, and PS3 are located are indicated. They are visualised in SeaView (Gouy et al., 2010).

removed from the rest. These were F1 and seq11. The F1 strains clustered most closely with the other F strains but was nevertheless more distant from these than before. That seq11 would be separate from other strains is not surprising considering it carries a large deletion. Since only 18 blocks were identified, losing any one or several through deletions would severely affect the similarity to the other strains. Looking at the block profiles in Figure 4.7 shows that seq11 only has a PS in seven out of the 18 blocks. However, apart from these large, obvious changes through inclusion of the 20 additional sequences, the two trees in Figures 4.5 and 4.8 differ considerably in general clustering. For example, previously, F2 was far removed from the other F strains, which clustered with the two B1 strains, whereas now all F strains cluster together and are far removed from B1 strains. These differences illustrate the large effect of the number of strains included and the resulting number of PS blocks given a certain conservation threshold.

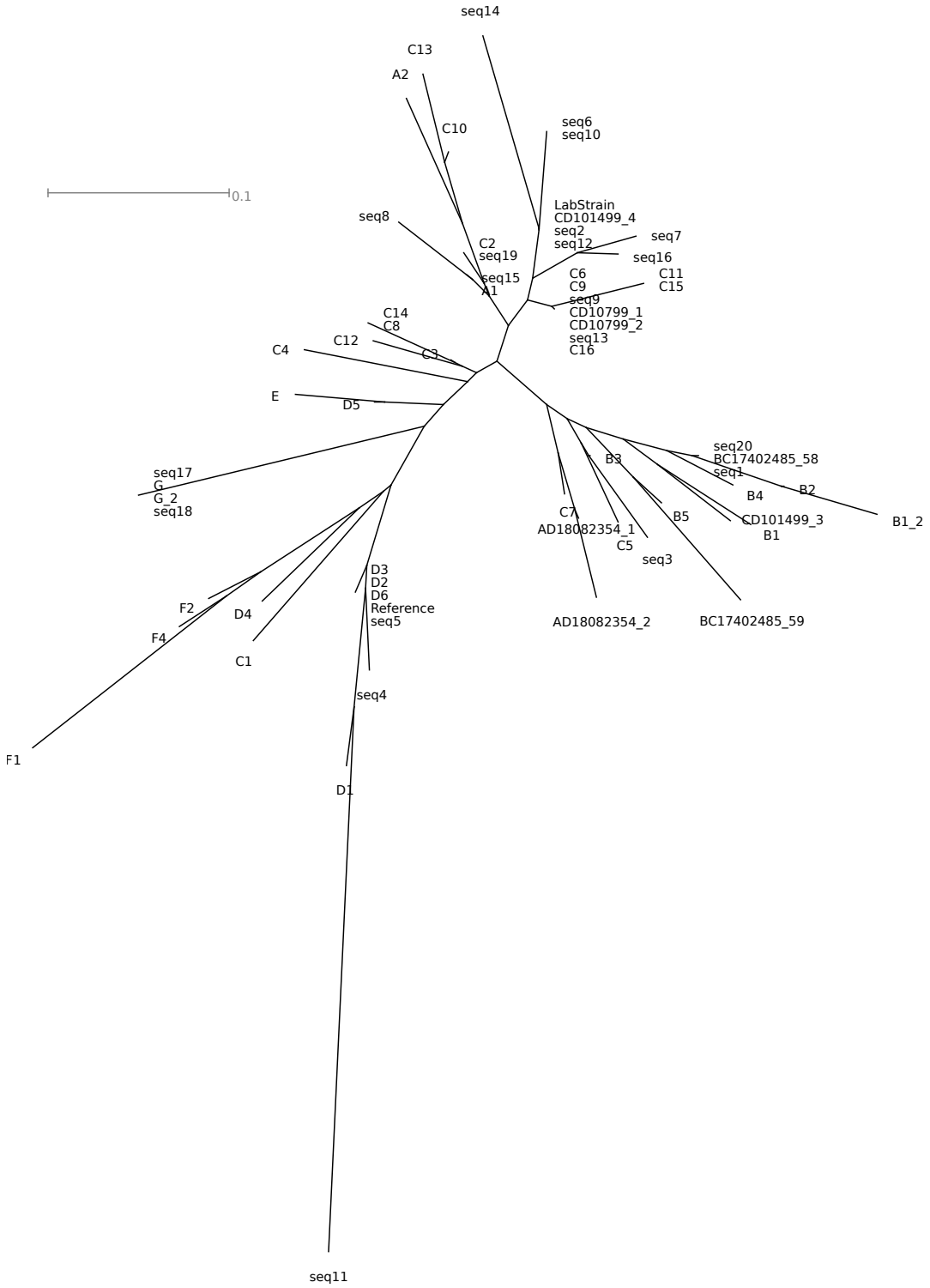

FIGURE 4.8: **Phylogenetic trees of PS profiles of HBV genomes utilised in PS identification and genotypes.** For the PS profile tree (left) the PS blocks were utilised. The NJ tree was reconstructed in SplitsTree 4 (Huson and Bryant, 2006) using Hamming distances. Included are all reference (sub)genotypes as used for Figure 4.5 as well as the laboratory strain, the NCBI reference strain, and the 20 genomes randomly selected to be used for PS identification in Chapter 3 "Identification of a Packaging Signal Motif in HBV".

4.5 Longitudinal and Regional Study Data

Above, the PS phylogeny method was applied to a small number of representative genomes from different genotypes. These are expected to be quite different and setting a conservation threshold is not trivial as can be seen from the drastic changes in clustering when more sequences were included. Therefore, instead of looking at distantly related sequences, I applied the method to four data sets with evolutionarily close sequences. They stemmed from either longitudinal studies, where the viruses from the same people were sequenced at different time points or family members who infected each other over time (mother to child), or regional studies.

4.5.1 20 Patients from one Region in Japan by Michitaka et al. (2006)

The first data set stemmed from a regional study. Michitaka et al. (2006) analysed the origin of genotype D strains in Ehime in Western Japan, where it is endemic. They obtained 20 complete sequences and compared them phylogenetically to D strains from other parts of the world, as well as representative strains from other genotypes. The set of sequences used in the study and here with their respective accession numbers and aliases as used here is shown in Appendix A Table A.4. When reconstructing the phylogenetic tree, the authors found that the Ehime strains clustered closely together, distinct from other D strains, implying a common origin. Molecular evolution analysis revealed that these first started to spread in Ehime around 1940 and peaked in the 1970s (Michitaka et al., 2006).

The set of sequences used in Michitaka et al. (2006) provides the opportunity to work with sequences that have diverged fairly recently compared to subgenotypes. This can provide insights into the speed of PS evolution in HBV. Since these strains diverged less than 100 years ago, not much change is expected on the PS level. The same sequences, including reference strains, as in Michitaka et al.

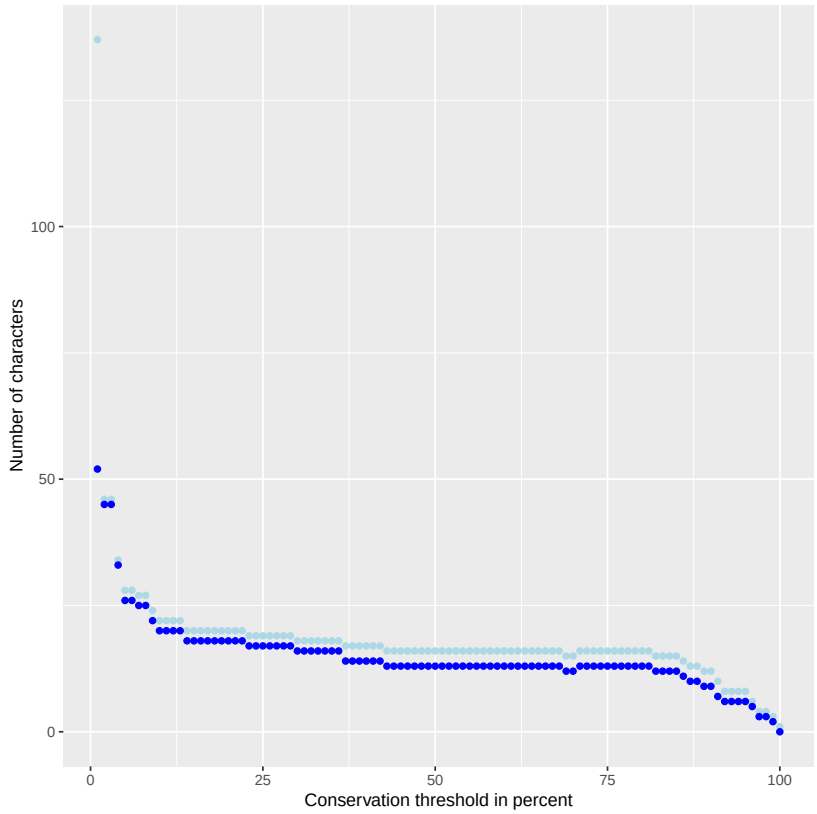

FIGURE 4.9: **Impact of conservation threshold on number of characters in Michitaka data set.** Conservation thresholds from 1% to 100% were tested for the data set by Michitaka et al. (2006). The number of total characters, i.e. blocks, (light blue) and informative characters (blue) they resulted in are shown.

(2006) were utilised and processed as described above. The number of blocks by conservation threshold was plotted to identify a suitable threshold value (Figure 4.9). This curve was notably flatter around the middle values than the one for different HBV genotypes only. To enable a split between Ehime and other strains, the threshold needed to represent the number of sequences versus references. Here, about 20 Ehime and 40 reference sequences were used. This means that only 33% of the sequences were study sequences. Consequently, using a threshold of 34% or higher would make it impossible to identify PSs present specifically in the Ehime strains. The threshold was therefore set to 30% rather than 50%, which allowed for slightly more informative characters as seen in the graph. With this, 18 PS blocks were identified, of which 16 were informative. These were located at nucleotide positions 177–198, 265–283, 474–498, 499–513,

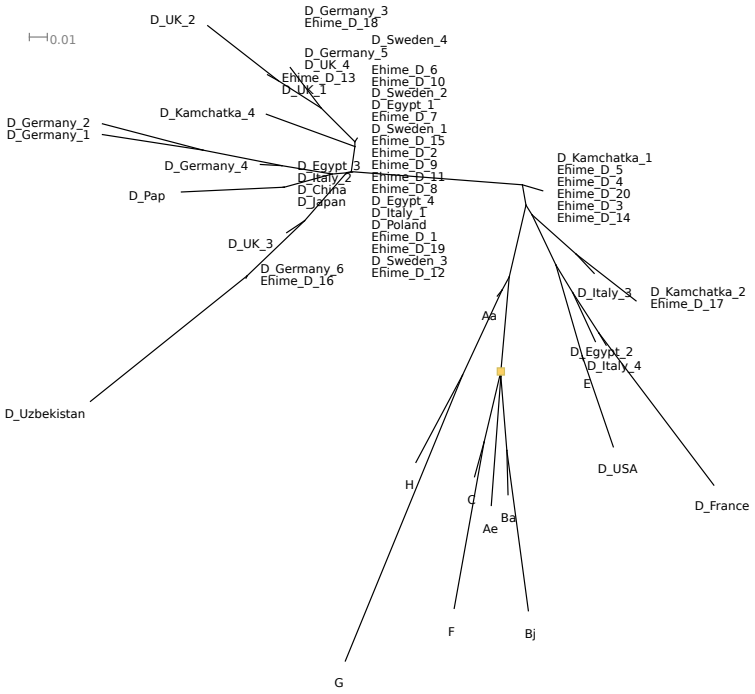

FIGURE 4.10: **Phylogenetic tree of PS profiles of 20 HBV patients from one region in Japan and reference genotypes.** The NJ tree was reconstructed in SplitsTree 4 (Huson and Bryant, 2006) using Hamming distances. Sequences included 20 different HBV patients with genotypes D from one region in Japan compared with genotype D sequences from other parts of the world and additional reference genotypes. Data from Michitaka et al. (2006).

571–594, 1028–1055, 1191–1218, 1278–1296, 1303–1327, 1328–1342, 1504–1529, 1620–1637, 1707–1731, 2107–2120, 3069–3097, 3115–3142, 3201–3225, and 3226–3262. They were largely the same as but not completely identical to the ones identified in Section 4.4 when the sequences from the paper were included. Despite adjustment of the threshold there was no distinct clustering of the Ehime strains compared to other genotype D strains (Figure 4.10). While they all stayed fairly closely together, they were separated to some extent and intermixed with genotype D strains from other origins. This indicates that some degree of PS evolution occurs even during a relatively short time span of only a few decades to a century as seen through the split of Ehime strains. On the other hand, divergence was also limited as seen through the intermixing of other genotype D strains.

4.5.2 Mother and Three Children by Sede et al. (2014)

The next data set included sequences from a mother and her three children, one daughter and two sons (Sede et al., 2014). Sede et al. (2014) obtained three sequences for the mother and two for each of the children at different time points. By comparing the sequences also in the context of the immune status, they found that little divergence happened in the initial stages of infection when HBeAg is still abundant, the immune-tolerant phase, while less conservation was observed once the respective host started clearing HBeAg, the immune clearance phase (Sede et al., 2014). This indicates that there is little evolutionary pressure for the virus in a new host until the immune system springs into action. In addition to the familial samples, the study included representatives from different genotypes as reference strains. All strains with alias and accession numbers are summarised in Appendix A Table A.5. These data provide an opportunity to investigate PS evolution on an even shorter time scale compared to the regional set above. Moreover, with the different time points, it could also be seen whether PS evolution is also affected by immune status of the host as general sequence is.

As above, the sequences and references were converted to pgRNA form, folded, and PS profiles generated. As explained earlier, the conservation threshold needs to be adapted to the data set to allow for sufficient resolution. As seen in Figure 4.11, the number of blocks increases even more rapidly in this data set with thresholds below approximately 30%. Whilst 50% would be more conservative, it is not suitable for these data. In this case there is a high number of reference sequences (45) compared to study sequences (nine). Keeping the default threshold of 50% would require at least 27 strains to have a PS at any nucleotide to generate a block. This means that PS blocks specific for the study sequences, of which there are only nine, would not be considered. Therefore, for the next step of identifying conserved PS blocks the threshold needed to be adjusted as follows: To better represent the ratio of study sequences to references (9:45), a conservation threshold of 5% rather than 50% was used. However, this meant

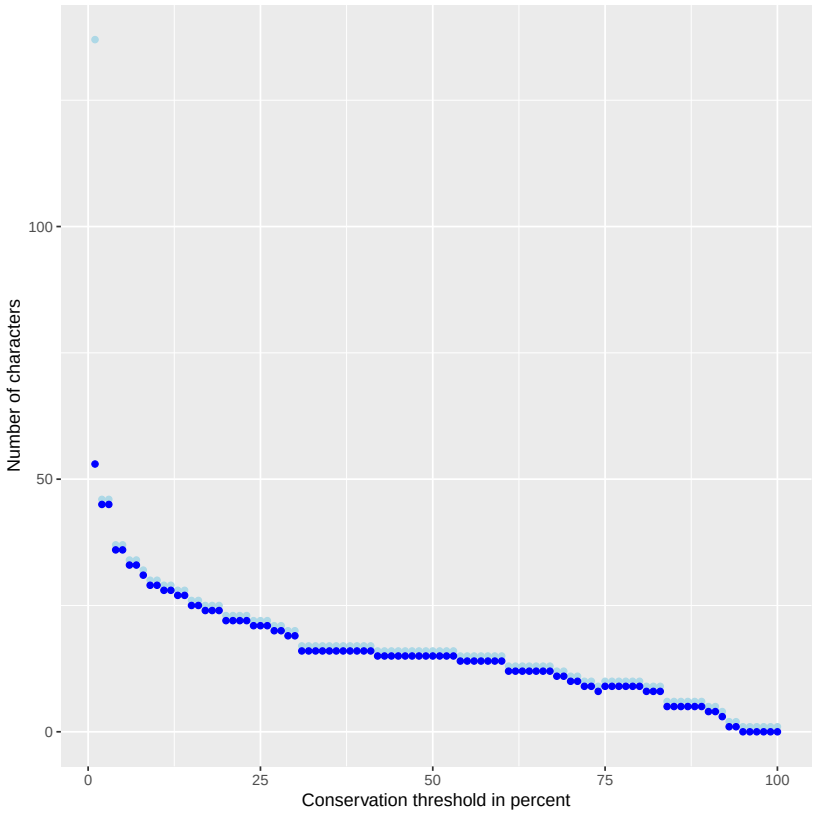

FIGURE 4.11: **Impact of conservation threshold on number of characters in Sede data set.** Conservation thresholds from 1% to 100% were tested for the data set by Sede et al. (2014). The number of total characters, i.e. blocks, (light blue) and informative characters (blue) they resulted in are shown.

going into the territory of quickly increasing block numbers. Unsurprisingly, this resulted in a larger number of blocks: 37 blocks, all except for one informative. Their respective nucleotide positions were at 170–198, 263–294, 331–355, 361–385, 387–406, 473–497, 498–522, 523–533, 571–594, 598–619, 802–828, 857–893, 929–946, 953–974, 1023–1055, 1081–1105, 1106–1118, 1188–1220, 1253–1277, 1278–1299, 1303–1327, 1328–1348, 1386–1416, 1433–1457, 1494–1529, 1614–1644, 1707–1731, 1748–1773, 2107–2120, 2305–2332, 2349–2385, 2774–2799, 3069–3097, 3111–3139, 3160–3183, 3192–3216 (PS1), and 3217–3263 (non-informative).

The block profiles were supplied to SplitsTree 4 (Huson and Bryant, 2006) to generate NJ phylogenetic trees using Hamming distances (Figure 4.12). Interestingly, in this data set the genotypes clustered clearly together. While in the previous data sets there was a lot of intermixing, here there is none with

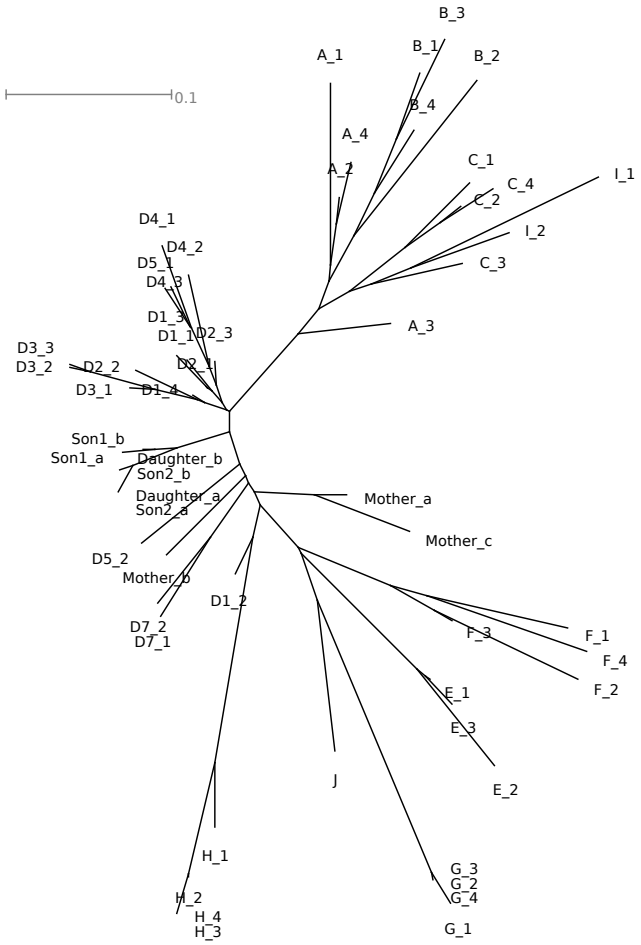

FIGURE 4.12: **Phylogenetic tree of PS profiles of HBV sequences from one mother with her three children and reference genotypes.** NJ tree using Hamming distances including sequences from eight different patients in 1979 and 2004 and reference genotypes was constructed in SplitsTree 4 (Huson and Bryant, 2006). Data from Sede et al. (2014).

each genotype clustering exclusively with other representatives. Similarly, the study samples of genotype D were all within the D cluster. Despite the higher resolution of the genotype level, the closely related familial sequences did not all cluster closely together. While the children's samples were all close, the mother's appeared separately. This indicates that some divergence on the PS level may occur early in transmission regardless of immune status as there was not much difference between the time points as opposed to the results of the original study.

4.5.3 Eight Patients at Two Time Points 25 Years Apart by Osiowy et al. (2006)

The next data set stemmed from a longitudinal study that looked at the changes of the HBV virus in eight independent patients between 1979 and 2004. These patients were in the immune clearance phase of infection with no detectable levels of HBeAg and showed no symptoms at the beginning of the study in 1979 (Osiowy et al., 2006). As seen in Sede et al. (2014) more genetic diversity accumulated after immune clearance, so studying patients already in this phase ensured a clear baseline for the 25 year follow-up. The authors analysed the sequences for substitution rates, regions of hypervariability, and regions of high synonymous to non-synonymous mutations. They found a higher substitution rate than previously published: 7.9×10^5 versus 1.5 to 5×10^5 substitutions per site per year. The most variable regions are in the C gene and at the overlap between the P and preS/S gene. These differ, however, in their synonymous to non-synonymous substitution ratios. While the P gene has the highest, the C gene has the lowest ratio. This points towards high pressure for conservation of Pol, while the opposite is true for HBcAg, which is under more pressure to evolve (Osiowy et al., 2006). For phylogenetic tree building, the study only included a small number of representative genotype strains as references. All strains used for tree-building with alias and accession numbers are shown in Appendix A Table A.6.

The data set from Osiowy et al. (2006) provided an opportunity to look at the evolution of PSs within a number of hosts after 25 years. Both samples were taken at immune clearance stage of infection when according to Sede et al. (2014) more sequence evolution takes place. The sequences shown in Table A.6 were processed as before to identify blocks of conserved PSs. Here the threshold was kept at 50%, where the curve was relatively flat (Figure 4.13) This resulted in 17 PS block located at nucleotide positions 170–185, 263–286, 377–402, 477–500, 866–893, 1030–1055, 1100–1115, 1273–1296, 1325–1342, 1498–1518, 1615–1643,

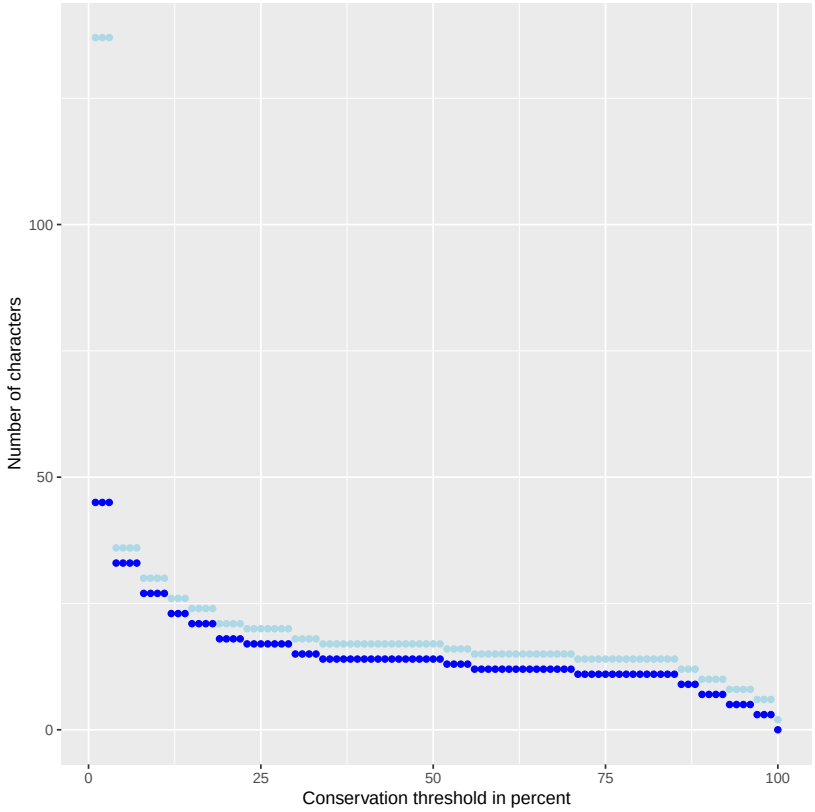

FIGURE 4.13: **Impact of conservation threshold on number of characters in the Osiowy 2006 data set.** Conservation thresholds from 1% to 100% were tested for the data set by Osiowy et al. (2006). The number of total characters, i.e. blocks, (light blue) and informative characters (blue) they resulted in are shown.

1707–1731, 2107–2120, 3072–3096, 3112–3139, 3198–3222, and 3223–3259. 14 of these were informative.

The block profiles were supplied to SplitsTree 4 (Huson and Bryant, 2006) to generate a NJ tree using Hamming distances as before (Figure 4.14). Many of the study sequences clustered closely together. For patient3, patient6, and patient7 there was no change at all in the PS profiles, the blocks were identical between sequences from 1979 and 2004. The other patient samples diverged between the two time points and did not cluster together. Comparing to the distances of the different reference strains, it can be seen that some patient viruses, patient1, patient5, and patient8, diverged as much over 25 years as different genotypes. Taken together this phylogeny suggests that PS evolution is variable and occurs in leaps. This leads to a third of the patient viruses not changing at all over

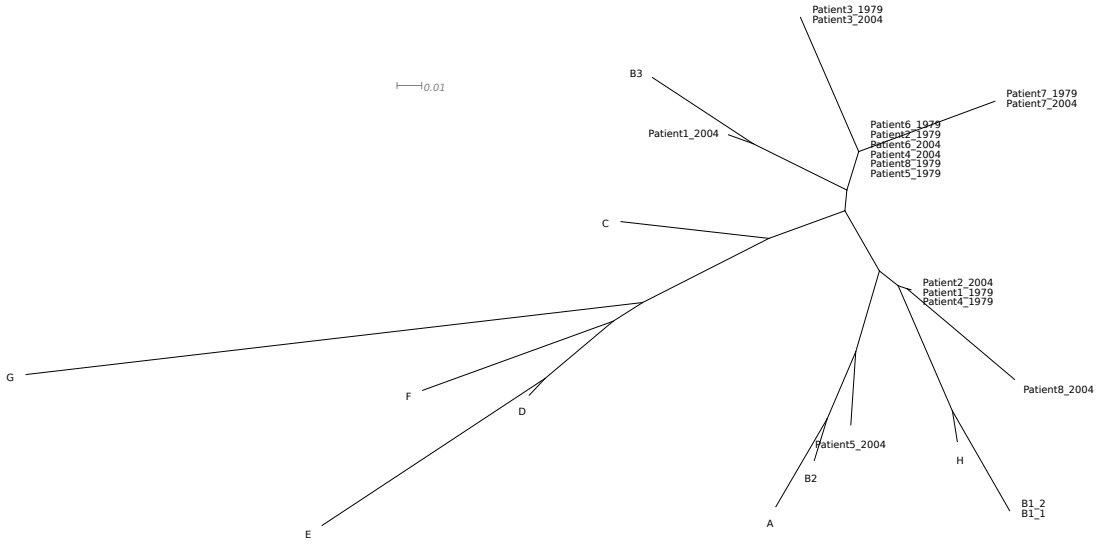

FIGURE 4.14: **Phylogenetic tree of eight HBV patients and reference genotypes.** NJ tree using Hamming distances including sequences from eight different patients in 1979 and 2004 and reference genotypes was constructed in SplitsTree 4 (Huson and Bryant, 2006). Data from Osiowy et al. (2006).

25 years while another third diverged largely. It is possible that this is due to PSs acting in groups and thus several would have to change to reach another local fitness maximum. With only 17 blocks and 14 different ones, a change to any single one would appear as a large effect on the phylogeny. It is, therefore, also possible that with only 14 informative characters not enough resolution was achieved to properly study PS profile evolution over this period of time.

4.5.4 One Patient at Ten Time Points over Nine Years by Osiowy et al. (2010)

The final data set utilised here included viral sequences from only one patient, who was followed over nine years collecting a total of ten time points between 1999 and 2008 (Osiowy et al., 2010). This patient was infected with recently discovered genotype I virus. During the study period the patient received two separate courses of an antiviral, which did not clear the infection. HBV resurfaced when the first drug treatment was stopped and an emergence of stable nucleotide changes was observed a few years afterwards coinciding with this (Osiowy et al.,

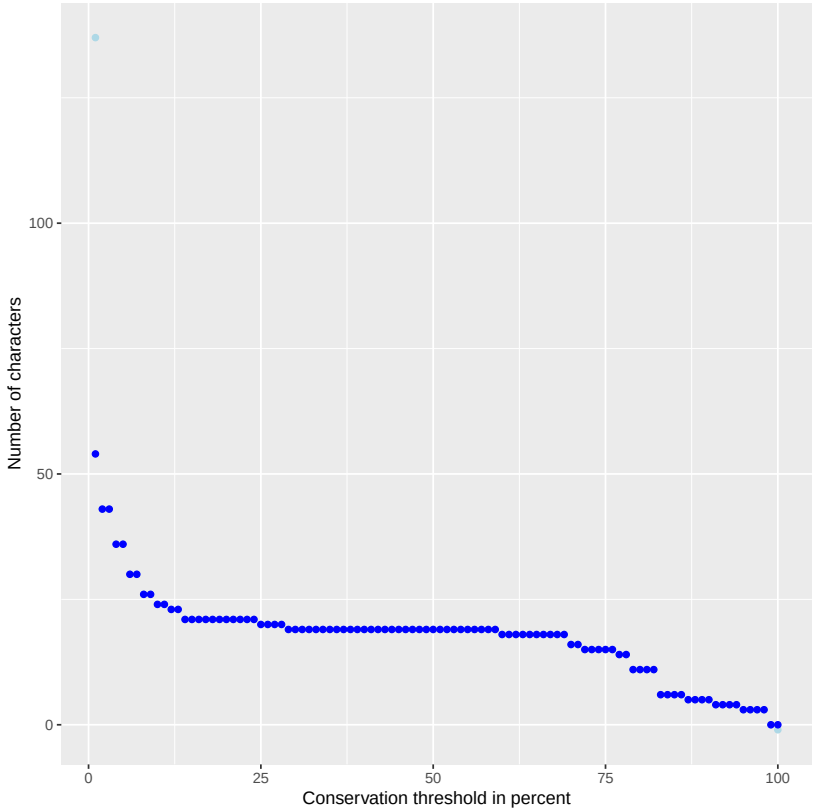

FIGURE 4.15: **Impact of conservation threshold on number of characters in the Osiowy 2010 data set.** Conservation thresholds from 1% to 100% were tested for the data set by Osiowy et al. (2010). The number of total characters, i.e. blocks, (light blue) and informative characters (blue) they resulted in are shown.

2010). In addition to the sequences obtained from the patient (Isolate1–10) the authors included other genotype I strains from Asia and a few representatives from other genotypes as reference strains for phylogenetic analysis. All strains with alias and accession numbers are shown in Appendix A Table A.7.

The set of sequences provided through the study by Osiowy et al. (2010) provides the most short term view on evolution. As opposed to previous data sets evolution is only followed in one host omitting the effects of transmission and establishing of infection in a new host. Furthermore, the resolution is higher compared to the study by Osiowy et al. (2006) since more samples were taken within shorter time intervals. To test how quickly PS profiles would evolve within one host, the sequences from the study together with the references were processed as above to identify conserved PS blocks. To better represent the ratio

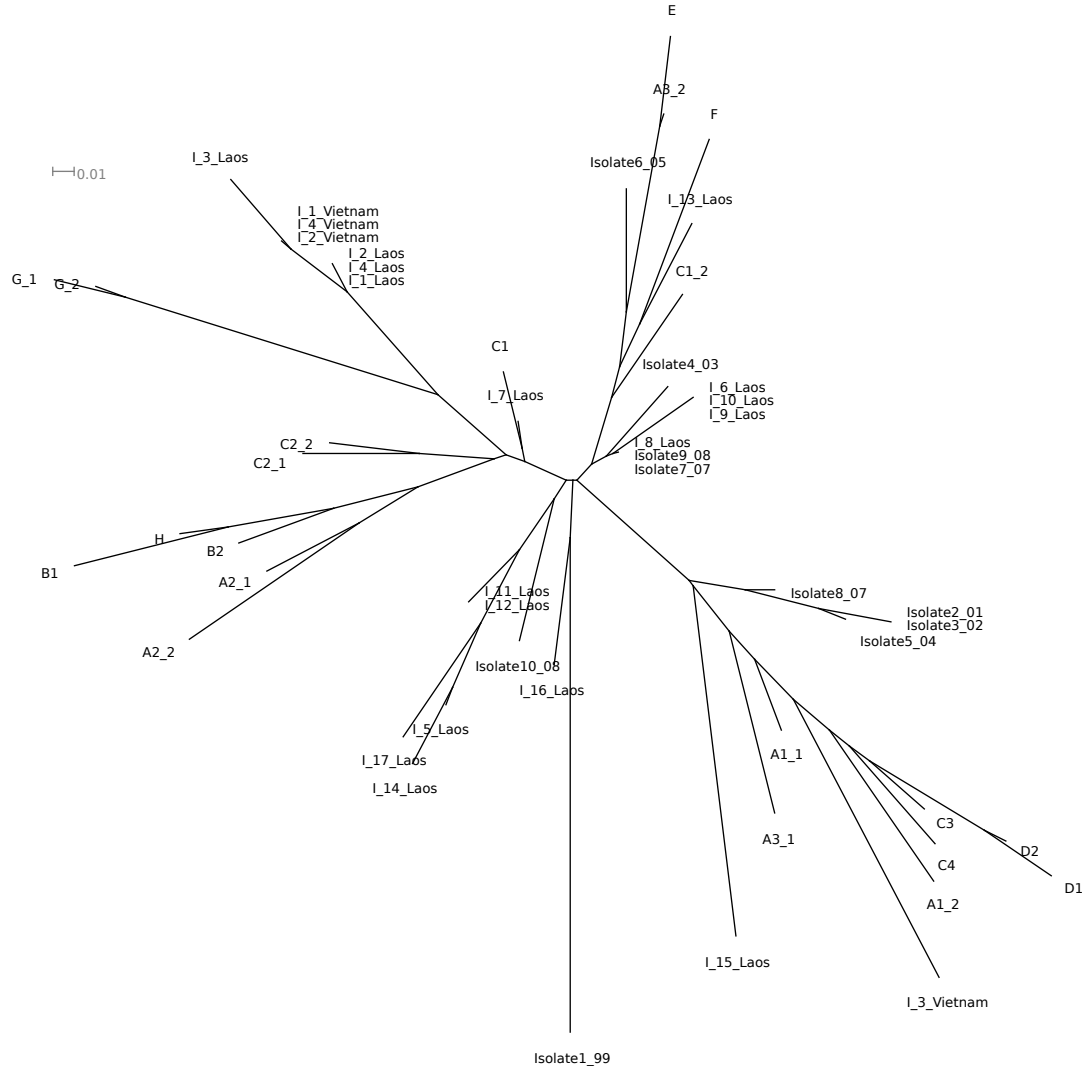

FIGURE 4.16: **Phylogenetic tree of PS profiles of one HBV patient over nine years and reference genotypes.** NJ tree using Hamming distances including sequences from eight different patients in 1979 and 2004 and reference genotypes was constructed in SplitsTree 4 (Huson and Bryant, 2006). Data from Osiowy et al. (2010).

of study sequences versus references (10:40) a threshold of 20% instead of 50% was used. This still coincided with the relatively flat part of the curve, whilst having a higher resolution than 50% (Figure 4.15). 21 blocks were identified, all of which were informative. They were located at nucleotide positions 170–188, 263–286, 473–502, 861–888, 1024–1048, 1099–1115, 1188–1216, 1266–1296, 1303–1327, 1328–1344, 1433–1457, 1498–1519, 1614–1643, 1707–1731, 2107–2120, 2306–2331, 2775–2798, 3070–3096, 3112–3139, 3198–3222, and 3223–3260.

A phylogenetic tree was reconstructed in SplitsTree 4 (Huson and Bryant, 2006) using the NJ method with Hamming distances (Figure 4.16). With 21 characters the sequences clustered partially by (sub)genotypes but not as well as when 37 blocks were used (see Section 4.5.2 Figure 4.12). Different representatives of G, C2, and A2 were close, while the two A3 strains separated considerably with A3_2 clustering with E and F. Generally the different A and C strains were spread across the phylogenetic tree and did not form a separate genotype cluster as in Section 4.5.2. The same is true for the more study-relevant reference strains of genotype I. The sequences from Laos were spread across the tree but three of the four Vietnam strains had identical profiles. While a part of the I strains formed two separate clusters (I_1–4_Laos, I_1,2,4_Vietnam and I_5,11,12,14,16,17_Laos), one clustered with C1 (I_7_Laos) and one with F (I_13_Laos), and two were isolated (I_15_Laos and I_3_Vietnam). The sequences from the patient themselves did not display a meaningful pattern over the years. The first isolate from 1999 (Isolate1_99) clustered most closely with I_16_Laos and the last isolate from ten years later (Isolate10_08). The two subsequent isolates (Isolate2,3_01,02) had identical profiles and were close to isolates 5 and 8 (Isolate5,8_04,07) but were considerably different from the first and the fourth (Isolate4_03), which clustered most closely with different strains from Laos and isolates 7 and 9 (Isolate7,9_07,08). Isolate6_05 did not cluster with the others or even other I strains, but with A3 and E. There appear to have been reversible changes in the PS profiles of the HBV in this patient over time. The first sample from 1999, which is quite distant from the rest, was isolated before the patient received antiviral treatment in 2000 (Osiowy et al., 2010). This may explain the large shift to the next isolate. The following changes were partly reversible with an outlier at isolate 6.

4.6 Comparing Compact Packaging Signal Profiles

During the phylogenetic analysis of HBV strains it became apparent that only few PSs are present in the virus. Even with a very low conservation threshold only 37 blocks were identified. While this makes it more difficult to reconstruct phylogenetic trees by the method developed in Chapter 2, it provides an opportunity to manually examine the PS profiles for conserved PSs and patterns. To obtain a quick look at the distribution of PSs, compact versions of the PS profiles were generated. "Compact" here means that the PS profiles, which have the same length as the respective genomic sequence, are compressed to the most essential information: the pattern of PS distribution that can then be viewed all at once. This transformation method made use of the predicted SL length and shortened every PS to a single character while the space between them was shortened by a factor of the SL length. This is a crude method and relative lengths are not well preserved. It was therefore necessary to manually adjust these profiles to align corresponding regions. One way to enable that was to use anchor structures in the sequences. These were SLs that should occur in every sequence at the same genomic position. In HBV ε lent itself as anchor point. The alignment and identification of patterns in the profiles was also made easier through the conversion to pgRNA form. This ensured an anchor at both 5' and 3' end of each sequence.

The compact PS profiles showed that certain positions have a high degree of conservation between genotypes (Figure 4.17). These would correspond to the PS block positions observed even for the strictest conservation threshold. Interestingly, there were many highly conserved PS sites at the 3' end of the pgRNA. Most strains had a pair and a set of three PSs there. The notable exception was genotype G, which also had several PSs at the 3' end but they were very different structures compared to the other genotypes. Two sites at the 5' end also showed a high degree of conservation as well as several sites dispersed

FIGURE 4.17: **Compact Packaging Signal Profiles of HBV (sub)genotypes.** Manually aligned profiles are shown for the same set of (sub)genotype sequences used in section 4.4 visualised in SeaView (Gouy et al., 2010). PSs are shown as "C"s (green), ε as "T"s (blue) and remaining sequence as "A"s (red).

towards the middle. The latter half of the pgRNA appears notably sparse in PSs. This particular spread of PSs along the genome may be of functional importance during the assembly and packaging process.

4.7 Discussion

The method to reconstruct phylogenetic trees based on PSs developed in Chapter 2 "Phylogenetic Algorithms" was applied here to different HBV data sets. These data sets provided an opportunity to investigate PS evolution on different time scales starting at comparing the differences between genotypes and ending with the changes over ten years in one host.

To reduce noise conservation thresholds were utilised so that PSs were only considered when they occurred in at least that many strains in the set. Plotting the conservation thresholds against the number of blocks illustrated the landscape of this variable. Unsurprisingly, the number of PS blocks was largely dependent

on the conservation threshold with high thresholds yielding a low number of blocks and vice versa. For most data sets the curve showed a pronounced flat portion around 50%, which was the default value. This means that using a somewhat higher or lower value would not have affected the number of blocks by much or at all. The default value was usable for the comparison of different subgenotypes but was too high for many of the other data sets, because they included a small set of study sequences and a larger set of reference sequences. Therefore, when longitudinal or regional study strains were used with respective reference sequences lower thresholds were applied to allow separate clustering of these strains. Setting the threshold higher than the percentage of study sequences would have made it impossible for blocks specific for this subset to be picked up.

While for most data sets it was possible to choose a conservation threshold within the flat region of the curve, for one this was not possible due to the large relative number of reference strains requiring a threshold of only 5%. This resulted in 37 PS blocks, approximately double the number found with more conservative thresholds. This has important implications for the block profiles and the reliability of the trees. At the very least there were eleven characters. This corresponds to $2^{11} = 2048$ different possible combination for the block profiles. However, realistically and through use of the conservation threshold, the combinations with few or no PSs are unlikely to actually be present. Whilst there is potential for separation of the genotypes based on which combination they utilise, achieving reasonable resolution was nevertheless difficult since PSs specific for a certain genotype or subgenotype would not be considered for the sake of avoiding noise. Moreover, a change in any one block could have a strong impact on the clustering. This resulted in strains of the same genotype not clustering well together with 50% but doing so at 5% conservation threshold, where 37 blocks were generated. On the other hand, better clustering does not necessarily mean that more blocks are superior. 37 is still a fairly low number of characters so irrelevant ones could have a strong impact on the phylogeny. The risk when setting

the threshold too low is an increase in noise and that PS profiles become little more than compacted sequence information. This means that small mutations can make single SLs appear or disappear, which may not have much functional importance, but would weigh in greatly in the phylogeny. What this means is that the underlying SL selection algorithm is not perfect so conservation is also used to filter out artefacts, which would bias the phylogeny.

How the outcome of the phylogeny method can be influenced by the data and threshold could be seen from the first data set, which was for comparison of different (sub)genotypes. When only the representative strains were used with a 50% threshold, eleven blocks were generated, and strains of the same genotype did not necessarily cluster together. However, when the 20 sequences from Chapter 3 were included, the block number increased to 18 with the same threshold and clustering of the original sequences changed. Note that 14 of the 20 study sequences were of genotype C. With almost half the sequences being of genotype C the new data set was further biased towards that genotype. Genotype C has the highest number of subtypes (Mulyanto et al., 2010, 2011, 2012), which means that it is not only overrepresented in the original data set but also that there is a considerable amount of variability within this genotype. It may, therefore, not be surprising that there is also significant variation on the PS level. Generally, the study sequences clustered with other sequences of their genotypes but some intermixing still occurred. Like other structural/functional elements PSs are likely to evolve within stricter boundaries as there are fewer places to change and not all combinations viable. Therefore, convergent evolution of different genotypes is as much possible as divergence of subgenotypes.

To get a clearer picture of the way PSs evolve, the scope was changed to more closely related sequences starting with a set of genotype D strains from a particular region, Ehime, which are thought to have been introduced to that region approximately 100 years ago. Despite forming a separate cluster from other genotype D reference strains in the original study (Michitaka et al., 2006),

the Ehime strains intermixed with these when trees were reconstructed based on PS profiles. Assuming that they all have one D strain common ancestor this means that since the introduction of this strain 100 years ago they diverged on the PS level within the boundaries of D strains. This points towards evolution in jumps that partially coincide with genotypes, which makes sense given that a new genotype is defined by being at least 8% different from any other. In this time frame there may not have been enough time to diverge from genotype D and converge towards other genotypes on the PS level.

Zooming in even more closely the viral sequences isolated from a mother and her three children at different time points were analysed (Sede et al., 2014). Interestingly, the children's sequences formed a cluster together, separate from the mother's sequences. This was despite the use of a very low conservation threshold and a relatively high number of PS blocks used for tree building. The strains from the children did not separate clearly based on time points. Strains isolated in early infection from two children were on the same leaf and only slightly separated from the respective strains isolated later, which were also on one leaf in the phylogenetic tree. This indicates an evolutionary jump when infection is established in a new host and less evolution later on, which stands in contrast to the findings for genomic sequence, where more diversity is found later in infection during the immune clearance phase (Sede et al., 2014). The observed patterns are in line with the colonisation-adaptation trade-off (CAT) model, which describes the idea that different viral strains are dominant at different points in infection, because the properties that are important to colonise a new host are different from those required to adapt to it or evade its immune system (Lin et al., 2015). A virus may lose some fitness, i.e. replication efficiency, to adapt to a host but needs high replicative efficiency to colonise a new host. Thus, it makes sense for PSs to vary more in the early stages of infection since PSs are not directly affected by the immune system response, i.e. no antibodies are directed against them. Instead they are important for establishing infection and colonisation in

the first place through their fitness contribution.

This idea that less evolution of PSs occurred in the immune clearance phase was further supported by looking at a set of sequences from different patients at two time points, both at late infection stages (Osiowy et al., 2006). Many but not all samples from different time points were on the same leaf, i.e. had identical PS profiles. However, some evolved substantially during the 25 years that lay between the time points indicating, again, a jump-like evolutionary pattern.

The most detailed view on short term PS evolution was gained by analysing a set of viral sequences isolated from one patients at different time points of the course of ten years (Osiowy et al., 2010). The most striking result was the large shift in PS profiles after treatment with a reverse transcriptase inhibitor. Since this type of drug should not directly affect PSs or their evolution, it is most likely that this shift was a side effect of a general large change in sequence of the viruses in the face of this evolutionary pressure. Otherwise, there were many reversible fluctuations observed over time showing that PS profiles are not necessarily stable over short amounts of time but may vary between different stable states.

The phylogenetic trees in this study were reconstructed using the NJ method and Hamming distances. When testing different phylogenetic methods for their ability to reconstruct a known human immunodeficiency virus (HIV) transmission tree with 13 taxa, Leitner et al. (1996) found that all methods performed equally well. Instead, it was the part of the HIV genome used for the analysis, which had the highest effect on accuracy (Leitner et al., 1996). These results in HIV stand in contrast to recent work in HBV. Godoy et al. (2020) found that the choice of evolutionary assumptions and phylogenetic methods markedly affects the topology of a rooted HBV phylogenetic tree. This disagreement may stem from the fact that two different viral species were examined; however, the relatedness of the taxa may also play a role. Leitner et al. (1996) attempted to reconstruct a transmission tree, indicating closely related sequences similarly to the ones I analysed on the longitudinal data sets, whereas Godoy et al. (2020)

compared HBV on the genotype level using a protocol that is more similar to the first set of comparisons I performed. It is therefore possible that phylogenetic trees are more or less sensitive to the tree building method depending on the level of relatedness between the taxa.

NJ as used here is a distance-based method, which means that differences in characters are first translated into a distance matrix using some evolutionary model, e.g. Hamming distance (Saitou and Nei, 1987). As such it can be applied to any type of character as long as it is possible to calculate a distance matrix (Yang and Rannala, 2012). This matrix is then used to reconstruct the phylogenetic tree rather than the original characters (Saitou and Nei, 1987). NJ is the most widely used distance-based method, because it produces good results while being computationally efficient. However, a suitable substitution model is crucial and the method can struggle with accuracy when the compared characters are too distant (Yang and Rannala, 2012). One of the most accurate methods is thought to be maximum likelihood (ML), which produces the most probable trees based on the characters directly. However, it is very computationally intensive and also requires a suitable evolutionary model similar to NJ (Yang and Rannala, 2012). The fact that NJ usually produces good results and can be applied to any type of data, whereas this may not be the case for probability methods such as ML, made it a suitable starting point in this initial test of the PS-based phylogeny method. However, as stated before, the method can only be as good as the distance matrix it is based on, so using a suitable evolutionary model is crucial. Here I used Hamming distances, which are a simplistic tool to calculate evolutionary distance. To date, there is no better substitution model for PS profiles. There are some approaches for RNA sequences, which also take their structure into account, but they do not work under the evolutionary assumptions that are likely to govern PS SLs. Instead, a novel evolutionary model specific for this application would have to be developed, which is not trivial as was discussed in detail in Section 2.5. Experiments on PS evolution, and studying further

samples with known relatedness, may provide necessary information to develop such a model. Until then, given the small amount of characters and the fact that they are only present in one of two states, i.e. PS and not PS, Hamming distances are considered the most appropriate model available.

The novel PS-based phylogenetic analysis was based on few characters only in HBV. When the conservation threshold was relaxed, more PS blocks were identified, so that there were more characters available for phylogeny, which changed clustering of reference sequences to an extent. The effect of the number of taxa and characters on tree accuracy has been studied in the past. More characters tend to improve accuracy, whilst there is disagreement on the effect of including more taxa, which may be due to different ranges of taxa and characters studied (Graybeal, 1998; Bremer et al., 1999). The consensus is, however, that too few characters result in difficulties resolving an accurate phylogenetic tree (Bremer et al., 1999; Scotland et al., 2003). The question, thus, remains whether trying to reconstruct phylogenies from PSs in viruses such as HBV that can only provide relatively small numbers of characters, is meaningful.

The most comparable type of character to PS profiles are traditional morphological characters. Despite molecular features such as genetic sequences, which can provide hundreds or thousands of characters, being the main approach for phylogenetic reconstruction nowadays, morphology is still used to derive phylogenies, and methods are being developed and improved (Wright et al., 2016). Even when morphological characters are often not numerous enough for useful tree resolution, they have been shown to be useful in improving phylogenies based on molecular characters by providing a framework for analysis (Scotland et al., 2003). As Wheeler et al. (2013) argued, basing phylogeny solely on molecular characters, in this case DNA sequence, ignores a plethora of additional information about species evolution that can be found in, e.g., morphology or behaviour, and can give misleading results. For viruses, one of these additional features to study could be their PS profiles.

As seen in the recent work by Godoy et al. (2020), parts of the evolutionary history of HBV genotypes are still debated. More information gained from alternative characters may improve our understanding of how this virus has evolved in humans once it is possible to root a PS-based phylogenetic tree and thus combine the resulting information about the topologies with PSs, providing a framework analogous to the use of morphological characters by Scotland et al. (2003). Alternatively, PS information could be added to sequence information for tree building similar to the ProfDistS program, which uses both sequence and RNA structure information to reconstruct phylogenies (Wolf et al., 2008). For now I have shown that it is possible to reconstruct phylogenetic trees using PSs as characters. By doing so, our knowledge has been expanded on how these functional features evolve at different time scales and how they differ in that respect from the sequence that underlies them. Especially, using data from short-term longitudinal studies revealed how PSs are exposed to different evolutionary pressures at different stages of infection, which fits with the CAT model. Over longer time periods, especially on the (sub)genotype level we found which PSs are more or less conserved. Changes in the variable PSs appeared to occur in reversible jumps and may indicate co-evolution of some sets of PSs, which would be interesting to investigate more closely in the future. Additionally, there appeared to be a pattern of a small number of conserved PSs when profiles were compared between genotypes: a few at the 5'-end of the pgRNA, a few at the 3'-end, and some dispersed. This pattern of conservation may provide insight into the functioning of the virus, which will be further investigated in Chapter 5. The conclusions here are only directly applicable to HBV. However, highly conserved PSs are hypothesised to represent crucial additional functions, which themselves are likely conserved in related viruses. Therefore, the pattern of highly conserved PSs may be shared to an extent between closely related viral species that also share other functional elements such as WHV in the same genus or even DHBV in the same family. In how far the evolution of PSs is specific to HBV and related

viruses or a common feature of PSs remains to be investigated. For that, further longitudinal studies in other viruses would be necessary. In Chapter 6 the phylogenetic method will be applied to ssRNA bacteriophages MS2 and BZ13, which will elucidate PS phylogenetics of viruses with known PS affinity tiers and that does not reverse-transcribe its RNA within the capsid and that therefore present larger PS ensembles. Moreover, it gives the chance to compare PS evolution on an even higher level, namely between species of the same genus.

Chapter 5

Nucleation of Assembly in HBV

In silico modelling of viral capsid assembly indicates the selective advantage of few high affinity packaging signals (PSs) that trigger the process (Dykeman et al., 2014). This allows assembly and packaging to proceed efficiently in an ordered fashion and avoids trapping in dead-end intermediates. If there are more high affinity sites, trapping occurs as packaging is initiated from several places. On the other hand, if there are no such sites, initiation is slow and not localised to a specific site. Biologically it makes sense that one region in the genome is the designated first contact point for capsid protein (CP), forming an assembly nucleation complex. This then sets off a cascade for further binding of CPs to PSs as well as of CPs to each other. The role of this nucleation complex involves switching from translation or replication to packaging in a single-stranded RNA (ssRNA) virus as in MS2, or possibly setting the geometry of the capsid like a crystal seed when the first few CPs bind together.

In this chapter a putative nucleation complex in hepatitis B virus (HBV) will be investigated and characterised. Due to the position of this region, I will propose a novel hypothesis for the interplay of packaging and reverse transcription in HBV. This will be expanded on and applied to predict PSs in other viruses of the *Hepadnaviridae* family.

5.1 Evidence for a Nucleation Complex in HBV

Both potential roles of a nucleation complex are relevant to HBV. As explained in detail in Section 3.1.2 the pre-genomic RNA (pgRNA) of HBV also functions as messenger RNA (mRNA) for core protein (HBcAg) and DNA polymerase (Pol). It is therefore covered in translating ribosomes at all times after leaving the nucleus. Binding of HBcAg to the RNA especially at a number of stem-loop (SL) sites requires freeing it from ribosomes. In the discussion of Chapter 3, I mentioned the importance of ε and Pol for this purpose. It has been suggested that with this ε is performing similar functions as mentioned above, but this ignores the potential importance for setting the geometry of the HBV capsid *in vivo*. The structure of capsids is observed as either T=3 or T=4 icosahedral symmetry in experiments (see Section 3.1.2). The ratio of T=4 to T=3 particles in virus isolated from infected human liver cells is 13:1, i.e. 93% of capsids exhibit T=4 symmetry. When expressed in *Escherichia coli (E. coli)* this drops to 3:1 (75%) and further decreases to 1:1 when the protein carboxy terminal end is truncated after amino acid 149 (Kenney et al., 1995). The truncation prevents the protein from binding nucleic acid. The findings by Kenney et al. (1995) were further corroborated by other work that found that almost all capsids isolated from infected individuals exhibit T=4 symmetry (Roseman et al., 2005). This strongly indicates that it is the T=4 isomorph that is infectious. In *in vitro* re-assembly experiments at 20°C, the proportion of T=4 capsids is considerably higher when truncated capsid proteins are used. Instead of 50%, 95% of capsids display T=4 symmetry with the protein truncated at amino acid 149. Further truncation results in increasing percentages of T=3 capsids (Zlotnick et al., 1996). The carboxy terminal end is arginine rich and has been shown to be essential for RNA binding and DNA synthesis (Gallina et al., 1989; Nassal et al., 1990; Nassal, 1992). Whilst this tail appears to be dispensable for capsid formation in an *E.*

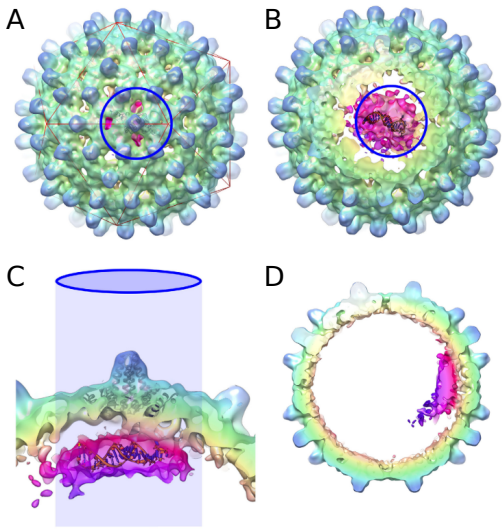

FIGURE 5.1: **CryoEM structure of HBV capsid with inner density.** (A) Outer view of HBV T=4 capsid. Within the blue ring the inner density corresponding to PS1 copies bound to the capsid can be seen in pink through the holes in the structure. (B) A slightly opened view of (A) with the inner density better visible and one SL fitted. (C) A side view of a fitted SL. (D) A cross-section of the capsid with the inner RNA density shown in pink on the side. The images were taken from Figure 5 in Patel et al. (2017).

coli expression system, its presence and the subsequent presence of RNA inside assembled capsids correlates with a higher stability of the particles (Birnbaum and Nassal, 1990).

Experimental validation of PS1-triggered re-assembly of CPs in HBV also included cryo-electron microscopy (cryo-EM) of the resulting virus-like particles (VLPs) (Patel et al., 2017). By using asymmetric reconstruction instead of icosahedral averaging our collaborators were able to identify localised density within the capsid that would correspond to PS1 oligonucleotides (oligos) in contact with CP (Figure 5.1). 2–4 PS1 SLs could be fitted into this density. Note that in this experiment sufficient copies of PS1 oligos were present to contact each CP. If HBV assembly required PS-CP contacts all over the capsid, as is seen in some ssRNA viruses such as MS2, the density would be uniformly distributed under the contact sites. The fact that there is a single spot of density indicates that only a small, critical set of PSs make initial contact forming a nucleation com-

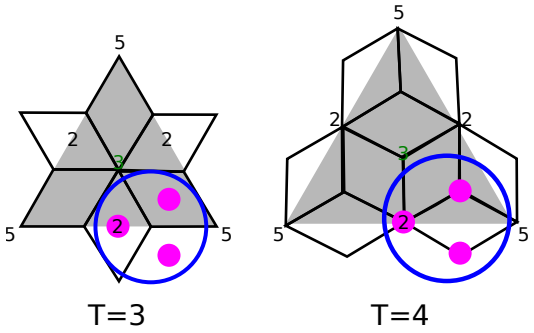

FIGURE 5.2: **Placement of density observed in cryoEM in tiling of T=3 and T=4 capsid.** One triangle (grey) of the icosahedral structure with 5-, 3-, and 2-fold axes of symmetry is shown for a T=3 (left) and T=4 (right). The CP dimers are visualised with black parallelograms. The positions of the inner densities observed in the cryoEM are shown as pink circles surrounded by a blue circle as in Figure 5.1. In a T=3 symmetry the density points would be under the dimers whereas in a T=4 symmetry they are between dimers.

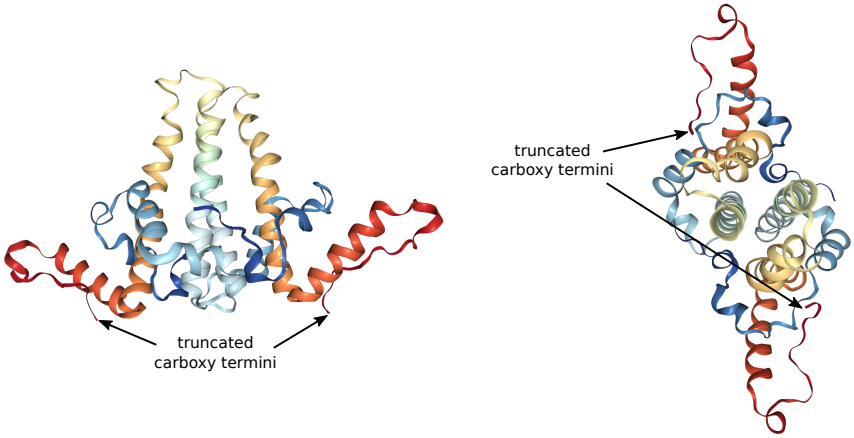

FIGURE 5.3: **Crystal structure of HBV capsid protein dimer.** RSB PDP structure 4BMG of a capsid protein dimer is shown (Rose et al., 2018). The crystal structure was resolved by Ferguson et al. (2013) at 3.0Å resolution. The protein is truncated at amino acid 149 and is missing the carboxy-terminal tail, which extends into the interior of the assembled capsid. The carboxy-terminal ends from which the tail would continue are marked with arrows. They are located at the long sides of the dimer. A view from the side (left) and from the top (right) is given.

plex and thereby triggering the assembly process. These would have to be in close proximity on the pgRNA. Additional weaker sites may provide further points of contact, but are not necessary and more variable as they are not seen after image reconstruction.

Indication of how these few PS-CP contacts may be involved in ensuring a

T=4 geometry comes from their placement inside the capsid. Looking from different angles, collaborators were able to pinpoint the approximate locations of the contacts relative to the symmetry axes. One was along the 2-fold and the others towards the direction of the 3-fold axis from there (personal communication) as shown schematically in Figure 5.2. In T=4 this places the contacts between the dimers, which means they can stabilise interdimer interactions (Figure 5.2, right). Applying the same geometry to T=3 would place them underneath the dimers (Figure 5.2, left). In a virus such as MS2, which requires a conformational switch of some CPs, this type of interaction would be beneficial. However, no such switch happens in HBV, and there is no role for PSs in dimer formation and such contacts would therefore not be helpful.

The truncated carboxy terminus in the crystal structure of HBcAg can be seen on the long opposite sides of the dimers (Figure 5.3). It can therefore be expected that the missing carboxy-terminal tail would extend into the capsid on the sides of the dimers. Due to their inherent positive charge, they would repel each other and thereby benefit from contacts with negatively charged RNA to neutralise the charges and stabilise interdimer contacts. This effect is paired with sequence-specific interactions to define the nucleation complex. Comparing with the latest, highest resolution electron microscopy structure of the entire capsid made up of truncated HBcAg, the carboxy-terminal ends now appear to extend into the space in the middle of hexamers and pentamers (Figure 5.4). This would indicate that stabilising RNA contacts would be most needed in the middle of hexamers and pentamers. Since none of the available structures include the carboxy-terminal tails, it can, of course, only be speculated how exactly they sit within the capsids. They are very flexible and may extend in either direction to find contact with RNA.

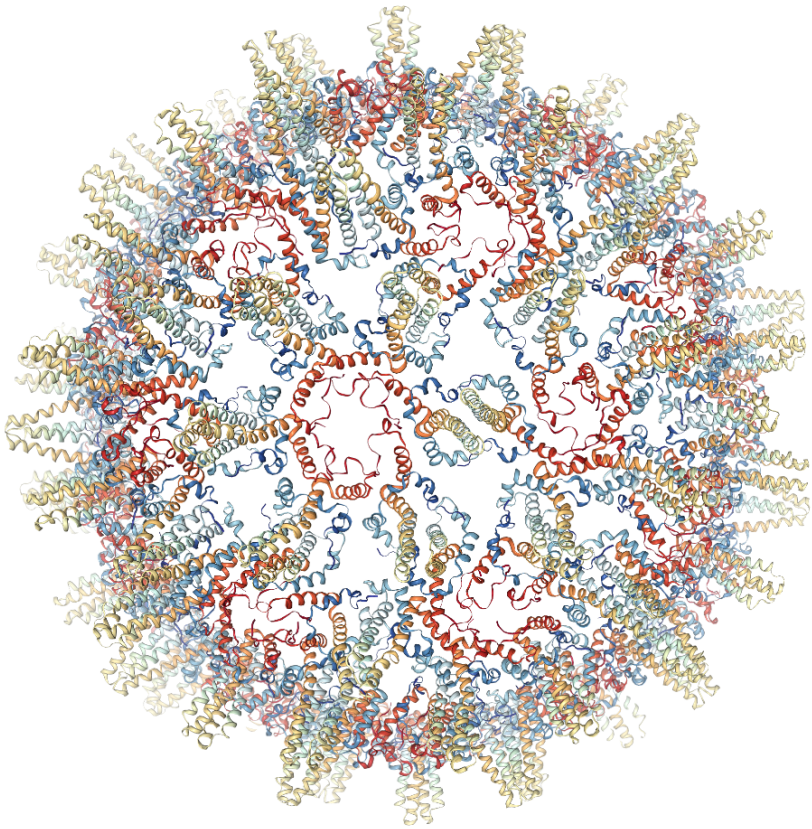

FIGURE 5.4: **Electron microscopy structure of whole HBV capsid.** RSB PDP structure 1QGT of the complete capsid is shown (Rose et al., 2018). The structure was resolved using electron microscopy by Böttcher and Nassal (2018) at 2.66Å resolution. The carboxy-terminal tail is missing from the structure as it is too disordered to resolve. The single capsid proteins are coloured in the same gradient meaning that the carboxy-terminal ends are red in all molecules in the capsid whilst the amino-terminal ends are dark blue.

5.2 Identification of Conserved PS Groups

The phylogenetic analysis of HBV based on PS profiles carried out in Chapter 4 already provided some insights into conserved PSs. Comparing the compact PS profiles of different (sub)genotypes revealed several groups of PSs that may be involved in a nucleation complex. To be considered a group the SLs had to occur in at least 50% of the HBV (sub)genotypes in Table A.3 and be within 100 nucleotides of each other. This ensured that no fragment was longer than 150 nucleotides, a reasonable length for structure prediction and experiments. The window choice excluded only one putative PS around nucleotides 820–870 whilst its respective putative partner was included in a downstream fragment. Whilst the cryo-EM fit with 2–4 PSs, only groups of two and one of three were found to be conserved. Further analysis was carried out on these sites to determine which would be viable candidates.

5.2.1 Fragment Analysis

The seven identified sites were first investigated for their ability to fold into the PS pairs on the same RNA fragment (Table 5.1). This was necessary to ensure they could be tested experimentally by collaborators. Therefore, the sequence of the laboratory strain (NC_003977.1) was used in the first instance.

Fragments of various lengths containing the respective putative PSs were folded in Mfold (Zuker, 2003). Since the fragments were taken out of the context of their neighbouring sequence, which may affect local folding, suboptimal structures up to 500% worse than the minimum free energy (MFE) structure were also taken into consideration. Four of the fragments folded into only one SL with an RGAG motif (Figures 5.5, 5.6, 5.7, and 5.8) whilst two did not fold into any at the given range of suboptimality (Figure 5.9). Only one fragment folded into two RGAG SLs and showed several alternative SLs for one PS (Figure 5.10). Interestingly, this was the fragment that included PS1 (nucleotides 1722–1788).

TABLE 5.1: Fragments with at least two putative PSs.

Fragment	Position in pgRNA	Sequence in the lab strain
1	127–255	5'-CCCGTATAAAGAATTTGGAGCTTCTGTGGAGTTACTC TCTTTTTTGCCTTCTGACTTCTTTCCTTCTATTCGAGATC TCCTCGACACCGCCTCTGCTCTGTATCGGGAGGCCTTAGA GTCTCCGGAACA-3'
2	985–1083	5'-ATTCTATATAAGAGAGAAACTACACGCAGCGCCTCAT TTTGTGGGTCACCATATTCTTGGGAACAAGAGCTACAGCA TGGGAGGTTGGTCTTCCAAACC-3'
3	1213–1314	5'-CTGGCCAGAGGCAAATCAGGTAGGAGCGGGAGCATTT GGTCCAGGGTTCACCCCACCACACGGAGGCCTTTTGGGGT GGAGCCCTCAGGCTCAGGGCATATT-3'
4	1390–1494	5'-CCTCTAAGAGACAGTCATCCTCAGGCCATGCAGTGGA ACTCCACAACATTCCACCAAGCTCTGCTAGATCCCAGAGT GAGGGGCCTATATTTTCCTGCTGGTGG-3'
5	1576–1697	5'-ACCGAACATGGAGAGCACAACATCAGGATTCCTAGGA CCCCTGCTCGTGTTACAGGCGGGGTTTTTCTTGTTGACAA GAATCCTCACAATACCACAGAGTCTAGACTCGTGGTGGAC TTCTC-3'
6	3030–3104	5'-CGTAGCATGGAGACCACCGTGAACGCCCACCAGGTCT TGCCCAAGGTCTTACACAAGAGGACTCTTGGACTCTCA-3'
7	3150–3231	5'-TTTAAAGACTGGGAGGAGTTGGGGGAGGAGATTAGGT TAAAGGTCTTTGTACTAGGAGGCTGTAGGCATAAATTGGT CTGTT-3'

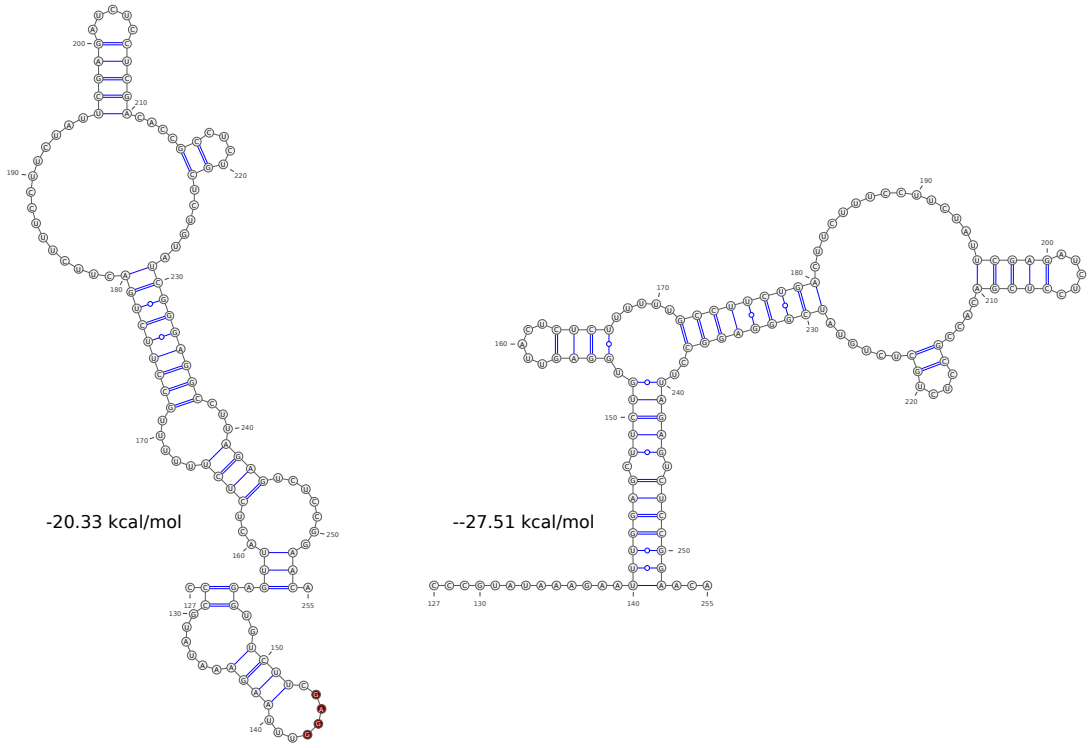

FIGURE 5.5: **Structures in fragment 1.** The fragment was folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. The MFE structure (right) and the lowester energy structure containing the RGAG motif marked in red (left) are shown with their respective energies. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

PS2 (nucleotides 2602–2633) and PS3 (nucleotides 2776–2798) are too distant to be involved and do not appear to form the nucleation complex. To further study potential pairs, the fragment was shortened at the 3' end to just the end of the second putative PS.

The identified region, named LS1, is located in the part of the X gene that does not overlap with another gene. This would mean that it is less mutationally restricted. Its multiple sequence alignment (MSA) generated in ClustalΩ (Goujon et al., 2010; Sievers et al., 2014) was examined more closely for conservation (Figure 5.11). There was a high degree of nucleotide conservation between the (sub)genotypes, which is consistent with it playing an important functional role in addition to coding for X protein (HBxAg).

Upon closer examination, it was found that LS1 contained three overlapping putative PSs that could take on the role of PS1 together with one highly conserved

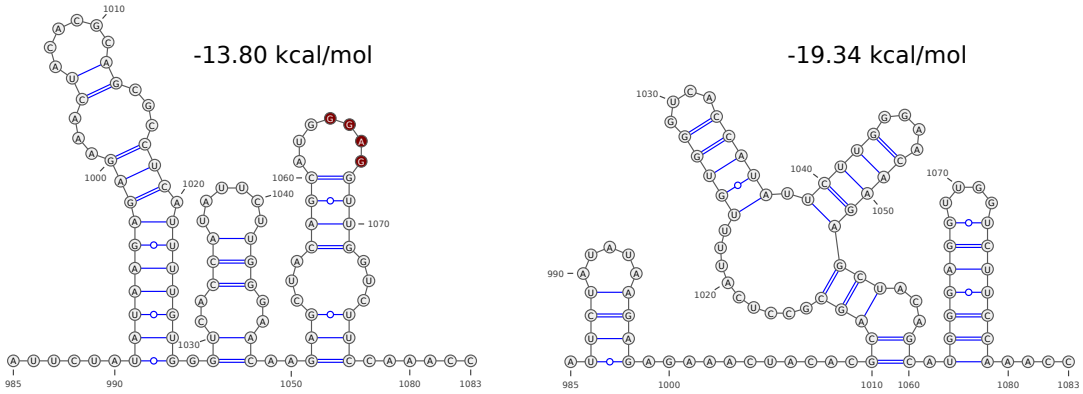

FIGURE 5.6: **Structures in fragment 2.** The fragment was folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. The MFE structure (right) and the lowester energy structure containing the RGAG motif marked in red (left) are shown with their respective energies. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

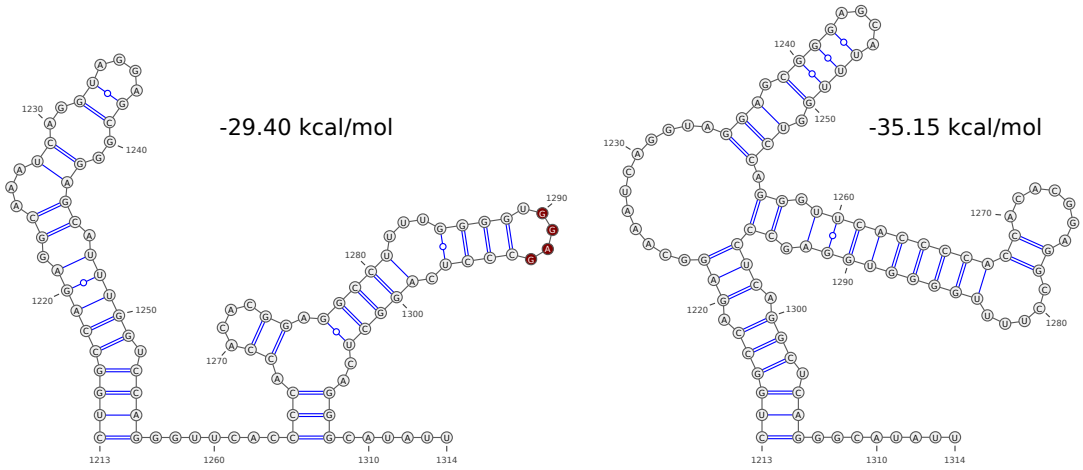

FIGURE 5.7: **Structures in fragment 3.** The fragment was folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. The MFE structure (right) and the lowester energy structure containing the RGAG motif marked in red (left) are shown with their respective energies. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

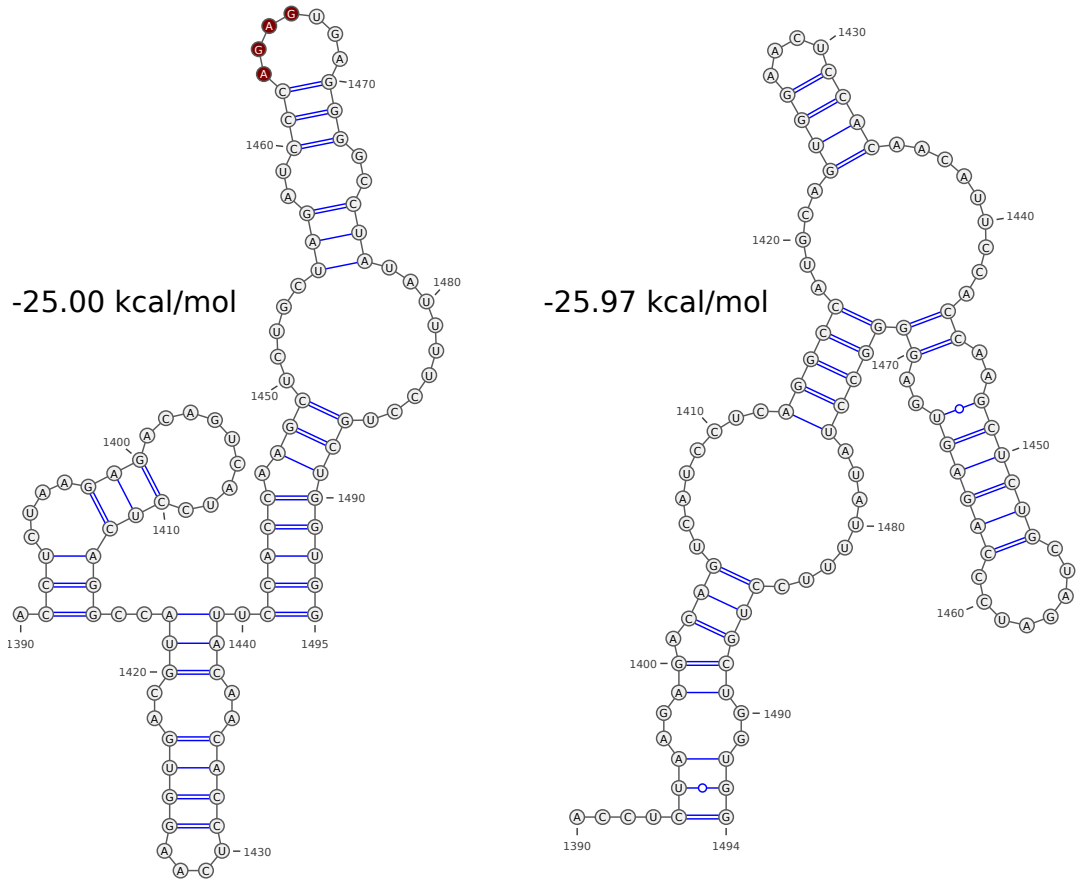

FIGURE 5.8: **Structures in fragment 4.** The fragment was folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. The MFE structure (right) and the lowester energy structure containing the RGAG motif marked in red (left) are shown with their respective energies. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

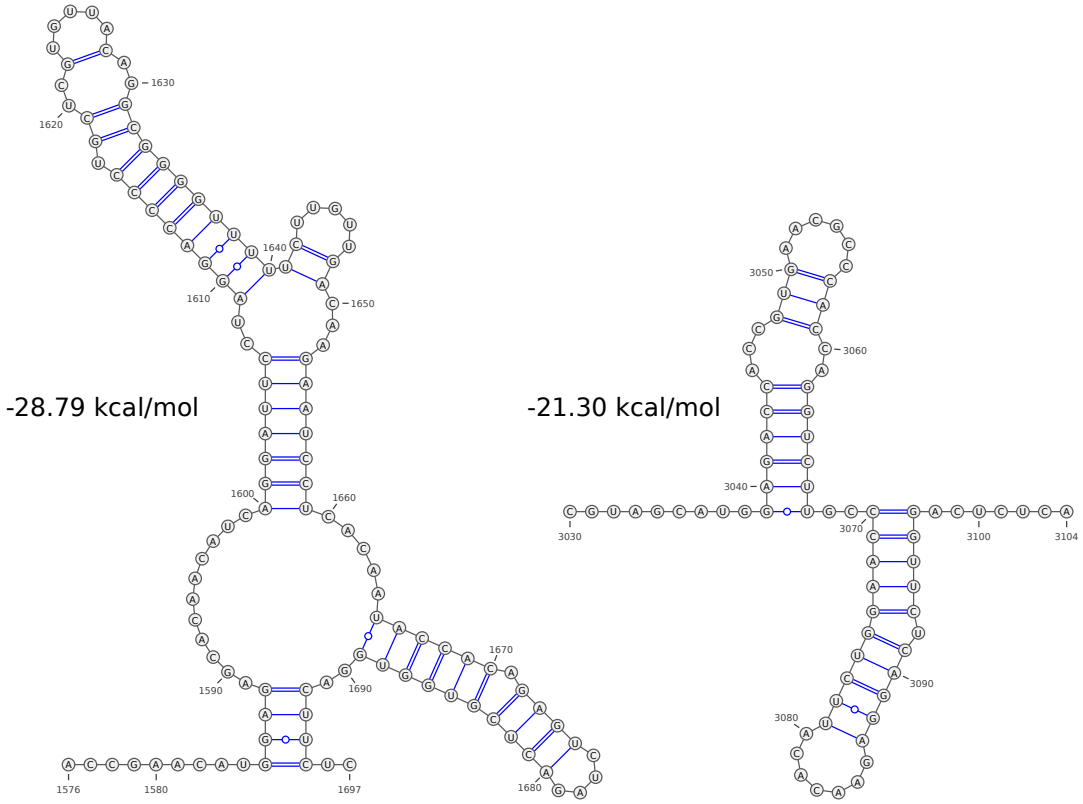

FIGURE 5.9: **Structures in fragments 5 and 6.** The fragments were folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. No structures containing the RGAG motif were found for either fragment. The MFE structures for fragment 5 (left) and fragment 6 (right) are shown with their respective energies. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

but lower stability secondary PS downstream (Figure 5.12). All of the alternatives show co-mutations of nucleotides to preserve base-pairing, e.g. A-U to G-C or G-U to G-C, with the exception of genotype B strains. Most genotype B sequences cannot fold into the first SL as shown in Figure 5.12, i.e. PS1 as described above, due to a change in two nucleotides in the upper helix. Instead they can form the third SL. To confirm that these SLs can function together as PSs, shorter fragments truncated 5' and 3' to only include the two respective SLs, could be used for capsid re-assembly experiments. The structures for the three combinations of "PS1" with the secondary PS as well as a version with three SLs together with possible stabilising mutations are provided in the Appendix (Figure A.1).

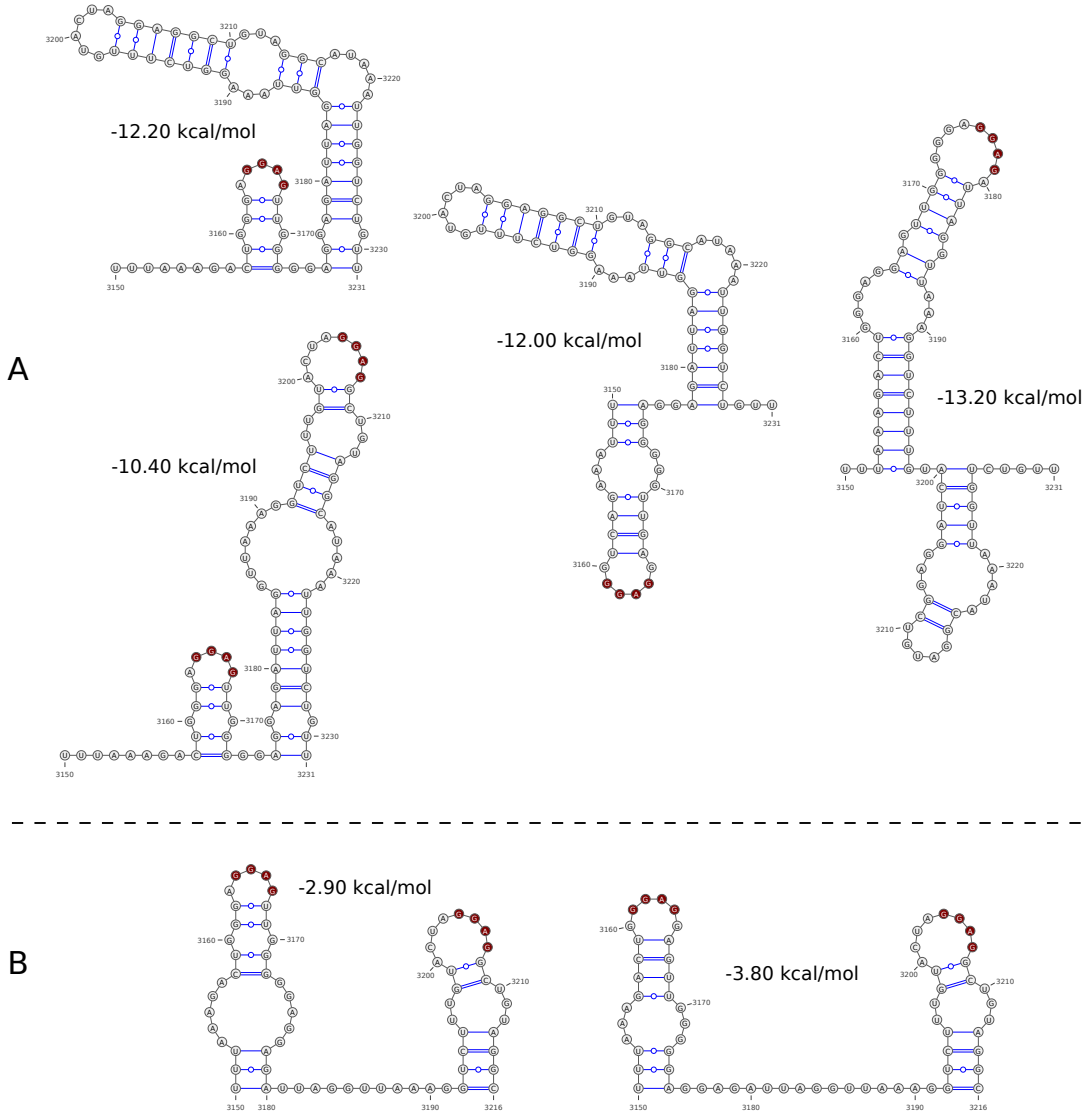

FIGURE 5.10: **Structures in fragment 7.** The fragment was folded in Mfold (Zuker, 2003) allowing for up to 500% suboptimality. Structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape. The RGAG motif is highlighted in red. (A) Stable structure with an RGAG were found with three alternative SLs: in AGGAG (left), in GGGAGG (middle) or in GGGAGGAGA (right) apical loop. The AGGAG loop also occurred in combination with another RGAG SL (left, bottom). (B) Two alternative RGAG SL pairs in a shortened fragment 7.

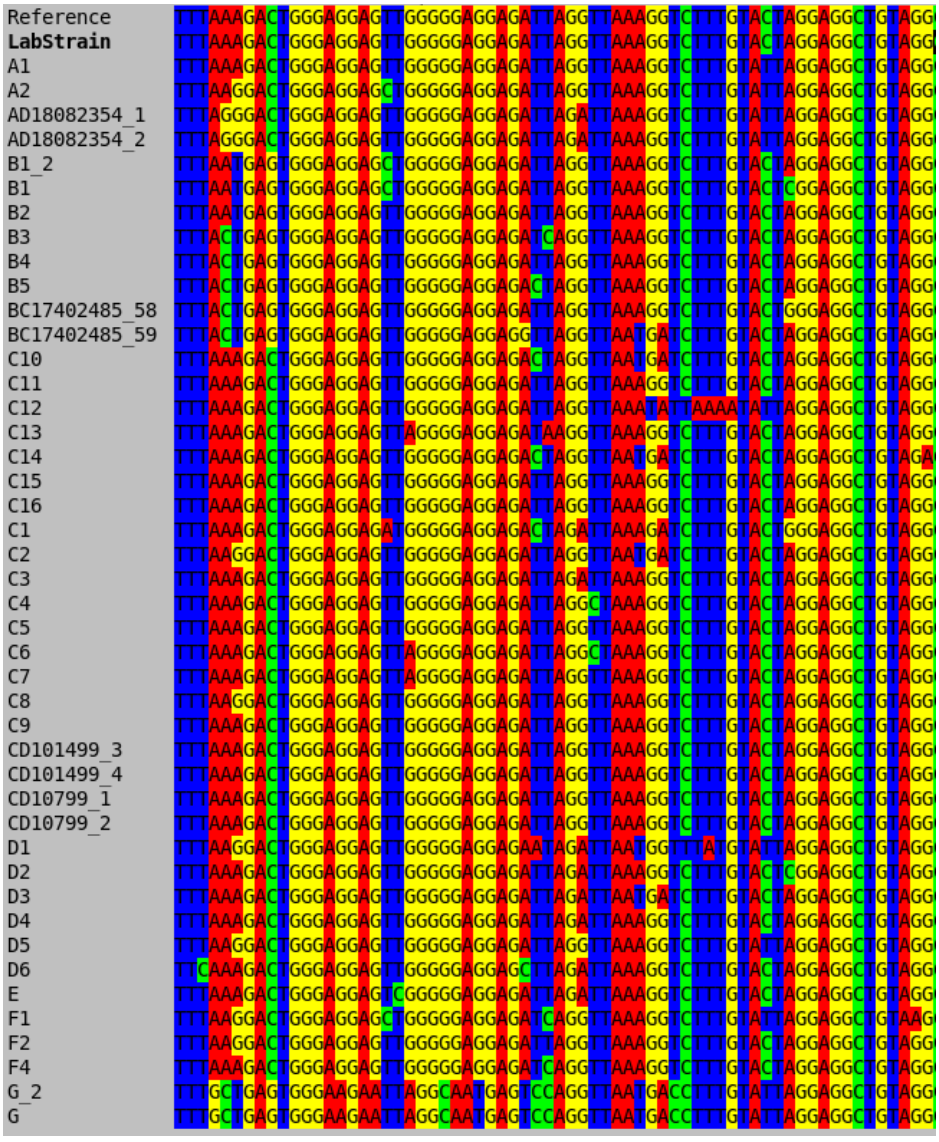

FIGURE 5.11: **Multiple sequence alignment of LS1 region.** The sequences were aligned using ClustalΩ (Goujon et al., 2010; Sievers et al., 2014). The alignment is visualised in SeaView (Gouy et al., 2010). The sequence of fragment 7 after shortening is shown for the NCBI reference strain, the laboratory strain, and a set of reference genomes of different (sub)genotypes (see Section 4.4).

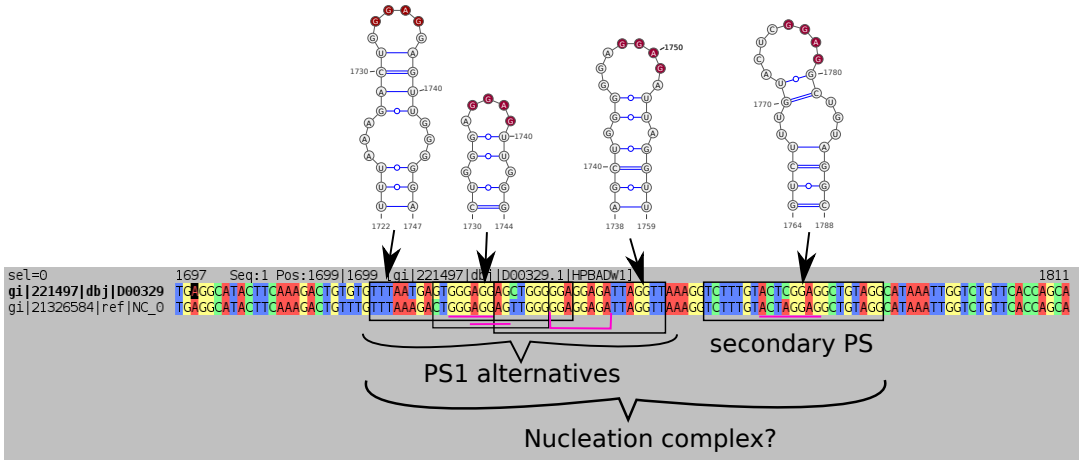

FIGURE 5.12: **Putative set of PSs involved in nucleation complex.** A section of an alignment of a genotype B (D000329) and the NCBI reference strain (NC_003977.2) of genotype D from positions 1697 to 1811 in SeaView is shown (Gouy et al., 2010). SL folds were generated with VARNA and are depicted above with the RGAG motif highlighted in red (Darty et al., 2009). Sequences involved in each SL are marked by black boxes and apical loops are underlined in pink. There are three alternative folds for PS1, each of which is preferred by different strains. The secondary PS is highly conserved among all strains but has a lower stability.

5.3 Knock-out of LS1 PSs in HBV

Thus far experimental validation of HBV PSs had been limited to *in vitro* re-assembly assays of isolated HBcAg overexpressed in *E. coli*. These naturally have their limitations. Not only is the protein expressed in a foreign cell system (prokaryote, rather than its natural eukaryotic host), the re-assembly of capsids also occurs without other components such as viral and host proteins present and at a lower temperature (20°C rather than 37°C). All of these factors may influence how the protein behaves and interacts with RNA. Therefore, we wanted to test the effect a knock-out of LS1 PSs would have on a fully replicating virus in a eukaryotic cell system. My task was to identify synonymous mutations that would ablate the above PSs whilst maintaining the coding sequence.

This endeavour proved to be non-trivial. As seen in Figure 5.13, the RGAG motif in PS1 encodes for W (UGG) and E (GAG). W is only encoded by UGG, whilst E is encoded by GAG and GAA. This only leaves the option to change

```
  1 M  A  A  R  L  C  C  Q  L  D  P  A  R  D  V  L  C  L  R  P
  1 ATGGCTGCTAGGCTGTGCTGCCAACTGGATCCTGCGCGGGACGTCCTTTGTTTACGTCCC

 21 V  G  A  E  S  C  G  R  P  F  S  G  S  L  G  T  L  S  S  P
 61 GTCGGCGCTGAATCCTGCGGACGACCCTTCTCGGGGTCGCTTGGGACTCTCTCGTCCCCT

 41 S  P  S  A  V  P  T  D  H  G  A  H  L  S  L  R  G  L  P  V
121 TCTCCGTCTGCCGTTCCGACCGACCACGGGGCGCACCTCTCTTTACGCGGACTCCCCGTC

 61 C  A  F  S  S  A  G  P  C  A  L  R  F  T  S  A  R  R  M  E
181 TGTGCCTTCTCATCTGCCGGACCGTGTGCACTTCGCTTCACCTCTGCACGTCGCATGGAG

 81 T  T  V  N  A  H  Q  I  L  P  K  V  L  H  K  R  T  L  G  L
241 ACCACCGTGAACGCCCACCAAATATTGCCCAAGGTCTTACATAAGAGGACTCTTGGACTC

101 S  A  M  S  T  T  D  L  E  A  Y  F  K  D  C  L  F  K  D  W
301 TCAGCAATGTCAACGACCGACCTTGAGGCATACTTCAAAGACTGTTTGTTTAAAGACTGG

121 E  E  L  G  E  E  I  R  L  K  V  F  V  L  G  G  C  R  H  K
361 GAGGAGTTGGGGGAGGAGATTAGGTTAAAGGTCTTTGTACTAGGAGGCTGTAGGCATAAA

141 L  V  C  A  P  A  P  C  N  F  F  T  S  A  *
421 TTGGTCTGCGCACCAGCACCATGCAACTTTTTCACCTCTGCCTAA
```

FIGURE 5.13: **Amino acid and nucleotide acid sequence of X protein.** The sequences for laboratory strain (NC_003977.1) are shown.

GGAGG to GGAAG. The possible alternative PS with apical loop AGGAG would thereby be changed to AAGAG, which is still an RGAG motif. Further changing the next E codon from GAG to GAA would result in AAGAA. Whilst neither GGAAG nor AAGAA are technically RGAGs any more, they are still very similar and from mutational experiments performed in Patel et al. (2017) they cannot be excluded as functional. Moreover, it is not straightforward to knock out these SLs by mutating the nucleotides in the helix of the structures. Whilst this may destabilise one putative PS structure, it can instead stabilise another, alternative SL with the RGAG motif. Further complications arise from the fact that the secondary PS (`GTCTTTGTACTAGGAGGCTGTAGGC`) largely overlaps with the *cis*-acting element ϕ (`CTAGGAGGCTGTAGGCA`). Changes to ϕ have disastrous effects on DNA minus-strand synthesis (Tang and McLachlan, 2002; Oropeza and McLachlan, 2007). To ensure replication competent virus, it must not be changed. Since this also includes the RGAG motif, an alternative method to mutating away the apical loop motif for knock-down was necessary.

5.3.1 Genetic Algorithm to Evolve PSs

A longer sequence fragment of 180 nucleotides was used with the intention of identifying mutations that could stabilise alternative structures making the PSs less favourable. A genetic algorithm was used to evolve this fragment to minimise the number of RGAG SLs. In a set of 2000 sequences, each was folded in a 40, 50 and 60 nucleotide sliding window and 100 structures were sampled from each window using Tfold in the partition function mode as described in detail in Chapter 2. The total number of structures with RGAG in either window was calculated. Note that the same SL could be counted several times between samplings, window sizes, and across windows. The number of RGAGs, therefore, does not represent the number of distinct SLs with that motif. However, one SL counted several times is more robust and more likely to occur in the fragment when used in an experiment. Theoretically, if every one of the sampled structures contained the motif in every overlapping window, the maximum number of RGAGs would be around 25,000. From a total of 2000 sequences, every round the 400 sequences with the lowest RGAG number were selected. These sequences were "evolved" in two ways to generate 1600 new sequences: (1) they were split into 10 fragments of 18 nucleotide length and randomly recombined preserving the correct order, and (2) random synonymous mutations were introduced into the recombinants at 1% of the codons. In order to preserve full function of the mutants, changes in the ϕ sequence were not permitted.

The method consisted of two programs: One fragmented the sequence into 40, 50, and 60 nucleotide windows, submitted those fragments to Tfold, and counted the total number of RGAG occurrences for each sequence (Algorithms A.10 and A.11). The second, given a set of 2000 sequences, a list of attributes (in this case RGAG number), and constraints sorted the sequences by their attributes, kept the 400 best ones, recombines them and introduced synonymous mutations at 1% of codons to generate 1600 new sequences (Algorithms A.12 and A.13). The pseudocodes for both are shown in the Appendix.

```
wt    AGGACUCUUGGACUCUCAGCAAUGUCAACGACCGACCUUGAGGCAUACUUCAAAGA
mut   ..............U.....................................G..

wt    CUGUUUGUUUAAAGACUGGGAGGAGUUGGGGGGAGGAGAUUAGGUUAAAGGUCUUUG
mut   U..CC.C.....G..U.....A..AC.A..C..A..A...C....G........C.

wt    UACUAGGAGGCUGUAGGCAUAAAUUGGUCUGCGCACCAGCACCAUGCAACUUUUUC
mut   ......................C.C.....U.....G.............C...

wt    ACCUCUGCCUAA
mut   ...AGC......
```

FIGURE 5.14: **Synonymous knock-down of RGAG PSs in LS1.** The top sequence shows the original, wildtype sequence (wt), whilst the bottom shows the result of 80 rounds of evolution to minimise RGAG stem-loops (mut). Changes in nucleotide sequence are marked with "^" underneath whilst identical nucleotides are replaced by a ".".

```
wt    AGGACUCUUGGACUCUCAGCAAUGUCAACGACCGACCUUGAGGCAUACUUCAAAGACUGU
mut   ..............A.................G...........U........G......

wt    UUGUUUAAAGACUGGGAGGAGUUGGGGGAGGAGAUUAGGUUAAAGGUCUUUGUACUAGGA
mut   C.U.....G...........A.....C.....A.............................

wt    GGCUGUAGGCAUAAAUUGGUCUGCGCACCAGCACCAUGCAACUUUUUCACCUCUGCCUAA
mut   ............................G.............................
```

FIGURE 5.15: **Synonymous knock-down of RRAG PSs in LS1.** The top sequence shows the original, wildtype sequence (wt), whilst the bottom shows the result of 80 rounds of evolution to minimise RRAG stem-loops (mut). Changes in nucleotide sequence are marked with "^" underneath whilst identical nucleotides are replaced by a ".".

Initially wildtype (wt) sequences had RGAG numbers of almost 10,000, which decreased to 17 after 80 rounds of evolution. No further reduction was possible. 25 nucleotide changes were present in the evolved sequence (Figure 5.14).

Taking this a step further, the algorithm was repeated for RRAG. Now the wildtype sequence counted 13478 RRAG SLs. After 80 rounds of evolution the number of motif loops plateaued and did not decrease further. The best evolved sequence still had 904 RRAG SLs across all windows. As opposed to the above, which resulted in 25 changes, this evolved sequence only had 11 nucleotide changes compared to the wildtype sequence (Figre 5.15). These variations are available for experimental testing.

5.4 A Proposed Double Role for ϕ

PS1 (nucleotides 1722–1747, nucleotides 1730–1744, or nucleotides 1738–1759) lies downstream of DR2 (nucleotides 1590–1600) and upstream of ϕ (nucleotides 1769–1791), the 3' DR1 (nucleotides 1824–1835) and ε (nucleotides 1847–1907) in pgRNA. Its potential partner PS (ACUAGGAG; nucleotides 1761–1790) (see Figure 5.12) highly overlaps within ϕ. This PS is from now on referred to as PSϕ. There is an almost complete overlap between the putative PS and the *cis*-acting element. Earlier I mentioned that PSs may not interfere with other RNA functions; thus, disqualifying this SL from being a PS. However, PS-mediated assembly and reverse-transcription do not happen in parallel but rather sequentially. This suggests that PSs may play another role in regulating the timing of reverse transcription. I am thus proposing the following model (Figure 5.16):

Some incarnation of PS1 and PSϕ form on the pgRNA and facilitate capsid assembly and RNA packaging. Interaction of CP bound by PSs with CP bound by the ε-Pol complex would bring the 5' and 3' end of pgRNA in close proximity. After packaging, PSϕ melts exposing ϕ and allowing it to interact with ε and ω. This in turn melts ε and allows for reverse-transcription to start. ϕ being unavailable to bind ε until after packaging of the pgRNA due to it being a CP-binding SL prevents premature DNA minus-strand synthesis. Genome replication can only commence after packaging. This hypothesis is supported by PSϕ being a highly conserved SL even in the most phylogenetically distant genotype G and having lower stability, making it easier to melt and expose ϕ following assembly. If the double function of the ϕ region is indeed true, it provides a means to predict PSs in other viruses with known ϕ without the need for systematic evolution of ligands by exponential enrichment (SELEX).

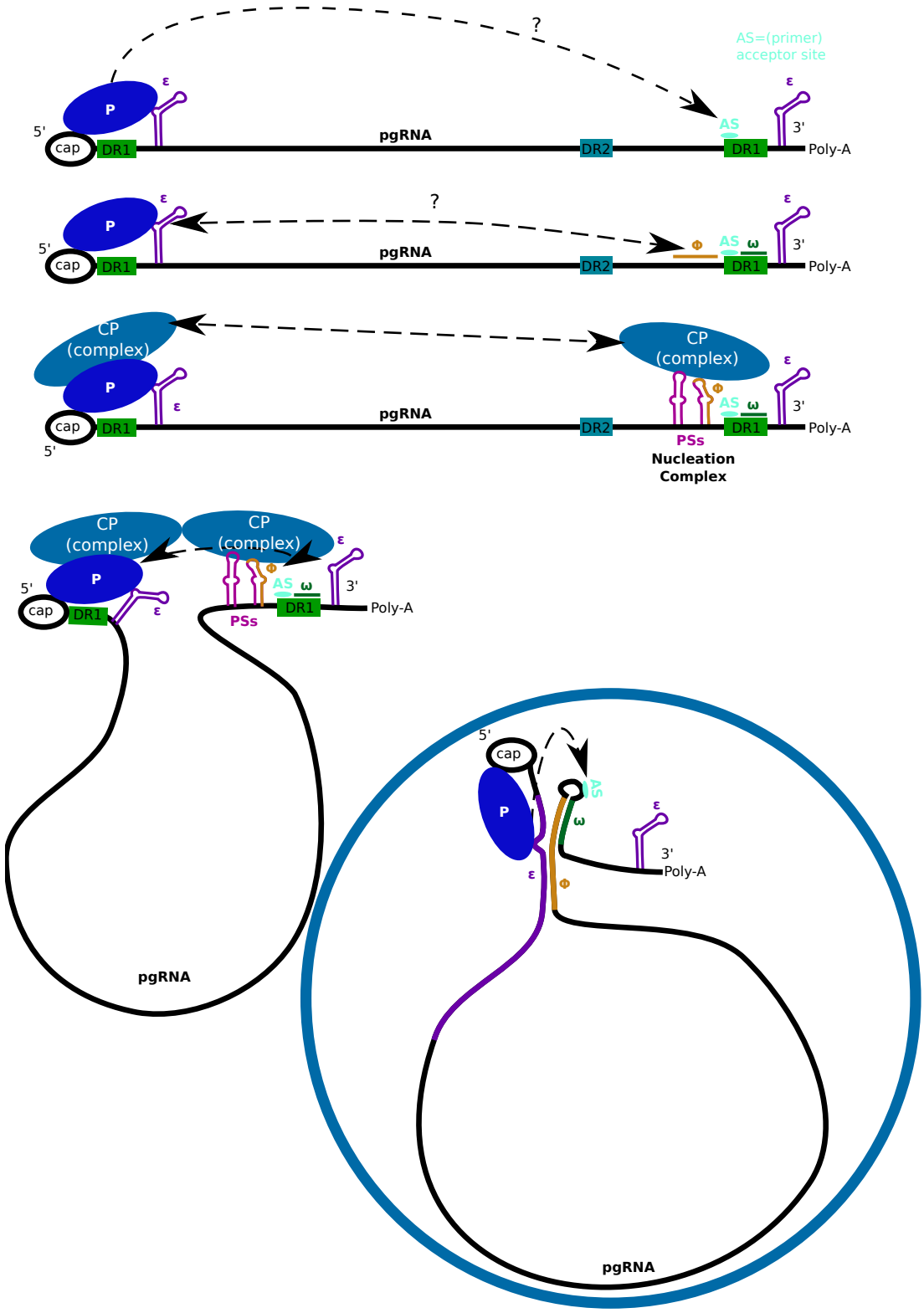

FIGURE 5.16: **Regulation of HBV reverse transcription by PSs.** In order to initiate DNA minus-strand synthesis, polymerase (blue ellipse "P") needs to bind to the primer acceptor site (AS, light blue). This is aided by pgRNA circularisation by interaction of ϕ with ω and 5' end ε. However, how these cis-acting elements come in contact is not known. I propose that PS (pink SLs) interaction with CP (grey-blue ellipse "CP")brings both ends of the pgRNA in close enough proximity so that when the ϕ PS melts it can interact with ε. The presence of ϕ PS thereby also prevents premature DNA synthesis.

5.5 Nucleation Complex Packaging Signals in Ancient HBV strains

TABLE 5.2: Accession numbers, approximate ages, and geographic locations for the ancient HBV strains.

Sample name	Accession number	Age in years	Origin
RISE563	LT992443.1	4488	Germany
DA222	LT992454.1	1167	Kazakhstan
DA195	LT992441.1	2645	Hungary
DA51	LT992444.1	2297	Kyrgyzstan
RISE254	LT992459.1	4009	Hungary
DA119	LT992440.1	1567	Slovakia
RISE386	LT992448.1	4188	Russia
DA27	LT992439.1	1610	Kazakhstan
DA29	LT992438.1	822	Kazakhstan
DA45	LT992442.1	2120	Mongolia
RISE387	LT992447.1	4282	Russia
RISE154	LT992455.1	3851	Poland

Nucleation complex PSs, especially if they are important for other functions as suggested by the ϕ hypothesis described above, have likely evolved a long time ago. They are therefore expected to be conserved not only among different genotypes found today but also in ancestral sequences. One way of testing this hypothesis is to consider ancient viral strains. Since HBV has infected humans for many thousand years, traces of ancient viruses can sometimes be found in archaeological samples. The twelve strains used here were published in Mühlemann et al. (2018). They were isolated from human teeth at different sites and from different time periods (Table 5.2). Due to the age of the samples, they were not perfectly preserved making sequencing difficult. Some strains are missing large parts of their sequence and most contain ambiguous bases to some extent. It was therefore not possible to use them for phylogenetic analysis in Chapter 4. However, fortunately the sequences of LS1 as shown in Figure 5.11 are mostly

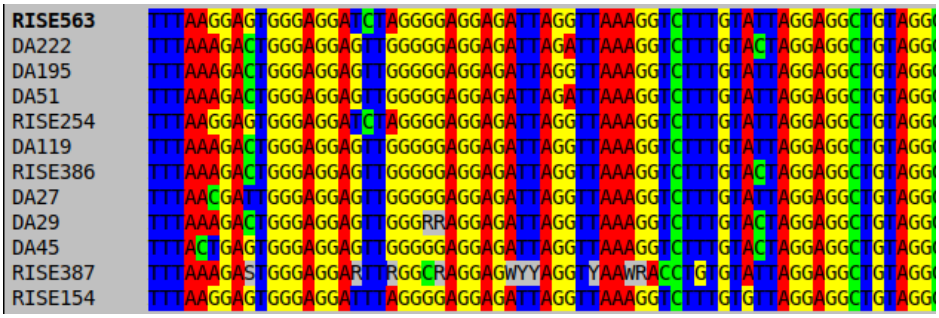

FIGURE 5.17: **Multiple sequence alignment of ancient HBV strains in LS1 region.** The sequences published in Mühlemann et al. (2018) were aligned using ClustalΩ (Goujon et al., 2010; Sievers et al., 2014). The alignment is visualised in SeaView (Gouy et al., 2010). The same region as in Figure 5.11 is shown.

resolved. A MSA of the twelve ancient HBV strains in the LS1 region is shown in Figure 5.17. Interestingly, not only the ϕ sequence but also the purine-rich regions that form the nucleation complex PSs are mostly conserved. All strains except for RISE387 could form some variation of the PS SL when folded in 60 nucleotide sliding windows (not shown). Since the folding algorithm cannot handle ambiguous bases, all windows containing such bases are skipped. This results in the lack of folds in this region for RISE387. These results show that the LS1 region and the PSs in it are highly conserved even for thousands of years. This provides further evidence for the functional importance of this region.

5.6 Prediction of PSs in Other *Hepadnaviridae*

The viral family *Hepadnaviridae* encompasses two genera: *Orthohepadnaviruses* and *Avihepadnaviruses*. More recently similar viruses were found in fish and amphibians (Hahn et al., 2015; Dill et al., 2016); however, they are not very well studied and will not be considered further here. HBV and other mammalian viruses belong to the *Orthohepadnaviruses*, whereas avian viruses such as duck hepatitis B virus (DHBV) and heron hepatitis B virus (HHBV) are members of the *Avihepadnavirus* genus. DHBV is often used as a model for HBV due to the many similarities. Most *cis*-acting elements known in HBV are also present

in DHBV including DR1, DR2 and ε. A functional element as important as ϕ is thus very likely to also be conserved. Tang & McLachlan have originally proposed a ϕ region for DHBV at nucleotides 2521–2542 directly upstream of DR1 (nucleotides 2543–2554) (Tang and McLachlan, 2002). In light of later findings regarding ω this region is improbable and was shown to not be involved in reverse transcription. Deletions of part of the region did not affect minus-strand DNA ((-)DNA) levels (Maguire and Loeb, 2010). Maguire & Loeb concluded that ϕ was in fact not conserved. However, they did not consider that the Tang & McLachlan may have identified the wrong region, which is much more probable considering the role ϕ plays in bringing together the Pol binding site in the 5' ε bulge and the primer acceptor site in the 3' DR1. Since at this point ϕ was not known in *Avihepadnaviruses*, it was necessary to first identify it before PSs could be predicted.

5.6.1 Annotation of *Hepadnaviridae* Genomes

Assuming that a region like ϕ exists in *Avihepadnaviruses* and other *Hepadnaviridae* that would interact with both 5' ε and a region downstream of 3' DR1 to bring both pgRNA ends into close proximity for primer translocation, I set out to identify base-pairing interactions between the pgRNA ends in these viruses. Apart from DHBV, woodchuck hepatitis virus (WHV), and HBV most of *Hepadnaviridae* are not well annotated: the positions of pgRNA start and end, DR1, DR2, and ε are not easily accessible. To have more viral species to corroborate putative ϕ and ω positions that could be identified in DHBV, these regions were annotated in all *Hepadnaviridae* species with complete published genomes.

5.6.1.1 Direct Repeats and pgRNA Start and End Positions

Since all *Hepadnaviridae* infect animals, they are transcribed in eukaryotic cells using host machinery. Their pgRNA would, therefore, have strong similarity to eukaryotic mRNA, which is especially useful for determining pgRNA start and end

sites. The canonical poly-adenylation (poly-A) signal is AATAAA (Levitt et al., 1989); however, HBV uses a slightly different motif, TATAAA (Simonsen and Levinson, 1983). Although the RNA does not directly end here, this corresponds to the position given in HBV and DHBV as end sites in the literature and was thus used for the other viruses as well. Despite encoding for several proteins all viral mRNAs, including the pgRNA, use the same poly-A site (Tiollais et al., 1985). The sequences were thus searched for ATAAAGAA, which unambiguously identified pgRNA end positions in all viruses (Table 5.3).

The transcription start site is expected to be about 30 nucleotides downstream of a TATA box. The TATA box is a *cis*-acting element that acts as promoter in many eukaryotic genes by recruiting transcription factors, which in turn recruit RNA polymerase II. In HBV the TATA box for the pgRNA is CATAAATT (Quarleri, 2014) and very similar sequences were identified in other mammalian viruses: CATAAAT(T/G). In DHBV it is TATATA and such sequences were found in all other avian viruses. However, the actual pgRNA start site is expected about 30 nucleotides downstream of the TATA box. It was identified based on the known sequences in DHBV and HBV due to high sequence identity within each genus: AAGA(A)TTACA in avian and ATCTTTTT in mammalian viruses (Table 5.3).

The direct repeats, DR1 and DR2, are two identical sequence stretches of 11 nucleotide length. DR1 is, moreover, located within the terminal repeat region, i.e. the 100 nucleotide sequence between pgRNA end and start position, which is on both ends of the RNA. In order to be able to easily check genomic positions, the FASTA files were altered to display the entire sequence in one line (lin_gen). A regular expression was utilised to search the sequences for positions that match TXCXCCXXTXX, where X is any nucleotide, and then comparing them to find identical matches in the correct places. Below is the short code that searches the linearised FASTA file for lines not containing (-v, for invert match) '>', i.e. the FASTA header. This line is then forwarded to a search for the above motif, which

prints only the matched sections (-o) and the genomic position (-b, byte offset).

```
> egrep -v '>' [lin_gen] | egrep -bo 'T.C.CC..T..'
```

This approach was sufficient for all viruses except for WHV. Whilst a DR1 was found in the correct place with the same sequence as in HBV, no matching DR2 was found in the initial search. The search sequence was therefore altered to allow variability at the 5' end whilst preserving the 3' end, which yielded a match (Table 5.3).

```
> egrep -v '>' [whv lin_gen] | egrep -bo '..C.CC..TGC'
1718:GTCACCTGTGC
1940:TTCACCTGTGC
```

5.6.1.2 Epsilon

ε is also located within the terminal repeat region as explained above for DR1. Comparing the published ε in HBV, WHV, DHBV, and HHBV (Kramvis and Kew, 1998), conserved structure and sequence elements were identified. Since for this element both sequence and structure is important, the genomic sequences were folded. RNA fragmentation, folding, and structure processing was performed as described in Chapter 2. Briefly, a window of 90 nucleotides was slid along the sequences in increments of 1 nucleotide. Each window was folded using an in-house implementation of the Mfold algorithm in partition function mode and the individual SLs extracted together with numerical information about their positions and stabilities. Next, the SLs were merged across windows, meaning that a SL in the same position was only kept once together with its highest stability across overlapping windows. The processed information for each SL included its structure in Vienna format, i.e. base-pairs as matched brackets and single-stranded nucleotides as dots, as well as the sequences of apical loops and bulges. These were used for a regular expression search. The regular expression matched a structure with only one bulge larger than 2 nucleotides that presented

a CU(X)UU(X)C(X) motif, and X*UGUX* in the apical loop, where X is any nucleotide and * is any number of repeats. The size of the apical loop is variable between species so it was not restricted. From the hits the most stable structures were selected. This method correctly identified ε in HBV, WHV, DHBV, and HHBV and was thus applied to all the other viral species (Table 5.3).

TABLE 5.3: Positions of epsilon, pgRNA start and end, and direct repeat (DR) positions and sequences. The primer acceptor site on DR is underlined.

Species	ε	**pgRNA**	**DR sequence**	**DR position**
HBV (human)	1857–1908	1818–1920	TTCXCCTCTGC	DR1: 1824–1835; DR2: 1590–1600
WHV (woodchuck)	1967–2025	1935–2037	TTCACCTGTGC; GTCACCTGTGC	DR1: 1940–1950; DR2: 1718–1728
GSHBV (Ground-squirrel)	36–97	5–107	TTCACCTGTGC	DR1: 11–21; DR2: 3200–3110
HBV (Chimpanzee)	1847–1907	1818–1917	TTCACCTGTGC	DR1:1824–1834; DR2: 1590–1600
HBV (gibbon)	34–94	5–104	TTCACCTGTGC	DR1: 11–21; DR2: 2959–2969
HBV (gorilla)	34–94	5–104	TTCACCTGTGC	DR1: 11–21; DR2: 2959–2969
HBV (organgutan)	1847–1907	1818–1917	TTCACCTGTGC	DR1: 1824–1834; DR2: 1590–1600

TABLE 5.3: (continued)

WMHBV (woolly monkey)	1847–1909	1820–1919	TTCACCTGTGC	DR1: 1826–1836; DR2: 1598–1608
HBHBV (horseshoe bat)	1687–1747	1661–1757	TTCACCTGTGC	DR1: 1667–1677; DR2: 1433–1443
RBHBV (roundleaf bat)	1687–1747	1661–1757	TTCACCTGTGC	DR1: 1667–1677; DR2: 1436–1446
TBHBV (tentmaking bat)	1662–1719	1634–1742	TTCACCTGTGC	DR1: 1640–1650; DR2: 1427–1437
BtHBV (“bat”)	1854–1914	1828–1924	TTCACCTGTGC	DR1: 1834–1844; DR2: 1600–1610
DHBV (duck)	2564–2623	2535–2779	TACACCCCTCT	DR1: 2541–2551; DR2: 2483–2493
HHBV (heron)	2562–2626	2535–2779	TACACCCCTCT	DR1: 2539–2549; DR2: 2482–2492
CrHBV (crane)	2557–2613	2526–2770	TACACCCCTCT	DR1: 2531–2541; DR2: 2474–2484
StHBV (stork)	2572–2628	2541–2785	TACACCCCTCT	DR1: 2546–2556; DR2: 2489–2599

TABLE 5.3: (continued)

ShGHBV (sheldgoose)	2590–2646	2559–2803	TACACCCCTCT	DR1: 2565–2575; DR2: 2507–2517
SGHBV (snowgoose	2563–2619	2532–2776	TACACCCCTCT	DR1: 2538–2548; DR2: 2480–2490
RHBV (Ross's goose)	2557–2613	2526–2770	TACACCCCTCT	DR1: 2531–2541; DR2: 2474–2484
PHBV (parrot)	2581–2637	2550–2794	TACACCCCTCT	DR1: 2555–2565; DR2: 2498–2508

5.6.2 Identification of ϕ and ω in *Avihepadnaviruses*

Locating ϕ relies on a number of initial assumptions: (1) it exists, (2) it is located near the 3' pgRNA end, (3) it interacts with 5' ε, and (4) it interacts with a region in short proximity downstream of 3' DR1, i.e. ω. In order to assess any interactions between the pgRNA ends a fragment was generated that brought the ends artificially in close proximity so they could be folded in Mfold. It consisted of pgRNA start until the UGU motif in the ε apical loop, followed by a 20 nucleotide long poly-uracil linker, and finally up to 75 nucleotides upstream of the primer acceptor site and all nucleotides downstream until the large bulge of the 3' ε. In HBV this corresponded to nucleotides 1818–1880 plus 1749–1865 in the laboratory strain, whereas in DHBV strain JX469898 it was nucleotides 2535–2596 plus 2465–2582. The sequences were folded in Mfold (Zuker, 2003) with constraints on leaving the artificial linker, the primer acceptor site, and the ε bulge single-stranded.

This method was first tested on HBV and reproduced the published interactions (Figure 5.18). Interestingly, the region between ϕ and ω formed a stem-loop

with the primer acceptor site (AS) presented in the 3' bulge. Furthermore, there appeared to be the possibility of base-pairing between the lower stems of 5' and 3' ε. Note that this interaction, however, is not essential for DNA synthesis. Deleting 3' ε does not markedly affect DNA levels (Quarleri, 2014).

Applying this approach to DHBV, similar interactions were identified, where the upper stem and part of the apical loop of ε base-pair with the region upstream of DR2 (Figure 5.19). The region downstream of DR2, including 2 nucleotides of DR2, base-paired with the region just downstream of DR1 similar to HBV ω. This would mean that the two base-pairing regions of ϕ in DHBV are separated by a single-stranded portion, including DR2, instead of being almost contiguous as in HBV. This arrangement is in line with the finding that DR2 plays a role in minus-strand synthesis in DHBV but not in HBV (Maguire and Loeb, 2010). A short part of DR2 is involved in base-pairing with putative ω. Removing these two nucleotides changes the overall structure and makes the interaction unfavourable. Interestingly, also here an interaction between the lower ε stems can be seen. This would place ϕ DHBV at positions nucleotides 2471–2475 and nucleotides 2492–2499. Repeating this experiments with the other *Hepadnaviridae* yielded similar results: all avian viruses showed $\varepsilon - \phi - \omega$ interactions as DHBV and all mammalian viruses including WHV behaved similarly as HBV except for slight shifts at the edges of the base-pairing regions. The corresponding positions are shown in Appendix A in Table A.8. The only variations were in the remaining structure (see Figure 5.20 for HHBV). In *Avihepadnaviruses* both the positions and sequences are highly conserved. The second ϕ part is fully conserved in the species tested. The first ϕ part is ACGGC in HHBV, DHBV and parrot virus, and ACAGC in goose and crane viruses. Note that the variable position (G/A) interacts with a U, meaning base-pairing is maintained.

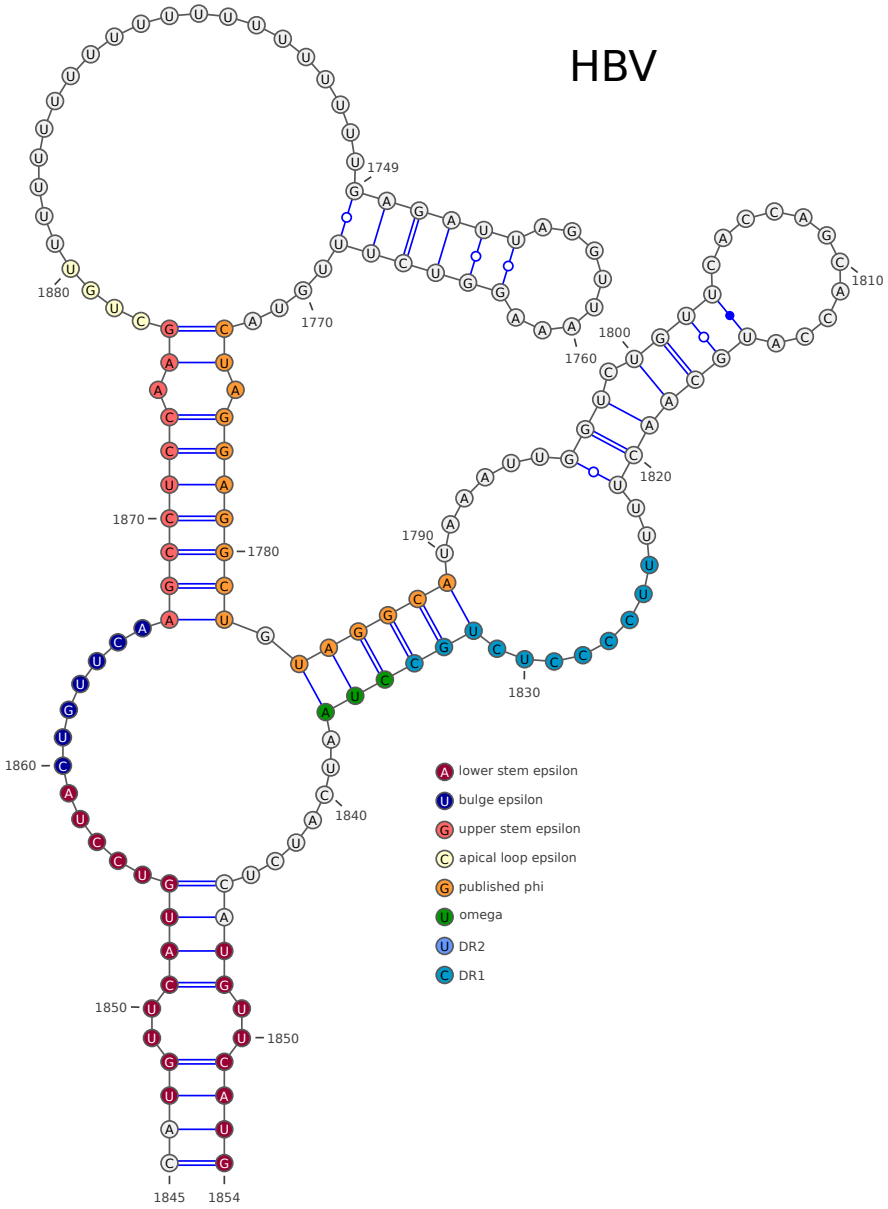

FIGURE 5.18: **5'–3' pgRNA interactions of HBV.** This structure shows the 5' ε up until the apical loop linked with 20 Us with about 100 nucleotides around the 3' DR1. The fragment was truncated on both ends to exclude single stranded ends. The upper stem of ε is shown in light red, the lower stem in dark red, the bulge in dark blue, and the apical loop in light yellow. DR1 is marked in light blue whilst DR2 is not visible in this section. ϕ and ω as published are shown in orange and green, respectively. One part of ϕ interacts with the upper stem of ε whilst the other base-pairs with ω, which partially overlaps with DR1. The primer acceptor site on DR1 is at the 5' end of DR1 and would be in the bulge of the formed stem-loop. The sequence was folded in Mfold (Zuker, 2003) whilst forcing the poly-U linker, ε bulge, and the primer acceptor site to be single-stranded.

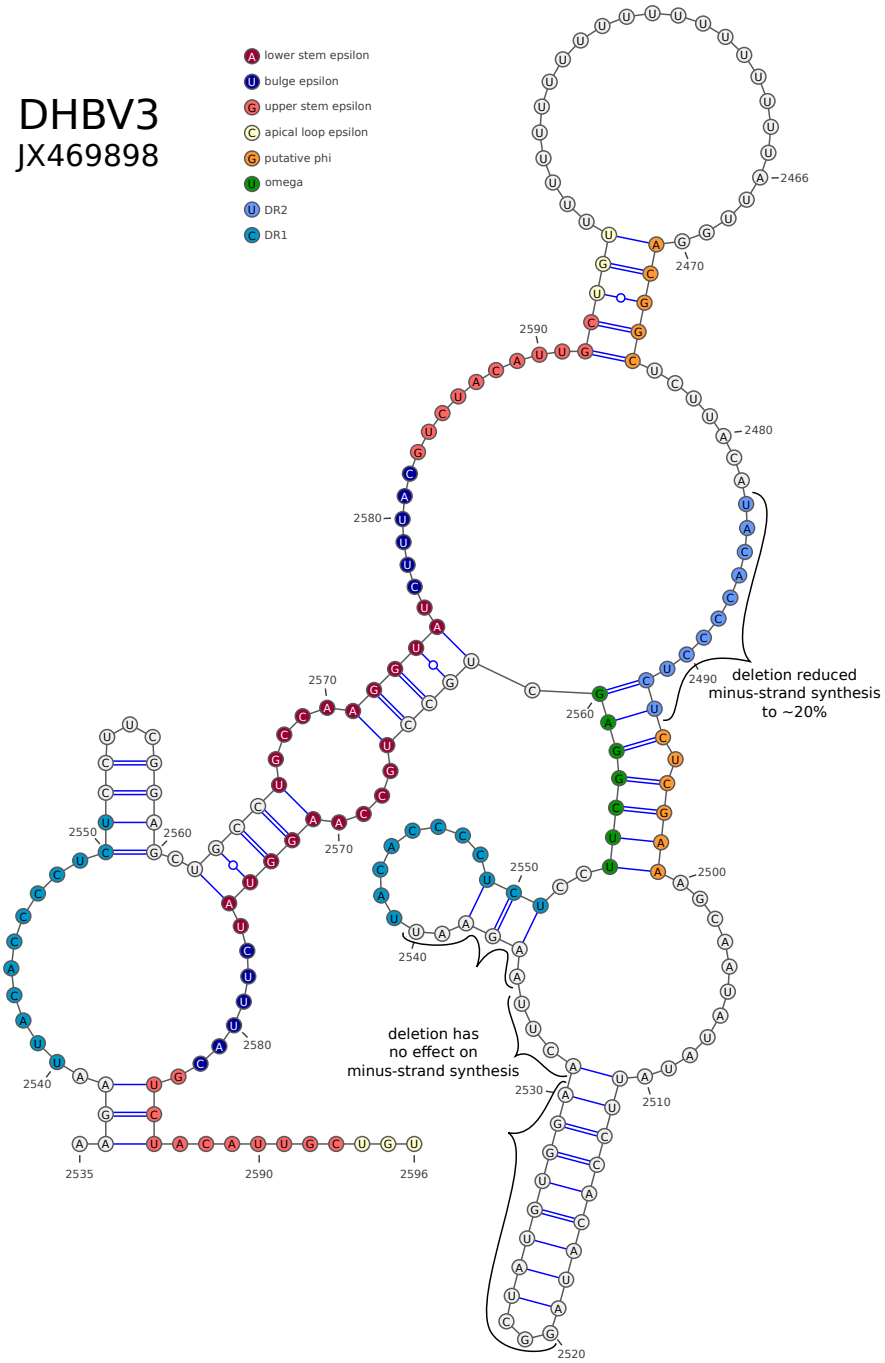

FIGURE 5.19: **Predicted 5'–3' pgRNA interactions of DHBV.** This structure shows the 5' ε up until the apical loop linked with 20 Us with about 100 nucleotides around the 3' DR1. The fragment was truncated on both ends to exclude single stranded ends. The upper stem of ε is shown in light red, the lower stem in dark red, the bulge in dark blue, and the apical loop in light yellow. DR1 and DR2 is marked in light shades of blue. Predicted ϕ and ω are shown in orange and green, respectively. ϕ is split into two parts separated by DR2, with which it partially overlaps. ω and DR1 do not overlap. The primer acceptor site on DR1 would be in the apical loop of the small stem-loop. Deleted regions from previous experiments are marked. The sequence was folded in Mfold (Zuker, 2003) whilst forcing the poly-U linker, ε bulge, and the primer acceptor site to be single-stranded.

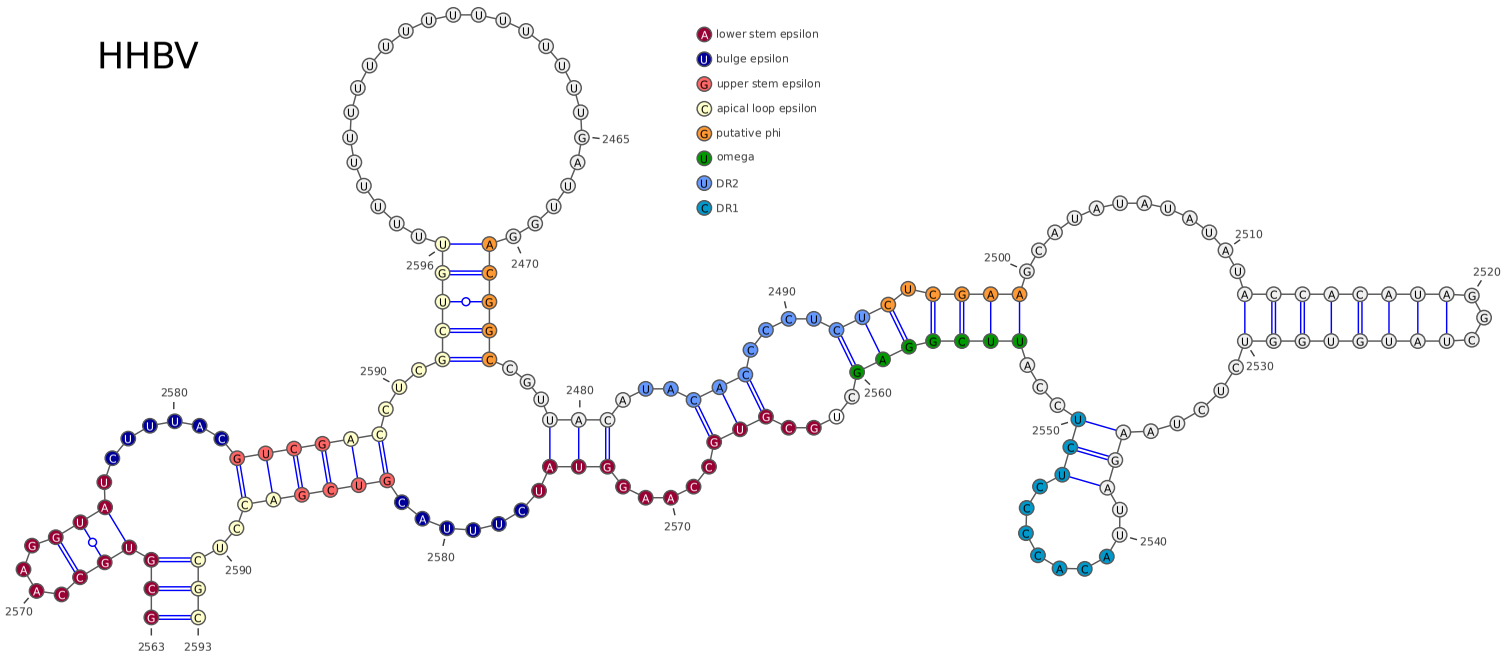

FIGURE 5.20: **Predicted 5'–3' pgRNA interactions of HHBV.** This structure shows the 5' ε up until the apical loop linked with 20 Us with about 100 nucleotides around the 3' DR1. The fragment was truncated on both ends to exclude single stranded ends. The upper stem of ε is shown in light red, the lower stem in dark red, the bulge in dark blue, and the apical loop in light yellow. DR1 and DR2 is marked in light shades of blue. Predicted ϕ and ω are shown in orange and green, respectively. ϕ is split into two parts separated by DR2, with which it partially overlaps. ω and DR1 do not overlap. The primer acceptor site on DR1 would be in the apical loop of the small stem-loop. As opposed to DHBV, HHBV pgRNA may form further interactions between DR2 and 3' ε. The sequence was folded in Mfold (Zuker, 2003) whilst forcing the poly-U linker, ε bulge, and the primer acceptor site to be single-stranded.

5.6.3 Suggestions for Experimental Validation of DHBV ϕ

Whilst experimentally testing the identified ϕ in DHBV was out of the scope of this project, I make a suggestion for an approach based on previous work on ϕ in HBV that would enable experimental testing of this hypothesis.

The position of ϕ in HBV has originally been identified through serial deletions (Tang and McLachlan, 2002; Shin et al., 2004) and the interactions characterised through a series of mutations (Abraham and Loeb, 2006; Oropeza and McLachlan, 2007; Abraham and Loeb, 2007). Since we already have a candidate region for ϕ in DHBV, doing serial deletions first would be unnecessary. Instead of completely deleting the proposed ϕ regions, they could be mutated to prevent base-pairing. Then the levels of minus-strand DNA in mutants versus wildtype would be measured. To avoid the need for synonymous mutations, this would require a transfection of cells with two plasmids: one for pgRNA and one for HBcAg and Pol protein expression. An experimental system similar to the one used in Maguire and Loeb (2010) would be useful. In HBV studies mutating away just 2–3 base-pairs (out of 15) could reduce minus-strand synthesis markedly (Abraham and Loeb, 2007). Even changing a G-C and a U-A base-pair to two G-Us in the ε-ϕ interaction reduced levels to 69% of wildtype, indicating sensitivity to small changes in stability. Below, I show the wt interactions of both parts of ϕ separately and suggested mutations (marked in red). Mutating different sites separately is common practice (Abraham and Loeb, 2006; Oropeza and McLachlan, 2007; Abraham and Loeb, 2007) and aids in pinpointing more important interactions. If a marked decrease in minus-strand DNA synthesis is detected, compensatory mutations in the interaction partner (ε or ω) can be introduced, which would restore the proposed base-pairing. An increase in minus-strand synthesis in the double mutants compared to the single mutants would show the importance of base-pairing between these regions for DNA synthesis. Note that

compensatory mutations in ε may affect packaging of pgRNA. There appears to be little consensus on which parts of ε can be mutated without affecting packaging so trying different mutants is advisable, although the UGU motif in the apical loop should be avoided. The part of ε that interacts with ϕ is largely the apical loop portion. Since the sequence is important for other functions of ε such as packaging (Knaus and Nassal, 1993), counter-mutations in ε to rescue the phenotype were not recommended. Specific suggestions for mutations are shown in Figure A.2 in Appendix A.

5.6.4 Prediction of PSs in Woodchuck and Duck Hepatitis B Viruses

If the ϕ region has the double function of also being a PS, knowing the location of ϕ in a given *Hepadnaviridae* virus allows the prediction of PSs for that virus. I used this idea to predict nucleation complex PSs in WHV and DHBV. Two assumptions were central to this task: (1) an overlap between ϕ and PSϕ and (2) another PS in close proximity upstream (PS1 equivalent).

Taking advantage of these assumptions variable fragments of 100 nucleotide length around the respective ϕ were extracted from a representative complete genome. For WHV, strain KF874493.1 and for DHBV, strain K01834.1 were utilised. The fragments were folded with Mfold (Zuker, 2003) allowing 500% suboptimality. This was in line with the SLs identified in HBV, which were also not the MFE structure on the respective fragment. The resulting structures were analysed for overlap with ϕ and apical loop similarity.

The first virus the idea was tested in was WHV. Due to HBV and WHV being in the same genus of *Orthohepadnaviruses*, a degree of similarity in the PS motifs could be expected. This is in line with observations in bacteriophages MS2 and BZ13 of genus *Levivirus*, where the top tier PSs have X(X)YA and X(X)RA in the apical loops, respectively, with these being swapped in the next tier. Additionally, this level of relatedness between HBV and WHV also meant

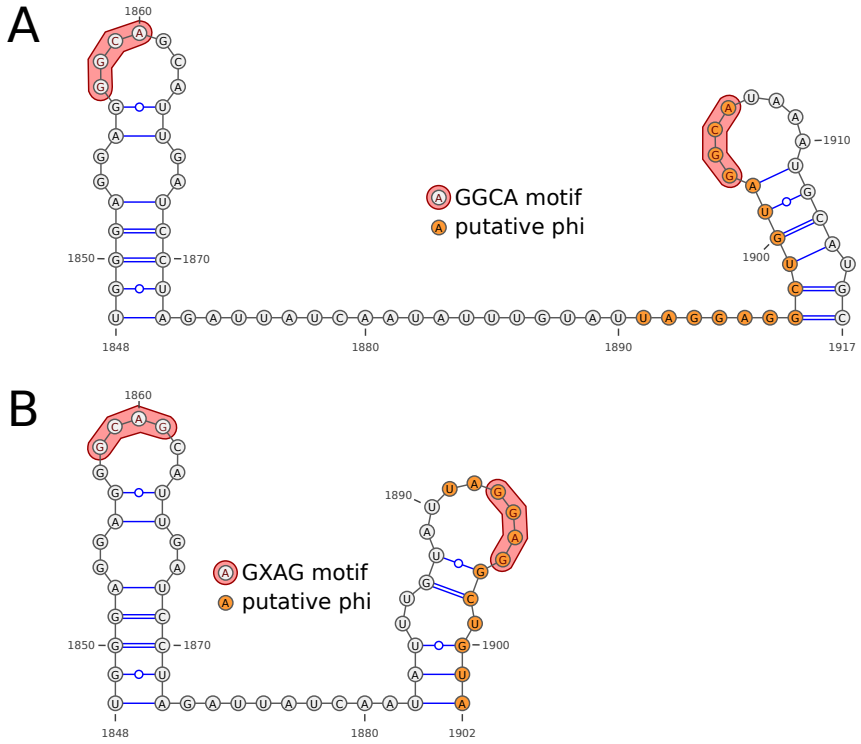

FIGURE 5.21: **Predicted nucleation complex PSs in WHV.** Two potential combinations of SLs are shown in A and B. The respective PS motifs are marked in red and the predicted ϕ is marked in yellow. The structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

that ϕ, whilst it has not been experimentally tested and published for WHV, is likely to occur in the same relative genomic location. These regions show high homology within the *Orthohepadnaviruses* and form nearly the same secondary structures (data not shown). This removes one level of uncertainty from the PSs prediction.

Two combinations of SLs were identified for WHV (Figure 5.21). Both included the same upstream SL, which would be the equivalent of PS1. This SL is very stable comparable to PS1 in HBV and is thus dominant in this region. For the second SL, however, there are two options that fulfilled the requirement of overlapping with the predicted ϕ and being partially similar in the apical loop to the first SL. The first combination would conclude in a GGCA PS motif (Figure 5.21A). Note that the actual motif would probably be less specific than that and allow for some variation. This is similar to HBV, where both PS1 and PSϕ

display the version GGAG of RGAG. The second combination, on the other hand, would result in a more variable motif: GXAG (Figure 5.21B). Interestingly, the putative PSϕ in this case would be almost the same SL as in HBV. This reflects how closely related these viral species are and may point towards cross-reactivity of their PSs and CPs.

Next, the method was also applied to DHBV. As explained earlier, ϕ was not initially known for this virus but was predicted above. The high degree of similarity in sequence and secondary structure between different members of the *Avihepadnaviruses* in that region is a good indication that the equivalent of ϕ is indeed located there and conserved within this genus (see Figures 5.19 and 5.20). Nevertheless, it has to date not been experimentally verified.

As opposed to WHV only one set of SLs was identified in DHBV as putative nucleation complex PSs (Figure 5.22). Interestingly, there were three SLs in close proximity rather than two as in HBV and WHV. They share the putative PS motif RCAA. Due to the split nature of the predicted ϕ in *Avihepadnaviruses* two of the SLs overlapped with it. However, the middle SL had only a small overlap in the outermost two base-pairs, which may melt. It is otherwise the most stable of the structures in this group. The third SL fully overlaps with the second part of the predicted ϕ. Moreover, it is a less stable structure, making it easier to melt and make ϕ available for binding to ω. Therefore, it is the more likely equivalent of PSϕ, whilst the middle SL is more like PS1. The additional SL may indicate a small variation in NC functionality in DHBV.

5.6.5 Compact PS profiles in DHBV with Predicted Motif

Having predicted a PS motif in DHBV it could be used for the phylogeny method developed in Chapter 2 and applied to HBV in Chapter 4. All available complete DHBV genomic sequences were downloaded from the NCBI database and processed as described in detail in Chapter 4 Section 4.3.1 and Chapter 2 Sec-

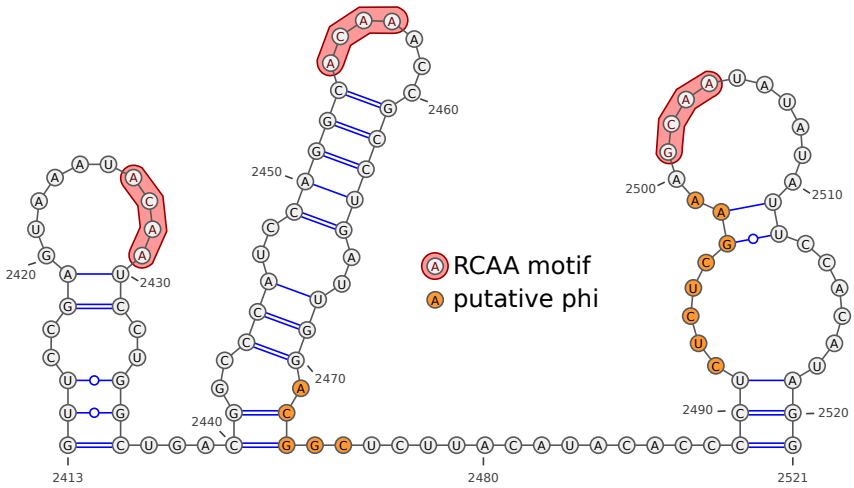

FIGURE 5.22: **Predicted nucleation complex PSs in DHBV.** This structure shows three putative PSs for DHBV. The proposed PS motif RCAA is marked in red in all SLs and putative ϕ is marked in yellow. One SL overlaps completely with the latter portion of ϕ and thus assumed to be the PSϕ equivalent, whilst the more stable SL upstream of it only overlaps slightly with the first part of ϕ at the very bottom of the helix, which may easily melt. The structures were visualised in VARNA (Darty et al., 2009) and edited in Inkscape.

tions 2.4.1.1–2.4.1.8. Briefly, the sequences were converted into pgRNA form using the start sequence of the TATA box, which is TATATA in DHBV, and the poly-A signal ATAAAGAA. In that form the sequences were fragmented in a one-nucleotide sliding window of 30 nucleotide length. These fragments were folded by sampling 10,000 times from the partition function, SLs extracted and merged across windows, i.e. only unique SLs kept in the list. This list of SLs was used for the SL selection program, which selected a set of non-overlapping SLs that, when added together, had the lowest overall energy. These energies also took into account PS affinities. SLs that displayed the RCAA motif in their apical loops had an energy bonus added corresponding to K_D of 15 nM. ε was used as anchor and always included. These SL sets were then utilised to generate PS profiles, i.e. pseudosequences where each genomic position is encoded as either part of a PS or not. So any genomic position that is involved in a PS would be encoded as “C” and all others as “A”. Compact versions of these profiles were created as described in Chapter 4 Section 4.6. Essentially, the sequences were compressed

FIGURE 5.23: **Compact Packaging Signal Profiles of DHBV strains.** Manually aligned profiles are shown for DHBV strains visualised in SeaView (Gouy et al., 2010). PSs are shown as "C"s (green), ε as "T"s (blue) and remaining sequence as "A"s (red).

by the factor of the expected SL length: Every PS was shortened down to a single letter "C" and the positions between PSs were shortened by dividing their length by the expected SL length, here 25, rounding to the nearest integer, and inserting that many "A"s. This method is crude and does not yield well aligned profiles so manual alignment of PS positions in SeaView was necessary (Gouy et al., 2010). The resulting compact PS profiles are shown in Figure 5.23. Similar to HBV there is a set of conserved PSs towards the 3' end of the pgRNA but most in the middle. In general the DHBV profiles appear less variable than the HBV ones indicating a closer relatedness and less diversity between the duck sequences. DHBV strains also had conserved PSs in the 3' half of the genome. These data show that given the RCAA PS motif, DHBV displayed some similarities in the pattern of PS distribution such as a set at the 3' end including PS1 and PSϕ equivalents and several groups in the middle, whilst also being distinct.

5.7 Discussion

Cryo-EM with asymmetric reconstruction performed by collaborators has shown a single spot of density within a capsid re-assembled in the presence of multiple copies of PS1 SLs. This density is located along the 2-fold and from there towards the 3-fold axes of symmetry and can fit 2–4 PS1 SLs. The fact that only this single point of density was observed led to the conclusion that HBV assembles via a nucleation complex of just 2–4 PSs. The location of these points towards a mechanism for how the virus assembles into T=4 capsids *in vivo*. From the

literature it is apparent that HBcAg can form into capsids of either T=3 or T=4 icosahedral symmetry. Both T=3 and T=4 capsids have been shown to be able to package RNA when expressed in *E. coli*. For both an inner density is visible in the micrographs (Crowther et al., 1994; Roseman et al., 2005). However, the ratio of these isoforms is largely affected by the expression system. The carboxy-terminal tail, which extends into the interior of the capsid and binds RNA, appears dispensable for capsid assembly, but may play a role for determining symmetry *in vivo*. Whilst *in vitro* also truncated capsid proteins form mostly T=4 capsids (Zlotnick et al., 1996), their relative percentage is highly reduced when capsids are isolated from *E. coli* and analysed directly (Kenney et al., 1995). In other viruses that display capsid polymorphism *in vitro* the dominant capsid isomorph found *in vivo* is the one that packages the viral RNA (Baker et al., 2002). Hence, whilst the virus shows a preference for T=4 capsids *in vivo*, it is not clear how this is regulated. The position of the PS density in the cryo-EM structure relative to the outer protein shell provides a possible mechanism for T=4 preference. In a T=4 capsid this density is situated at the inter-dimer interfaces. The PSs can thus stabilise these dimer-dimer interactions and by bringing together the first few dimers can set the geometry. Whilst it is not strictly possible to say where the PSs would be located within a T=3 capsid without actually imaging one, if the same positioning relative to the symmetry axes is applied to a T=3 capsid, the PSs would be located at the inter-dimer interfaces. Since HBcAg forms stable dimers without interacting with RNA, such an interaction would not provide additional stabilising benefits in the assembly of the capsid. This hypothesis provides one possible answer to the question why one isoform is preferred over the other *in vivo*.

Based on the experimental observations, I searched for a group of 2–4 putative PSs in close proximity to each other along the pgRNAs. Initially, seven such regions were identified when considering PS profiles generated for Chapter 4. However, only one, later termed LS1, folded into more than one RGAG-containing

SL when folded with Mfold. The idea was to find a relatively short genomic region, around 100 nucleotides in length that could be used for re-assembly experiments by collaborators to test the nucleation complex hypothesis. It was therefore essential to ensure that the respective fragment would actually fold into the putative PSs. The difference between the two structure prediction methods utilised here is twofold. For one, the PS profiles gave a stability bonus to RGAG SLs, which boosted them over competing structures, whilst Mfold only considers the pure SL folding energies. Since the fragments were to be used in re-assembly experiments, the RNA would be folding and re-folding in the presence of HBcAg. It was therefore not necessary for the PSs to be in the MFE structure for that fragment. However, even when allowing 500% suboptimality only LS1 folded into two putative PSs. The other difference is the window size. Which structures can form in a specific region largely depends on neighbouring structures. The size of the window that is folded can therefore determine whether certain structures appear (together) or not. For the PS profiles, 30 nucleotide sliding windows were utilised to allow most folds of reasonable size to be considered in the following selection step. For the fragment test, on the other hand, one large region around 100 nucleotides was folded with little flexibility at the 5' and 3' ends. Whilst two or more RGAG SLs may be able to form in isolation on a short neighbouring fragments, their combination may not be favourable as other, more stable SLs take their places. This does not mean that these are necessarily not real PSs *in vivo*. In the cell, the entire length of pgRNA is available and thus more 5' and 3' neighbouring sequence. Moreover, the RNA is likely to fold sequentially 5' to 3' as the last ribosome finishes after translation has been switched off. Whilst it might have been possible to stabilise the putative PSs on the other identified fragments using mutations to ensure their presence for experiments, the initial goal was to use only genomic sequence. It was therefore important to only use fragments that were predicted to fold into RGAG SLs without further manipulation. Further evidence for the importance of the LS1 region could be found

in its conservation in different HBV genotypes. Even ancient strains showed a striking conservation of the parts of the sequence involved in the putative PSs. Consequently, only LS1 was forwarded to collaborators for testing.

In addition to identifying nucleation complex PSs, another follow-up from the initial identification of PSs in HBV was to study the effect of their knock-out. To this end, I suggested a set of synonymous mutations in the LS1 region, which we considered the most important one due to its likely role in nucleation, to eliminate all RGAG SLs whilst preserving coding. This task proved difficult despite the PSs being located in an area that only codes for one gene (X). The codons used allow for little change on the sequence, so that a simple mutation of the PS motif was not possible. Instead, I used a genetic algorithm to evolve the sequence to reduce the relative stability of RGAG SLs. This was measured by folding the fragment in a 40, 50, and 60 nucleotides sliding window, sampling 100 folds from the partition function, and counting the number of times an RGAG SL is present. The sequences with the lowest numbers were kept each round and new sequences generated by introducing random, synonymous mutations in them. This was continued until a plateau was reached, when the number of RGAGs did not decrease further. Another restriction was that one of the two PSs in LS1 largely overlaps with the *cis*-acting element ϕ, which is essential for reverse transcription and cannot be mutated without affecting function. This made it impossible to completely knock-out all RGAG SLs, but rather a knock-down was achieved. Whilst this still results in the possibility for a putative PS to fold, the probability is much reduced. Nevertheless, is may be that the knock-down is not sufficient to have the desired effect and it is possible that no change in packaging would be observed in the mutant. To ensure a full knock-out non-synonymous mutations would have to be allowed. The only way for this to be possible, whilst maintaining a replicating viral system, is to utilise a helper plasmid. This method is commonly employed when testing packaging functions of different parts of the pgRNA in HBV. Two plasmids are utilised: one provides pgRNA to package and

the other expresses proteins needed for packaging, i.e. Pol and HBcAg, but lacks ε (Liu et al., 2004). One may even argue that the helper is not strictly necessary as HBxAg, whose gene is in LS1, is not known to be involved in packaging. Also, preservation of ϕ is only needed if several cycles of infection need to be studied. If the constructs are only to be tested in a single round of packaging, then the fact that they are incapable of reverse transcribing is irrelevant. Otherwise, it is possible to partially rescue reverse transcription by complementary mutations in ε and ω (Oropeza and McLachlan, 2007; Abraham and Loeb, 2007).

The discovery of LS1 led to another interesting observation, namely that the secondary PS, the one downstream of PS1, largely overlapped with ϕ. This inspired the ϕ hypothesis: The idea that the ϕ region performs a double function as PS first, and in promoting reverse transcription later. After translation has been shut off by Pol binding to ε, SLs start folding 5' to 3' along the pgRNA as the last ribosome finishes protein synthesis. Eventually, the nucleation complex PSs are formed close to the 3' end and are bound and stabilised by HBcAg. The location of these PSs may ensure that capsid assembly does not initiate until translation of that RNA is completed to avoid direct competitions between the functions or trapping in intermediates, as not the entire RNA is available for packaging. The PS-HBcAg complexes may then interact with an ε-Pol-HBcAg complex at the 5' end, bringing the two ends into close proximity. This provides an explanation for why only two PSs were identified in LS1, whilst 2–4 could be fitted into the density of the cryo-EM structure: the ε complex may provide the additional contact to help set the geometry. Interaction of all these players triggers assembly, which may also involve other, less important PSs. After completion of assembly and packaging, the PSϕ starts to melt and exposes parts of ϕ, which can begin interacting with ω and later ε. This switch of function is facilitated by the PSϕ not being a very stable SL unlike PS1. An advantage of this proposed double function is that it provides a mechanism for the virus to regulate the timing of reverse transcription. ϕ is not available whilst translation is on-going due to the

constant flow of ribosomes along the pgRNA. When translation is switched off, the sequence is immediately sequestered into a SL and bound by surrounding HBcAg proteins. Only after packaging of the pgRNA does the SL dissociate, melt, and become available for binding to ε and ω. This ensures that reverse transcription cannot commence until after the pgRNA has been packaged into a capsid. Since the pgRNA is transcribed by host RNA polymerase II (Rall et al., 1983), it appears like a cellular mRNA to the host cell. It contains a 5' cap and a poly-A tail at the 3' end, which enables it to also act as mRNA and be translated by host proteins. Thus, it would not trigger an innate immune response in the host cell as it does not appear foreign. DNA in the cytoplasm, however, is not a usual occurrence in eukaryotic cells and would be recognised as a pathogen-associated molecular pattern (PAMP) by toll-like receptors (TLRs). TLRs are pattern recognition receptors (PRRs) in mammals that recognise certain molecular features that tend to occur in bacteria and viruses, PAMPs, such as certain lipoproteins (Kawai and Akira, 2011). Upon binding to their respective PAMP, they trigger a signalling cascade, which results in the release of pro-inflammatory cytokines. For HBV the most important one is TLR9. It resides in vesicles in the cytoplasm of certain types of cells and recognises nucleic acids, most notably unmethylated CpG on DNA (Hemmi et al., 2000; Kawai and Akira, 2011). How dangerous this is for the virus can be seen through the fact that HBV down-regulates the expression of TLR9 in dendritic cells and other liver immune cells (Vincent et al., 2011; Wu et al., 2009). However, due to the location of this PRR in vesicles and its low expression in hepatocytes, it is unlikely to play a role in detecting free cytosolic viral DNA in the host cell itself. Recently, a novel molecular sensor of cytosolic DNA has been discovered: cyclic GMP-AMP synthase (cGAS). Through this molecule a signalling cascade is triggered, resulting in the release of antiviral interferons (Sun et al., 2010). Interestingly, cGAS is expressed in hepatocytes but is not activated by HBV infection; no measurable interferon is expressed upon infection (Verrier et al., 2018). Verrier et al. (2018)

also found that it is likely packaging the genomic RNA, which prevents it from being detected. When they transfected cells with naked viral DNA, a significant response was mounted. This supports the idea that having viral DNA free in the cytosol, would trigger immune activation, and that the virus has evolved to avoid this, possibly through the double function of ϕ and associated regulation of the timing of reverse transcription.

Functional elements as important as ϕ and PSϕ are likely to be conserved not only among HBV strains but also other related viruses considering that other *cis*-acting elements such as ε are also conserved among all *Hepadnaviridae* (Kramvis and Kew, 1998). Therefore, if ϕ is known in a virus, nucleation complex PSs can hypothetically be predicted from that region based on the overlap with ϕ. This idea was applied to first WHV, which is in the same genus as HBV and shares more features with it. Whilst ϕ has only been described and published for HBV, the region has high similarity in WHV and other *Orthohepadnaviruses* and can therefore be assumed to be the same. Taking this as basis, two putative pairs of SLs were identified as candidates for PSs participating in the nucleation complex of WHV. Computationally, there is no way of determining, which of the SL combinations, if any, is the correct set of nucleation complex PSs. Both also have properties of stability in common with the PSs in HBV, namely that the PS1 equivalent is a quite stable SL, whilst the PSϕs are more unstable folds. Less stability can allow the SL to melt easily when ϕ needs to base-pair with ϵ and ω. Therefore, the only way to verify the correct set of SLs is to test them both experimentally in a laboratory. Interestingly, both imply a PS motif that is similar to RGAG in HBV. This indicates a potential for cross-reactivity of WHV, HBV, and possibly all *Orthohepadnaviruses* PSs and CPs. Attempting re-assembly of WHV viral capsids with HBV PS1/PSϕ and HBV capsids with WHV PSs, once confirmed, would validate this idea.

Nucleation complex PSs were also predicted for DHBV. This involved the more complex process of first also predicting ϕ, which was previously thought to

not be present in *Avihepdnaviruses* (Maguire and Loeb, 2010). Considering the crucial role this region plays in HBV and its hypothesised extended role, it was unlikely to not be conserved. To identify a putative ϕ in *Avihepadnaviruses*, the 5' and 3' ends of the pgRNAs were linked by poly-U and folded. As expected, two interactions, one with ε and one with a sequence shortly downstream of DR1, i.e. ω, were found, which were conserved among all *Avihepadnaviruses*. As opposed to *Orthohepadnaviruses* these two regions were not contiguous, but separated by a short stretch of sequence including most of DR2. Whilst there is no experimental confirmation that this is in fact the correct region that functions as ϕ there are strong indications for it. For one, the interactions are conserved within the genus similar to the respective interactions in *Orthohepadnaviruses*. Secondly, it is in line with previous experiments by Maguire and Loeb (2010). The regions that they deleted or mutated without having an effect on reverse transcription were not part of the proposed interacting sequences. Instead, a small part of DR2 is overlapping with it. Without DR2 this interaction is not formed. When DR2 is deleted, minus-strand DNA synthesis decreases in DHBV and HHBV but not HBV (Maguire and Loeb, 2010). Experimental validation of the proposed ϕ and ω regions in DHBV would involve mutations in these sequences that disrupt base-pairing and testing the mutants for their ability to synthesise minus-strand DNA. Until such experiments can be performed, the regions are utilised for PS prediction. As opposed to WHV, only one set of SLs was identified that fulfilled the requirements and this set contained three putative PSs rather than two. Whether any of these are in fact functional and can trigger capsid assembly will be tested experimentally in re-assembly experiments.

The predicted PS motif in DHBV was utilise to generate compact PS profiles in order to gain insight into the pattern of distribution of PSs in this virus. Compared to HBV the profiles were less diverse illustrating the close relatedness of the duck virus sequences compared to HBV genotypes. Analogous to the HBV strains, the DHBV profiles showed a small group of PSs at the 3' end and many

towards the middle, whose positions were more conserved in DHBV. On the other hand, the DHBV strains were lacking the groups of PSs at the 5' end that are highly conserved in HBV. Due to the DHBV PS motif not being experimentally validated, these results need to be taken with a grain of salt. Once more evidence is available a closer comparison can be made between these viral species including reconstructing a phylogenetic tree. Until then, however, it is possible to test the phylogenetic method on two different species from another family, the *Leviviridae*.

Chapter 6

Application of Phylogeny to *Leviviridae*

In Chapter 2 a method for reconstructing phylogenetic trees from packaging signal (PS) profiles was presented. When it was applied to hepatitis B virus (HBV) in Chapter 4, some limitations and difficulties became apparent. The method makes use of a conservation threshold. It greatly influences the number of PS blocks and thus characters available for tree building. Therefore, in setting this threshold a careful balance needs to be found between resolution and noise: Setting it too high results in too much artificial similarity between the strains and little resolution. On the other hand, setting it too low can introduce noise and the inflation of informative characters lead to excessive distance between the strains. Ideally, the number of characters should be close to the number of expected PSs. However, PSs have only recently been identified in HBV and, as a result, not much is known about how many are needed in the virus for efficient capsid assembly. It was therefore difficult to find a suitable threshold. Even for the lowest threshold only 37 characters were found, which led to the conclusion that only few PSs are needed in HBV. Reconstructing a meaningful phylogeny from PS profiles is easier using a virus with better known PS numbers and properties. Such a virus is MS2 and other members of the *Leviviridae* family, which are thought to

utilise up to 60 PSs for assembly and packaging and have well-defined PS motifs. Additionally, *Leviviridae* provide the opportunity to apply the PS phylogeny method not only to strains of one viral species but to viruses of different species and thus reconstruct phylogenetic trees within a viral family.

Leviviridae is a family of single-stranded RNA (ssRNA) viruses infecting bacteria. It includes the genera *levivirus* and *allelovirus*. Especially *levivirus* MS2 and *allelovirus* Qβ are well studied members of this family. MS2 in particular has been a model virus used extensively for studying PS-mediated assembly. Much of the methods and insights used for the analysis of HBV were originally developed for MS2. One advantage of developing methods on these phages is that they have been extensively studied by several groups and much about the structure of their genomes and the contacts between RNA and capsid are known. Moreover, they are less complex so that the PS-mediated assembly mechanism is easier to study in such systems. Therefore, MS2 and related viruses will be the focus of this and following chapters.

6.1 Leviviridae

The first virus of the *Leviviridae* family was discovered in 1961 as a ssRNA virus that infects *Escherichia coli (E. coli)* (Loeb and Zinder, 1961). Since then this family of bacteriophages has been studied extensively due to their abundance and the relative ease to work with. *Leviviridae* are currently subdivided into four groups that are distinct in a number of properties including UV sensitivity (Watanabe et al., 1967b), filtration and elution patterns (Watanabe et al., 1967a), immunochemical and serological properties (Overby et al., 1966), and replicase cross-reactivity (Haruna et al., 1967). Group I includes MS2, group II GA, group III Qβ, and group IV SP (Watanabe et al., 1967a; Sakurai et al., 1968; Miyake et al., 1971; Sundram et al., 2006). Groups I and II belong to the genus *Levivirus*, whereas groups III and IV are members of the *Allolevirus* genus (Murphy et al., 1995). MS2 was the first biological entity to have its complete genome sequenced

albeit in parts (Jou et al., 1972; Vandenberghe et al., 1975; Fiers et al., 1975, 1976).

6.1.1 Genome Replication

These viruses replicate via a double-stranded RNA intermediate. The positive-sense RNA first serves as messenger RNA (mRNA) for synthesis of viral proteins such as replicase (Erikson et al., 1964; Godson and Sinsheimer, 1967). Replicase is an RNA-dependent RNA polymerase (Haruna et al., 1963). It competes with translating ribosomes for the same substrate and is capable of inhibiting new ribosome binding while already bound ones finish and detach (Kolakofsky and Weissmann, 1971a,b). Replicase then uses the plus-strand as template for minus-strand synthesis resulting in the double-stranded intermediate (Billeter et al., 1966). The enzyme is thereby template specific and only works on viral RNA with limited cross-reactivity within the *Leviviriudae* genera (Haruna et al., 1963; Haruna and Spiegelman, 1965; Haruna et al., 1967). From there more plus-sense RNA is made, which serves as further mRNA or genomic RNA for progeny virus (Weissmann and Borst, 1963; Billeter et al., 1966). As positive-sense RNA is continuously generated, the replication intermediate is not truly observed as double-stranded but rather has many plus strand tails protruding as they are constantly displaced (Erikson et al., 1964; Fenwick et al., 1964). The mRNA codes for three proteins in *Leviviruses* (Gussin, 1966; Horiuchi and Matsuhashi, 1970): coat protein (capsid protein (CP)), maturation protein, and replicase. *Allolevirus*es make an additional protein due to a read-through in the coat protein cistron (Garwes et al., 1969; Weiner and Weber, 1971; Horiuchi et al., 1971; Moore et al., 1971).

Replicase is the first viral protein that peaks after infection. It is followed by coat protein that in turn inhibits replicase and maturation protein expression (Lodish and Zinder, 1966; Viñuela et al., 1967; Nathans et al., 1969). This switch is achieved by binding of coat protein to the initiation site of the replicase cistron

(Sugiyama and Nakada, 1967; Ward et al., 1967, 1968; Robertson et al., 1968; Sugiyama and Nakada, 1968; Lodish, 1968; Eggen and Nathans, 1969; Bernardi and Spahr, 1972). More specifically, a dimer of CP binds to a PS stem-loop (SL) called TR (translation repressor) (Carey et al., 1983a,b; Beckett and Uhlenbeck, 1988; Rolfsson et al., 2008). The structure is essential for this function as melting RNA SLs results in increased replicase synthesis (Fukami and Imahori, 1971) due to unfolding of TR. While some changes to the sequence are tolerated, others significantly decrease CP affinity. In group I phages the TR binding motif requires four nucleotides in the apical loop with a pyrimidine (U or C) at the third and an A at the fourth position. Additionally, this PS also requires a 5' bulged A and two base-pairs above and three below it (Romaniuk et al., 1987). Low levels of maturation protein synthesis are maintained through synthesis from nascent plus-strand RNA, where the binding site has not yet been made (Robertson and Lodish, 1970; Kolakofsky and Weissmann, 1971a).

6.1.2 Packaging and Assembly

The capsids of *Leviviridae* display a T=3 tiling with 180 CP making up the icosahedral capsid (Vasquez et al., 1966). They consist of 60 identical triangular units, which each contain three different coat protein conformers: A, B, and C. These form 60 AB heterodimers, and 30 CC homodimers (see Figure 6.1 a and b) (Valegård et al., 1990; Golmohammadi et al., 1993). The inside of the capsid exhibits patches of positive charge (Valegård et al., 1990) that enable binding of the negatively charged RNA. Note, this effect is not only electrostatic but also has a sequence-specific component as explained above.

While the CPs of MS2 and other *Leviviridae* phages can self-assemble *in vitro* (Matthews and Cole, 1972), *in vivo* capsid assembly and genome packaging is aided by the genomic RNA (Hohn, 1969). The process is triggered by a CP dimer binding to TR (Hung et al., 1969; Ling et al., 1970; Beckett and Uhlenbeck, 1988; Beckett et al., 1988). Interaction with TR triggers the conformational

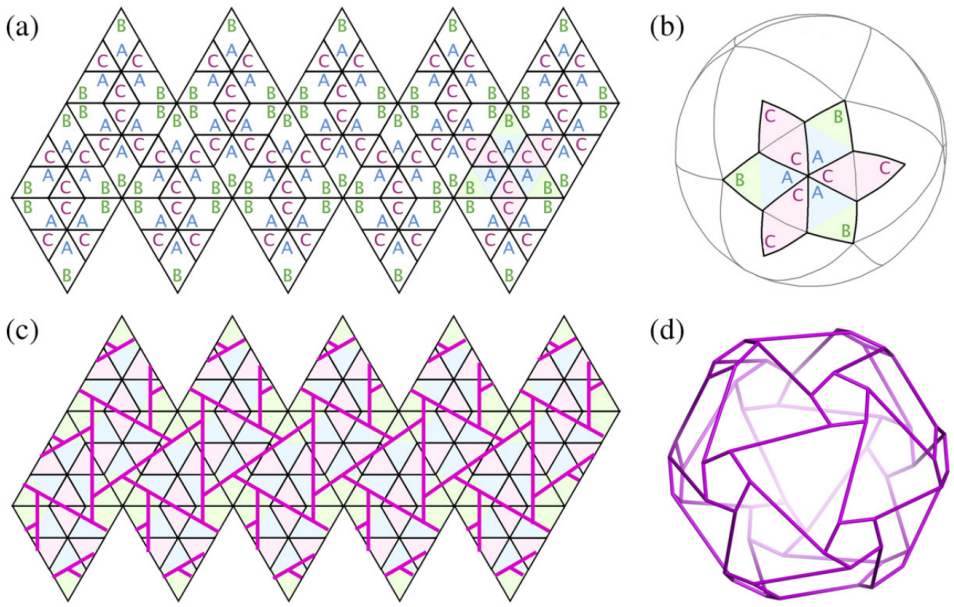

FIGURE 6.1: **MS2 capsid protein and RNA arrangements.** The tiling of MS2 coat protein conformers A, B, and C in 2D (a), and visualised on the capsid surface in 3D (b). AB dimers are arranged as pentamers around the 5-fold axis, whereas CC dimers are located on the 2-fold axes. If icosahedral averaging is applied to the imaging, the observed inner RNA density is also icosahedrally ordered. The location of the averaged density is visualised on top of the 2D tiling (c) and in 3D (d). In reality the RNA is thought to follow a Hamiltonian path along some lines of this lattice. Figure 6 as published in Toropova et al. (2008). Copy right cleared with Elsevier through Copyright Clearance Center, license number 4478220904481.

change in the coat proteins from CC homodimers to AB heterodimers required for assembly (Stockley et al., 2007). Normal-mode analysis of CP dimers in contact with TR confirmed this effect and revealed that there is little sequence specificity involved in it, which supports the idea for more PSs in MS2 in addition to TR (Dykeman and Twarock, 2010; Dykeman et al., 2010). In a complete capsid there are 60 AB dimers thus indicating that there would be 60 such PSs in the MS2 genome. Some have been identified or predicted. They are dispersed PSs in the phage genomes and have slightly different motifs and affinities (Basnak et al., 2010; Stockley et al., 2007; Knapman et al., 2010; Dykeman et al., 2013b; Rolfsson et al., 2016; Twarock et al., 2018). Cryo-electron microscopy (cryo-EM) of MS2 capsids re-assembled in the presence of several copies of TR shows an inner density beneath the protein layer corresponding to the bound RNA (Fig-

ure 6.1 c and d) (Toropova et al., 2008). Icosahedral averaging blurs the actual position of the RNA and results in the image of an icosahedrally ordered inner RNA layer. The RNA contacts are expected to follow a Hamiltonian path along some of the lines of this icosahedron. Further insights into the organisation of genomic RNA inside the capsid have recently been gained through cryo-EM with asymmetric reconstruction. As opposed to previous structures, these were not icosahedrally averaged and could therefore resolve asymmetric structural features. With TR being located around the middle of the genomic RNA it was hypothesised and modelled before that the capsid assembles in two hemispheres upstream and downstream of TR (Dykeman et al., 2011). In these cryo-EM structures (Koning et al., 2003; Dai et al., 2017) this was confirmed, and additionally 15 PS SLs were found and their relative location within the capsid resolved (Dai et al., 2017), again consistent with previous predictions (Dykeman et al., 2013b). These SLs will be referred to in this thesis as "Hong PSs" after the lab in which the structure was resolved.

6.2 Phylogenetic Trees of *Levivirus* PSs

6.2.1 Processing of Sequences

All complete genomic sequences of BZ13 and MS2 phages were downloaded from NCBI. There were five BZ13 and nine MS2 strains. Their names and accession numbers are shown in Table 6.1. For phylogeny the sequences were first processed separately as detailed in Sections 2.4.1.1 and 2.4.1.2. Briefly, using a sliding window approach the sequences were broken up into overlapping 30 nucleotide fragments shifted by one nucleotide each. Each of these was folded in Tfold in partition function mode sampling 10,000 structures. Single SLs were extracted and their occurrence among the sampled structures as well as positional information saved. Finally, the SLs were merged across overlapping windows adding the occurrences of each respective SL. In addition to normal Vienna structure

format, all SLs were also encoded into an own structure/sequence format. This enabled simple search for motifs consisting of specific sequence as well as structure elements.

TABLE 6.1: Accession numbers for the *Levivirus* genomes used for phylogeny.

Species	**Strain**	**Accession number**
BZ13	GA	NC_001426
BZ13	DL20	FJ483839
BZ13	KU1	AF227250
BZ13	DL10	FJ483837
BZ13	T72	FJ483838
MS2	DL16	EF108464
MS2	J20	EF204939
MS2	R17	EF108465
MS2	ST4	EF204940
MS2	MS2	NC_001417
MS2	DL52	JQ966307
MS2	DL1	EF107159
MS2	M12	AF195778
MS2	fr	X15031

6.2.2 Creating PS Profiles

Once lists of SLs have been created for each genomic sequence, PS profiles can be generated. The method is described in detail in Chapter 2 Sections 2.4.1.4–2.4.1.6. Briefly, the PS motifs for each viral species together with affinities were supplied to the program. As opposed to HBV in Chapter 4 it was possible to use different affinity tiers. For MS2 the same motifs were used as in Chapter 2 Section 2.4.2 (Figure 2.12). BZ13 is similar to MS2 in its PS motifs and the difference between them is mostly considered to be a switch from Y (pyrimidine) to R (purine) bases before the A in the apical loop (Figure 6.2). The affinities were based on Dykeman et al. (2013b) and were K_{D}s of 1.5 nM for top, 150 nM for medium, and 1500 nM for low affinity. Given these a two step SL selection process took place. The weighted activity selection (WAS) algorithm first selected

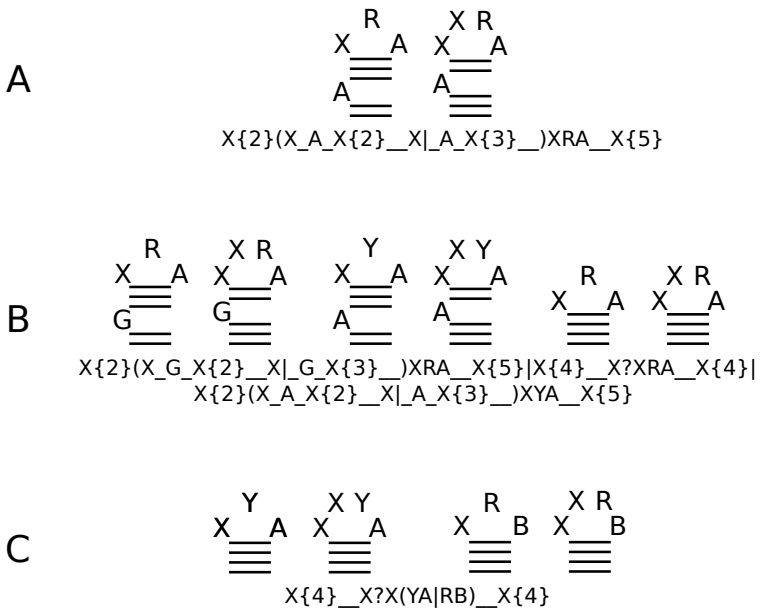

FIGURE 6.2: **BZ13 packaging signal search motifs.** Analogous to MS2 (Figure 2.12) the three motifs for the tiers of affinity used for the SL selection and phylogeny are shown as structure and corresponding search phrases, shown for high (A), medium (B), and low (C) affinity tier PSs.

non-overlapping SLs from the list that added together maximised overall stability. Stability in the first round was simply the ΔG of the SL with inverse sign such that a SL with low ΔG would have a high stability value. If a PS from the highest affinity tier was selected in this round, its stability value was changed to 10,000 to ensure it would be selected in the next step. Afterwards, the K_{D} values converted to ΔGs were added to the stabilities of each SL whereby non-PSs got an added value of -3.5 kcal/mol for unspecific binding. At the same time structures with a too low energy barrier for unfolding were excluded by assigning them a stability of -100. Then, another round of selection using the WAS algorithm was performed with the updated stability values. This final structure was then converted into PS profiles, which are pseudosequences encoding each nucleotide position as part of a high ("C"), medium ("G"), or low ("U") affinity PS or not part of a PS SL ("A").

6.2.3 PS Blocks and Phylogenetic Trees across Species

The method for PS blocks and reconstruction of phylogenetic trees from them is explained in detail in Chapter 2 Sections 2.4.1.7–2.4.1.9. Briefly, creation of PS blocks requires PS profiles for each sequence, a multiple sequence alignment (MSA) for all sequences, and a conservation threshold. This is the step when the sequences of MS2 and BZ13 were combined here. Genomic sequences and PS profiles were concatenated separately. The list of combined genomic sequences was input to ClustalΩ for MSA (Goujon et al., 2010; Sievers et al., 2014). This was utilised to shift the PS profiles accordingly to be able to compare corresponding genomic regions. The blocks were then created by going through the aligned PS profiles one nucleotide position at a time and checking if a proportion higher than the threshold of the sequences had a PS of any affinity at that position. If yes, a new block was started or the current one continued. If not, the current block was terminated or no new one started. The finished block lengths were divided by the expected SL length and split into two or more blocks if needed. Next, each defined block was considered and for each sequence it was determined whether it had a PS in that block or not. Block membership was encoded the same way as the PS profiles. These lists were utilised as input for SplitsTree4 to generate phylogenetic trees (Huson and Bryant, 2006).

6.2.4 Phylogenetic Trees of MS2 and BZ13

MS2 and BZ13 are different species of the *Levivirus* genus. Previously, the new method for phylogeny based on PS profiles had only been applied to different strains of one viral species. Broadening the application to more distantly related viruses comes with new challenges. The generation of PS blocks is dependent on an MSA of all strains and a conservation threshold. Reliable MSA become more difficult the lower the sequence identity of the sequences to be aligned. To check the similarity of the genomes used in this study the MSA was supplied to

TABLE 6.2: Percent of identical nucleotides of aligned *Levivirus* genomes.

Strain	**GA**	**DL20**	**KU1**	**DL10**	**T72**	**DL16**	**J20**	**R17**	**ST4**	**MS2**	**DL52**	**DL1**	**M12**	**fr**
GA	100	93	83	93	82	50	50	50	50	50	51	51	48	48
DL20	93	100	82	92	83	52	52	51	51	51	53	51	49	49
KU1	83	82	100	83	88	50	50	50	50	50	51	50	48	49
DL10	93	92	83	100	84	50	50	50	50	50	51	50	48	48
T72	82	83	88	84	100	50	49	49	49	49	50	50	47	48
DL16	50	52	50	50	50	100	95	92	92	92	88	95	86	76
J20	50	52	50	50	49	95	100	92	92	92	85	94	86	76
R17	50	51	50	50	49	92	92	100	97	96	83	92	87	76
ST4	50	51	50	50	49	92	92	97	100	99	83	92	87	76
MS2	50	51	50	50	49	92	92	96	99	100	83	92	87	76
DL52	51	53	51	51	50	88	85	83	83	83	100	84	80	71
DL1	51	51	50	50	50	95	94	92	92	92	84	100	87	76
M12	48	49	48	48	47	86	86	87	87	87	80	87	100	71
fr	48	49	49	48	48	76	76	76	76	76	71	76	71	100

the "Ident and Sim" tool of the Sequence Manipulation Suite (Stothard, 2000). The percent identity scores, i.e. the percentage of identical nucleotides in aligned sequences, for different strains of one species were between 82% and 93% for BZ13 and between 71 and 99% for MS2. Between species identity values were close to 50%. This means that after alignment only about 50% of positions are matched with identical nucleotides. The values for each comparison are shown in Table 6.2.

The MSA was utilised to shift the individual PS profiles into corresponding positions and assign PS blocks. The blocks were defined using a conservation threshold of 25%. Despite the relatively high threshold, this identified 96 blocks. As mentioned above, only 60 PSs are expected in each genome. Having more blocks than expected PSs indicates a lower level of conservation or more noise. The difference can be determined by looking at how many PSs were originally identified for each strain. As seen in Table 6.3 1–3 high, 12–23 medium, and 37–59 low affinity PSs were identified. The total numbers ranged from 56 up to 82 PSs. This shows that for all but two sequences the program identified an abundance of PSs. Most of these are low affinity PSs. It is possible that there are more of these than necessary and either can do the job for capsid assembly. This would make the mechanism more robust against mutations of some PSs. However, even when assuming the program predicted a suitable number of PSs, which is higher than 60 for most sequences, 96 blocks is still more than the highest number of PSs in any strain. This indicates lower levels of conservation between the strains, which is in line with the low percentage identity.

The PS blocks including affinity information were supplied to SplitsTree4 (Huson and Bryant, 2006) and a phylogenetic tree was reconstructed (Figure 6.3 left). For comparison another tree was reconstructed from the MSA (Figure 6.3 right). As before the neighbor-joining (NJ) method with Hamming distances was used. This means that any difference in characters was treated equally instead of applying a more sophisticated mutation model. As expected, both resulting trees

TABLE 6.3: Numbers of high, medium, and low affinity PSs identified in the *Levivirus* genomes.

	Strain	High	Medium	Low	Total
	GA	2	18	43	63
	DL20	1	20	43	64
BZ13	KU1	1	18	37	56
	DL10	2	16	45	63
	T72	2	23	46	71
	DL16	3	12	56	71
	J20	3	15	57	75
	R17	4	17	57	78
	ST4	3	19	59	82
MS2	MS2	3	18	59	80
	DL52	3	15	52	70
	DL1	3	18	55	76
	M12	3	13	50	66
	fr	4	16	55	75

separated BZ13 and MS2 strains but the relative distance was higher for the MSA tree. Also the splits within each species cluster were similar especially for BZ13, which was split into topologically identical clusters. This was less the case for MS2, where there were some changes to the tree topology. MS2 strain fr, which had the lowest percent identity to any other MS2 strain, was unsurprisingly most distant on both trees as well.

Due to the abundance of PSs in these strains a comparison by compact PS profile as done for HBV was not feasible. In order to better understand PS conservation between these two viral species, a comparison of PS positions in the complete structure by manual structure alignment was performed below.

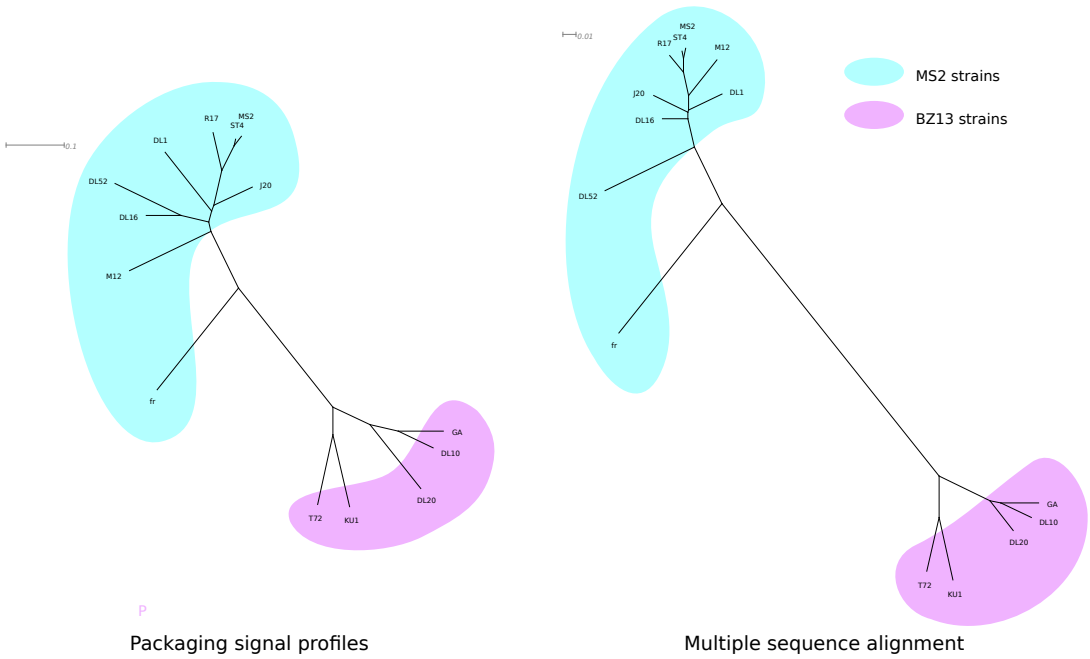

FIGURE 6.3: **Phylogenetic trees of MS2 and BZ13 bacteriophage strains based on PS profiles and genomic sequence MSA.** For the PS profile tree (left) the PS blocks were utilised. The conservation threshold was set to 25% for the PS blocks resulting in 96 blocks out of which 93 were informative. For the sequence based tree a MSA in ClustalΩ (Goujon et al., 2010; Sievers et al., 2014) was used. Both are NJ trees using Hamming distance and were created in SplitsTree4 (Huson and Bryant, 2006). BZ13 strains are highlighted in pink and MS2 strains are highlighted in blue.

6.3 Conservation of Packaging Signals in *Leviviridae*

The MSA of BZ13 and MS2 genomic sequences revealed only 50% sequence identity between the species. This makes the reliability of the alignment questionable and consequently also the PS-based phylogeny, which uses this alignment to match up corresponding PSs. To better understand conservation of PSs between species of the *Levivirus* genus or the *Leviviridae* family in general, they have to be matched without use of fast-evolving sequence. Since the method detailed in Chapter 2 and used above has not been developed to that extent, yet, a manual comparison of representative strains of MS2, BZ13 (KU1), and Qβ species was performed.

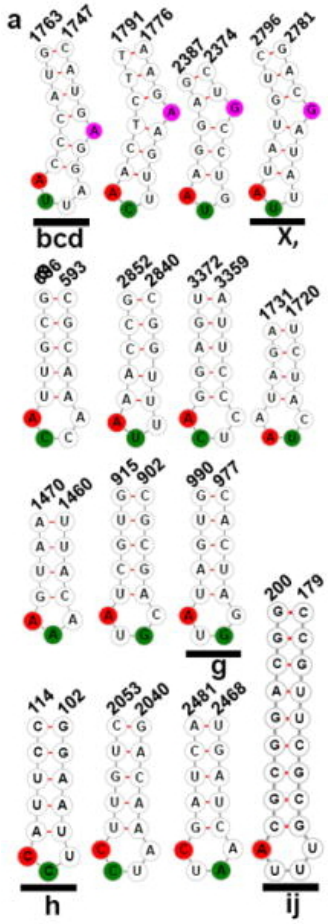

FIGURE 6.4: **Packaging signals identified by Dai *et al.*** These stem-loops were fitted into the cryo-EM densities seen within the capsid. Figure 3a as published in Dai et al. (2017). Copy right cleared with Elsevier through Copyright Clearance Center, license number 4525910684826.

As explained earlier, nucleotide sequence, especially viral genomic RNA, is very fast evolving and thus changes rapidly. To better compare between viral species they were aligned by RNA secondary structure as published in doctoral theses by the van Duin group at the University of Leiden (Olsthoorn, 1996; Beekwilder, 1996; Groeneveld, 1997) and the same nomenclature for the SLs was used. Despite the low sequence similarity the genomes were strikingly similar in secondary structure. The MS2, KU1, and Qβ structures were drawn simplified and adjusted to better illustrate equivalent SLs. For easier visualisation, the comparison was mostly done by RNA structural domain. Each SL was manually checked for the respective PS motif of that viral species. For Qβ, each SL with three nucleotides in the apical loop and the last one being an "A" was considered a PS. Special attention was given to the "Hong" PSs (Figure 6.4). Apart from TR these are the only ones in MS2 experimentally shown to be in contact with CP dimers in an assembled capsid. These PSs and equivalents in the other viruses were marked with red dots and other SLs with a PS motif with green dots in the simplified structures (Figures 6.5 and 6.6). This enabled easy comparison and determination of conservation within this viral family.

Aligning structures of MS2 and KU1 was comparatively easy. Their close relatedness (same genus) was apparent through the high degree of similarity on the secondary RNA structure despite low sequence similarity. Whilst some SLs were rearranged into more SLs such as in Figure 6.5B the general organisation was mostly the same. This was not the case fore Qβ, which is from the *Allolevivirus* genus. Due to partially large changes to the overall structure of the maturation (A) and coat protein (C) domains, alignment of equivalent regions required bundling some parts of the domains (Figures 6.5A and 6.6A). In MS2 and KU1 maturation protein expression is regulated through the interaction of the Shine-Dalgarno (SD) sequence with a region upstream, the upstream complementary sequence (UCS) (Poot et al., 1997). The maturation protein also serves as lysis protein in Qβ (Winter and Gold, 1983) requiring it to be expressed

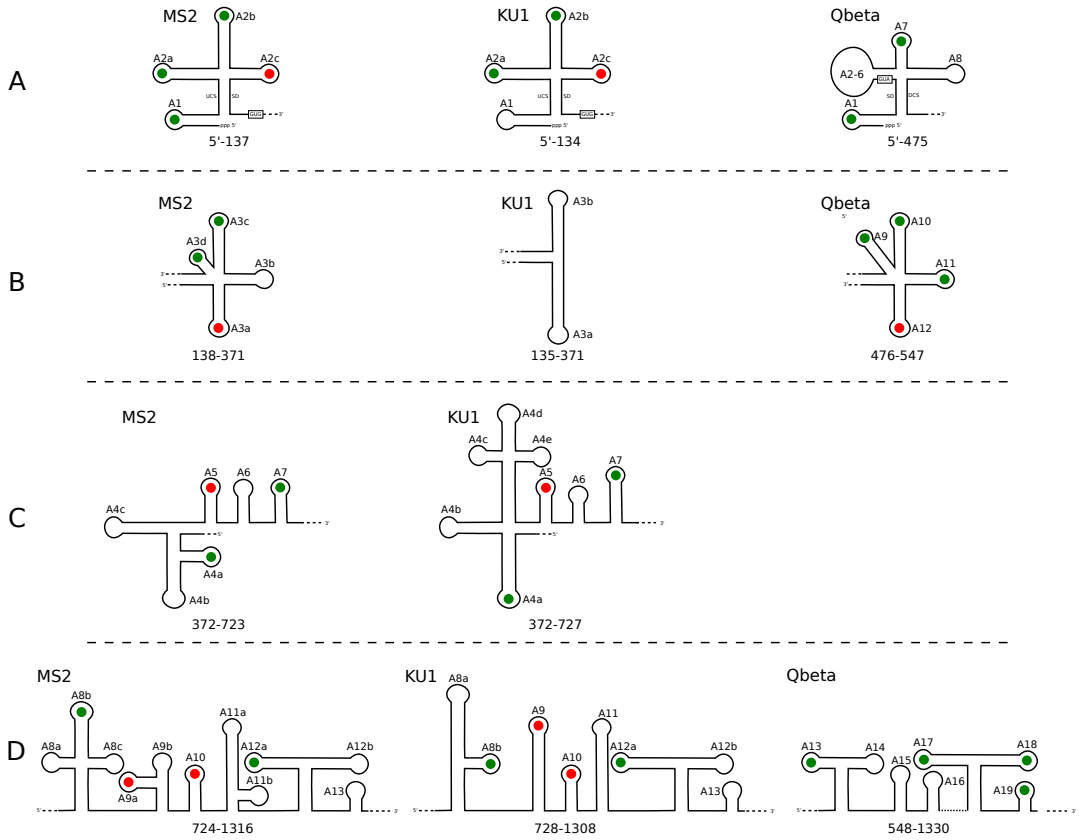

FIGURE 6.5: **Comparison of RNA structures and PS positions between MS2, KU1, and Qβ in maturation protein region.** The structures are organised to allow an approximate alignment of corresponding SLs. Stretched sequence is shown as dotted lines between continuous lines. Larger domains that have been left out and shown as large dotted circles. SLs with a respective PS motif are shown as coloured dots with red being Hong PSs and equivalent in the other viruses and green being all others PSs. The nomenclature is relative to Olsthoorn (1996), Groeneveld (1997), and Beekwilder (1996). (A) The first RNA structural domain is shown starting at the 5' end and continuing until nt 137, 134 and 475 in MS2, KU1, and Qβ, respectively. The structures are aligned by the position of the Shine-Dalgarno (SD) sequence and the respective upstream or downstream complementary sequence (UCS or DCS). (B) The next domain in the maturation protein region stretches until nt 371 in MS2 and KU1 or nt 547 in Qβ. The alignment follows from the previous domain. (C) The third domain shown in MS2 until nt 723 and in KU1 until nt 727 is not present in Qβ due to the large differences in tertiary structure. (D) The final domain in the maturation protein region ends before the start of the coat protein open reading frame (ORF). While MS2 and KU1 were very similar, alignment of Qβ required opening of a basepaired region between the sequences before A15 and after A17/A18. Additionally, the part between A16 and A17/A18 was stretched.

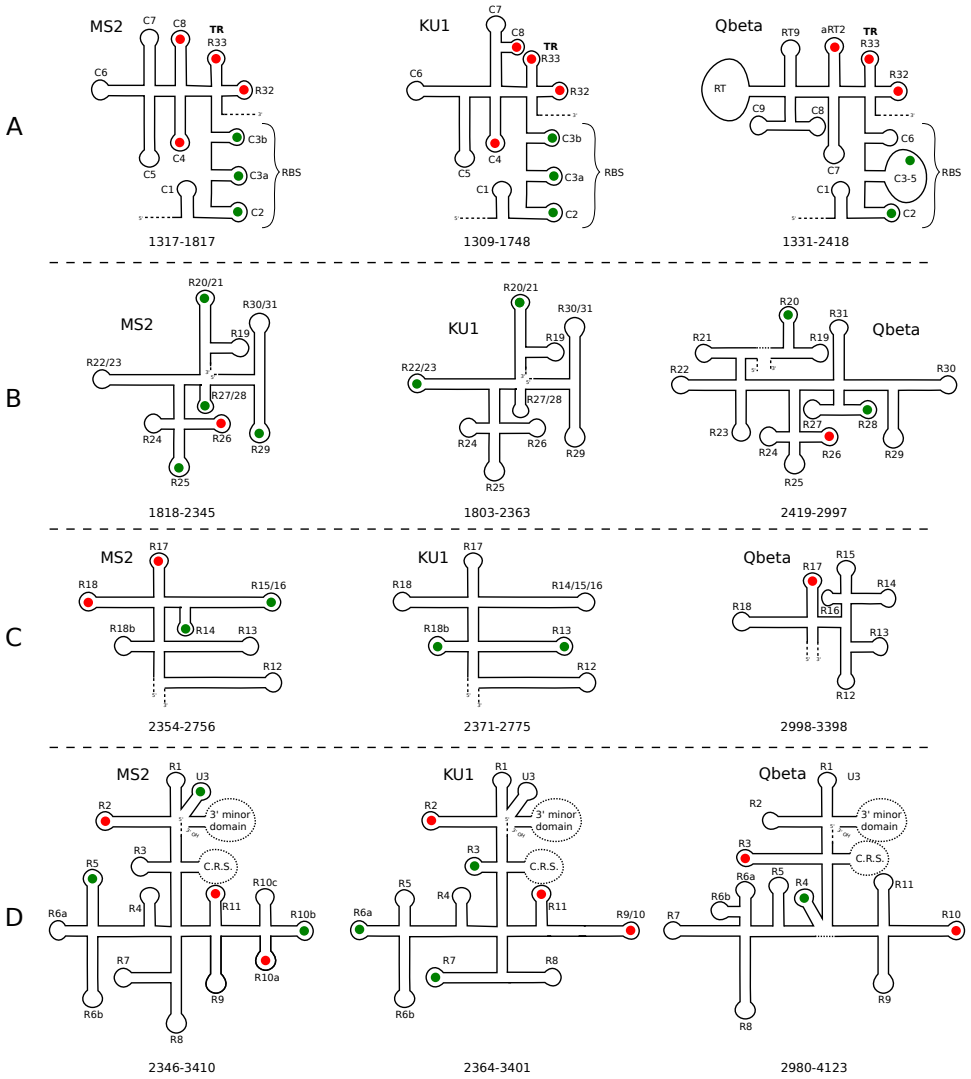

FIGURE 6.6: **Comparison of RNA structures and PS positions between MS2, KU1, and Qβ in coat and replicase region.** The structures are organised to allow an approximate alignment of corresponding SLs. Stretched sequence is shown as dotted lines between continuous lines. Larger domains that have been left out and shown as large dotted circles. SLs with a respective PS motif are shown as coloured dots with red being Hong PSs and equivalent in the other viruses and green being all others PSs. The nomenclature is relative to Olsthoorn (1996), Groeneveld (1997), and Beekwilder (1996). (A) The alignment of the RNA structure in the complete coat region is shown including the ribosome binding site (RBS) as well as the end of the replicase region and the readthrough (RT) region for Qβ. R33 is TR and equivalents in the other viruses. Due to RT in Qβ the alignment required condensing of parts of the structure as shown by larger loops. (B) The next domain in the replicase region is shown. It spans until nt 2345, 2363, and 2997 in MS2, KU1, and Qβ, respectively. To better visualise equivalent SLs, the 3' region encompassing R20 and R19 was slightly rotated and stretched in Qβ ensuring the same orientation of these SLs in all viruses. (C) The conserved replicase subdomain (C.R.S.) is shown. (D) The final domain of the replicase region is shown as well as the positions of the C.R.S. (see (C)) and the 3' minor domain (not shown). The orientation of the Qβ structure between R9 and R8 was changed for the alignment.

for longer. Therefore, the interaction partner is downstream of the SD sequence so a downstream complementary sequence (DCS) (Beekwilder et al., 1996). It is this interaction that was aligned in Figure 6.5A. Furthermore, Qβ contains an additional domain, the readthrough (RT) domain, between the coat and the replicase domains. It was also mostly bundled for the alignment (Figure 6.6A). These bundles can be thought of as analogous to indels in sequence alignments.

The alignment of structural domains revealed a high number of conserved PSs. All SLs with a PS motif are summarised in Table 6.4 with Hong PSs in bold. In total, 34 PSs were identified for MS2, 28 for KU1, and 16 for Qβ. These numbers are considerably lower for MS2 and KU1 than from the SL selection method described in Chapter 2. Here 20 PSs were found to be conserved between MS2 and KU1, whereas only 15 were conserved between MS2 and Qβ. Even fewer, namely nine, PSs were present in both KU1 and Qβ all of which were also in MS2. These included six Hong PSs. Interestingly, the conserved Hong PSs were TR and the SLs just up and down stream of it (C8/aRT2, R33 (TR), and R32) as well as one PSs at the 5' end (A2c/A7) and two at the 3' ends (R10a/9/10 and R2/R3). The positioning of the conserved PSs may point towards functional importance of these structures. Their position relative to maturation protein as seen in the structure by Dai et al. (2017) is shown in Figure 6.7. TR, its neighbours and the 3' most conserved PS are all in close proximity to each other and MP in the capsid. R10a/9/10, which is further upstream in the sequence, is also nearby, whilst the PS at the 5' end, A2c/A7, was mapped to the other side of the capsid.

TABLE 6.4: Packaging signals in global thermodynamic structures of MS2, KU1, and Qβ with Hong PSs marked bold.

MS2	**KU1**	**Qβ**
A1	–	A1
A2a	A2a	–
A2b	A2b	–
A2c	**A2c**	**A7**
A3a	–	**A12**
A3b	–	–

TABLE 6.4: (continued)

A4a	A4a	–
A5	A5	–
A7	A7	–
A8b	–	–
A9	A9	–
A10	A10	–
A12a	–	A17
–	A12b	–
–	A13	–
C2	C2	C2
C3a	C3a	C3
C3b	C3b	–
C4	**C4**	–
C8	**C8**	**aRT2**
TR(R33)	**R33**	**R33**
R32	**R32**	**R32**
R27/28	–	R28
R26	–	**R26**
R25	–	–
–	R22/23	–
R20/21	R20/21	R20
–	R18b	–
R18	–	–
R17	–	**R17**
R15/16	–	–
R14	–	–
–	R13	–
R11	**R11**	–
R10b	–	–
R10a	**R9/10**	**R10**
–	R7	–
–	R6a	–
R5	–	R4
–	R3	–
R2	**R2**	**R3**
U3	U3	–

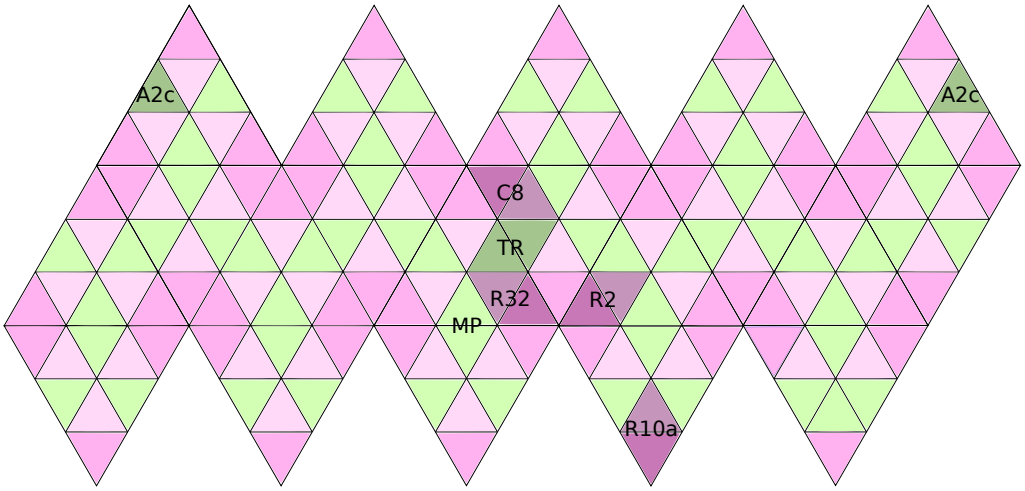

FIGURE 6.7: **Conserved Hong PSs in lattice.** AB dimers are shown in pink and CC dimers are shown in green in the lattice. Maturation protein replaces one CC dimer and is labelled MP. The position of the conserved PSs in the structure by Dai et al. (2017) relative to MP are shown shaded darker and labelled with SL name.

6.4 Discussion

Coliphages from the *Leviviridae* family have been studied extensively for many decades. Consequently, much is known about their RNA genome structure and PS-mediated assembly in them. In this chapter this was used to apply the PS-based phylogeny method introduced in Chapter 2 to viruses with known PS affinity tiers as well as comparing between viral species. The SL selection algorithm identified an abundance of SLs with PS motif, especially of low affinity. Only 1–3 high affinity PSs were found by the program. This is in line with previous research on PS-mediated assembly, which showed that assembly models performed better if PSs with different affinities were present on the RNA whereas RNA with only high affinity PSs performed worst (Dykeman et al., 2013a). An overabundance of low affinity PSs demonstrates a robustness of the mechanism against mutations in many SLs. How many of these low affinity structures are actually needed specifically is not known. It is also possible that additional PSs are the result of alternative assembly paths. PS-mediated assembly in MS2 is thought to follow a

Hamiltonian path (Dykeman et al., 2013b). Over 40,000 such paths are possible to give rise to the MS2 capsid; however, it is unlikely that the virus employs all or any of these. Rather further biological restrictions, such as the maturation protein contact sites at the 5' and 3' ends of the genomic RNA, would limit the number of actually used options. Instead of always using one particular path, the virus may utilise a small set and which path of these is used depends on the subset of PSs employed or vice versa.

Combining the PS profiles of MS2 and BZ13 to identify conserved PS blocks proved challenging. To allow for separation among the BZ13 strains a lower threshold of 25% was chosen. With this over 90 blocks were identified, which means more blocks than expected PSs for either strain. A test of percent identity in the MSA utilised for the blocks showed 50% identity between the species. Whilst this is not considered a bad value per se, it does illustrate the lower level of similarity between the species and challenges of matching up the correct regions. This is crucial for PS-based phylogeny because the conservation threshold used for the blocks requires the PSs to at least be at overlapping aligned nucleotide positions. More overall blocks than PSs indicates that the PSs were not matched fully, which can either be due to little conservation or unsuccessful alignment. To improve the phylogeny algorithm and make it better applicable to more distantly related viruses it will have to be made independent of an MSA of the genomic RNA sequences.

The phylogenetic trees reconstructed from the PS blocks and the MSA itself were very similar with largely the same topology. For either a clear separation of the species was seen albeit to a lesser extent relatively in the PS-based trees. Among the MS2 strains fr was the most removed from the rest, which was not surprising considering it also had the lowest (around 70%) sequence identity to the other strains. Hamming distances were utilised for tree building. These treat every difference between sequences the same. Whilst this was appropriate when applied to HBV, where only one affinity was used, it is less clear here. Since

affinity tiers were encoded in the block membership as well, which was the input to the tree building program, changes in affinity were treated the same as going from having a PS to none. Biologically, however, these shifts should not be considered equivalent. For the virus completely losing one of the few high affinity PSs would be much more detrimental than losing a low affinity one or downgrading medium to low affinity. In the future it would, therefore, be better to incorporate an own mutation model, which better reflects the respective transitions.

To understand conservation between different *Leviviridae* species better the comparison was done manually. Only MS2 (genogroup I), BZ13 (genogroup II), and Qβ (genogroup III) were included. For NL95 (genogroup IV) a complete structure was not available for comparison. A much lower number of PSs was found when only considering the SLs from the published global structure compared to the SL selection algorithm. This is due to the algorithm folding locally on 30 nucleotide overlapping fragments meaning that no single SL can be larger than that, whereas the phage structures contain much larger structures. Whilst this may seem like the algorithm is overshooting the number of SLs, a previous study suggested that local refolding may occur in parts of the MS2 genome revealing additional PSs (Dykeman et al., 2013b). These PSs available through refolding may account for the additional ones found by the SL selection program.

From all PSs identified in the secondary structures of the genomic RNAs only nine were conserved among all three viruses, MS2, KU1, and Qβ. Among these were six Hong PSs. The PSs are located in interesting positions in the genomes: Three are TR and its neighbouring PSs, and the others are at each RNA end. Inside the capsid in the asymmetric structure, TR and its neighbours as well as the PS closest the 3' end are located in close proximity to each other and maturation protein (Dai et al., 2017). The level of conservation, which extents even to Qβ from a different genus, and the close proximity to each other in the 3D structure indicate a functional importance of these PSs. The other conserved PSs likely perform other important functions, which were out of the scope of this

project to study.

Maturation protein is integrated in the capsids and takes the place of a homodimer thus breaking the icosahedral symmetry (Dent et al., 2013). In all mentioned viruses it is essential for infection of a new host cell as it binds to the F-pilus (Crawford and Gesteland, 1964; Lodish et al., 1965). To ensure that it is indeed present in progeny viruses and not accidentally left out when the capsid is built up, it binds to the genomic RNA in two places (Shiba and Suzuki, 1981; Rumnieks and Tars, 2017). In MS2 the interaction sites were mapped to positions 388–398 and 3510–3520 in the genome so close to either end of the RNA (Shiba and Suzuki, 1981). Maturation protein is therefore likely to be incorporated early on into the growing capsid.

TR stands for translation repressor because binding of CP to this SL inhibits translation of replicase. This is thought to be the first step in the initiation of capsid assembly. The results from the conservation analysis point towards more PSs being involved in the nucleation of assembly. We, thus, hypothesise that a nucleation complex is formed from several CP dimers bound to the above mentioned PSs and from maturation protein. How this insight could affect capsid assembly was tested in the following chapters in the form of computational MS2 assembly model.

Chapter 7

Gillespie Model of Virus Assembly Using Extended Nucleus

7.1 Modelling Chemical Reactions

Typically the mathematical modelling of chemical reactions involves a set of coupled ordinary differential equations. For very simple systems, which involve coupled 1st order reactions, these can be solved analytically resulting in a closed form solution for the kinetics for each species $C_i(t)$ as

$$\begin{bmatrix} C_1(t) \\ C_2(t) \\ C_3(t) \end{bmatrix} = \mathrm{e}^{\boldsymbol{A}t} \begin{bmatrix} C_1(0) \\ C_2(0) \\ C_3(0) \end{bmatrix}.$$

$\boldsymbol{A}$ is a matrix containing the transition rates. This way the state of molecular species in a system can be predicted at any point in time. However, most systems interesting enough to study are too complex for this approach. Already a relatively simple reaction involving two chemical species reacting together to form

a third one is a 2nd order reaction. Instead, they require solving the equations numerically using computers. Numerical simulations can be further divided into two general types: deterministic or stochastic (probabilistic) (Mira et al., 2003).

7.1.1 Deterministic Models

Deterministic simulations are still the most widely used for modelling chemical reactions. The amount of chemical species present is predetermined by the rates and initial conditions. As opposed to stochastic models there is no random element involved. This way they can predict for each point in time the exact amount of each modelled molecular species. This is achieved by using a set of coupled differential equations, which are solved analytically if possible or numerically otherwise. These equations describe the changes of concentrations of molecular species over time using reaction rate constants for each type of reaction a species can undergo, e.g. production, degradation, or reaction into another species. It is assumed that given a solution to these equations, the behaviours of the molecular species can be predicted exactly. Thus, given the same starting conditions this type of model produces the exact same results in every run (Mira et al., 2003).

To exemplify, consider the following 2nd order reaction of TR binding to a capsid protein (CP) dimer:

$$\mathrm{TR} + \mathrm{CP} \underset{K_\mathrm{B}}{\overset{K_\mathrm{F}}{\rightleftarrows}} \mathrm{CP} : \mathrm{TR}$$

The on and off rates K_F and K_B are defined respectively as:

$$\frac{K_\mathrm{F}}{K_\mathrm{B}} = \mathrm{e}^{\beta \Delta G}$$
$$\frac{K_\mathrm{B}}{K_\mathrm{F}} = K_\mathrm{D} = 1.2 \text{ nM}$$

For this example the coupled ordinary differential equations are

$$\begin{aligned}\frac{\mathrm{d}[\mathrm{TR}]}{\mathrm{d}t} &= -[\mathrm{CP}][\mathrm{TR}]K_{\mathrm{F}} + [\mathrm{CP}:\mathrm{TR}]K_{\mathrm{B}} \\ \frac{\mathrm{d}[\mathrm{CP}]}{\mathrm{d}t} &= -[\mathrm{CP}][\mathrm{TR}]K_{\mathrm{F}} + [\mathrm{CP}:\mathrm{TR}]K_{\mathrm{B}} \\ \frac{\mathrm{d}[\mathrm{CP}:\mathrm{TR}]}{\mathrm{d}t} &= +[\mathrm{CP}][\mathrm{TR}]K_{\mathrm{F}} - [\mathrm{CP}:\mathrm{TR}]K_{\mathrm{B}}.\end{aligned}$$

Solving these equations enables one to find the concentrations of each chemical species, here TR, CP and CP:TR, at equilibrium, the steady state.

7.1.2 Stochastic Models

Rather than assuming that a certain reaction is going to occur based on its reaction rate, the stochastic models work under the assumption that reactions occur randomly following a set of probabilities due to particles undergoing random Brownian motion. These probabilities are related to the properties of the system so that any reaction has a reaction probability constant, which is obtained from the likelihood of collision, analogous to the reaction rate constants used in deterministic models. Chemical species are modelled as "hard spheres" that periodically collide. The probability of this collision can be solved (see Gillespie (1977)) giving a probability of collision based on diffusion rate and number of particles. Each run of a stochastic model given the same starting conditions produces a different trajectory of the evolution of chemical reactions over time. If run many times, the average is the same as the single outcome of a deterministic model (Mira et al., 2003). Another important difference between deterministic and stochastic approaches is that the deterministic, through the rate laws applied, assumes continuous change in the chemical species. However, molecules can only react, be produced or degraded as whole units meaning that steps would need to be discrete rather than continuous. This makes a stochastic simulation a more accurate description of reaction kinetics because it can consider reactions with single particles (de Levie, 2009). The deterministic approach assumes systems

large enough where one can consider the overall behaviour of the system as average and small stochastic fluctuations are evened out. Under such conditions it produces accurate enough results (Samoilov and Arkin, 2006); however, the model quickly becomes inaccurate when dealing with systems that are strongly influenced by just a few molecules such as nucleation (de Levie, 2009; McAdams and Arkin, 1999; Martinez-Urreaga et al., 2003; Freeman, 1984).

7.1.3 The Gillespie Algorithm

In a spaciously uniform system, i.e. "well-stirred", of defined volume a stochastic model can be built using a chemical master equation. This describes the change in probability over time for the system occupying any possible state. The more possible states there are, the more complex the equation becomes. It is therefore only realistically solvable for simple systems, i.e. 1st order systems. In order to work with larger, more complex systems other algorithms are utilised to simulate these.

The "stochastic simulation algorithm", better known as the Gillespie algorithm, provides a method for stochastic simulation without the need for a master equation. It is a Monte Carlo method (Gillespie, 1977; Martinez-Urreaga et al., 2003) in that it uses probability density functions (PDFs) to stochastically sample a series of reactions and their time of occurrence. The PDF describes the probability of a random variable to take on a particular value. In practical terms, the probability is calculated from the area under the curve. For a continuous variable this is only meaningful for a range of values as the integral from a to a and thus the probability of any discrete value is zero (Kiran and Kiran, 2017, Chapter 27). Additionally, the Gillespie algorithm is also a type of Markov process, where the behaviour of the system is independent of previous states.

To illustrate how the Gillespie algorithm works, let us consider a system of N different chemical species or molecules, whose number in some fixed volume V is given by X_i with $i = 1, ..., N$ and which can undergo M different chemical

reactions R_μ with $\mu = 1, ..., M$. If the molecules are considered well mixed, i.e. evenly distributed over the volume, and are modelled as reactions when undergoing a hard sphere collision, then the probability that reaction R_μ occurs in some time interval $\tau + \mathrm{d}\tau$ can be given by a joint PDF for the statistically independent random variables τ and μ

$$\mathrm{P}(\tau, \mu) = \begin{cases} a_\mu \exp(-a_0 \tau) & \text{if } 0 \ \leq \tau < \infty \text{ and } \mu = 1, .., M \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

where the times step τ to the next reaction is an exponential random variable and the index of the firing reaction μ is a random integer variable (Gillespie, 2007). Here P(τ,μ) is the probability that reaction R_μ will fire in the next time increment τ while a_μ is the rate at which reaction μ occurs per unit time. The quantity a_0 is the sum over all reaction rates a_μ, i.e.

$$a_0 \equiv \sum_{\nu=1}^{M} a_\nu \equiv \sum_{\nu=1}^{M} h_\nu c_\nu \ . \tag{7.2}$$

Here, h is the number of unique combinatorial ways that the reactants can combine for the given state. For instance, if in reaction R_μ molecules A in V could react with molecules B into a complex AB, then $h_\mu = X_\mathrm{A} X_\mathrm{B}$ so if there were 10 molecules of A and 10 molecules of B in V, then $h_\mu = 100$. The stochastic reaction constant c is the rate per unit of time that a single reaction between any of the specific molecules characterising reaction μ will occur, e.g. in this example it could be A_4 with B_7 or A_8 with B_2 or any of the other $X_\mathrm{A} X_\mathrm{B}$ possible combinations. It should be noted that in most chemical reaction cases, in particular the ones considered for virus assembly in this thesis, a_μ can be approximated as a time-independent constant.

From the PDF in Equation (7.1), the probability that a specific reaction R_μ will take place between two time points a and b, can thus be obtained by calculating the area under the curve between a and b for that μ:

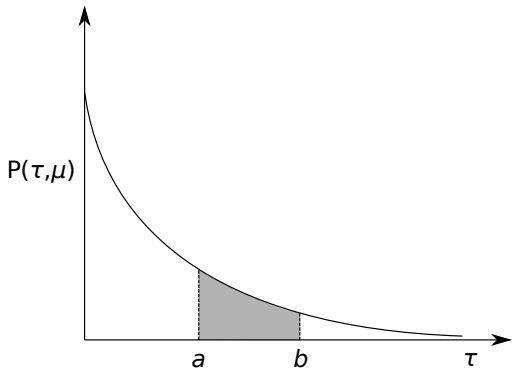

FIGURE 7.1: **Example of a reaction probability function P(τ,μ) for a reaction μ.** The probability for reaction μ to fire in a time interval [a,b] is given by the area under the curve between these two points (grey).

$$\int_a^b \frac{a_\mu}{a_0} a_0 \exp(-a_0\tau)\mathrm{d}\tau. \tag{7.3}$$

This is illustrated in Figure 7.1. The probability of reaction R_μ to fire at any point in time is thus

$$\int_0^\infty a_\mu \exp(-a_0\tau)\mathrm{d}\tau, \tag{7.4}$$

which is the same as $\frac{a_\mu}{a_0}$. As there are M different reactions in the system, there are M different reaction probabilities and the sum of all of these is 1:

$$\sum_{\nu=1}^{M} \frac{a_\nu}{a_0} = 1 \ . \tag{7.5}$$

The two independent variables τ and μ, the time step to the next reaction and the reaction, are at the heart of the Gillespie algorithm. In simple terms, both these variables are repeatedly sampled resulting in a series of reactions as the total simulation reaction time progresses. The sampling is based on the PDF in Equation (7.1). There are different approaches for this. In the original Gillespie paper (Gillespie, 1977) and in this thesis the inversion generating or Direct method was used. It works by using a random number r from a uniform distribution between 0 and 1 and solving the inverse of the cumulative distribution function (CDF) for r (Gillespie, 1991, Chapter 1). The CDF represents the prob-

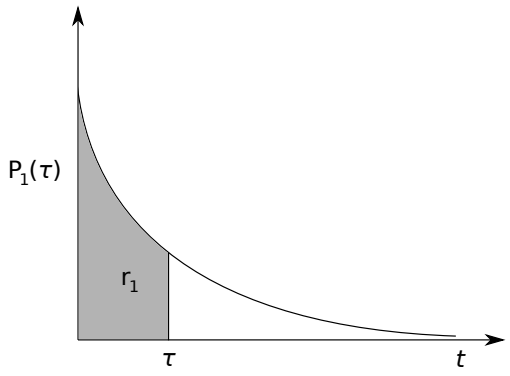

FIGURE 7.2: **Illustration of the meaning of random number r_1.** The probability density function $P_1(\tau)$ is shown. The area under the curve between 0 and a time point τ (grey) represents the probability of any reaction firing in that time span, which is equal to the random number r_1. The τ for which the area under the curve equals r_1 would be the time to the next reaction.

ability of the random variable being equal to or lower than the respective value on the x-axis. For easier calculation, the function is separated into its marginal PDFs, which are sampled using two random numbers r_1 and r_2. $P(\tau, \mu)$ can be rewritten as:

$$\begin{aligned} P(\tau, \mu) &= a_\mu \exp(-a_0 \tau) \\ &= \frac{a_\mu}{a_0} a_0 \exp(-a_0 \tau) \end{aligned} \tag{7.6}$$

and since τ and μ are statistically independent the joint probability function can be separated into its marginals using $P(\tau, \mu) = P_1(\tau)P_2(\mu)$:

$$\begin{aligned} P_1(\tau) &= a_0 \exp(-a_0 \tau) \\ P_2(\mu) &= \frac{a_\mu}{a_0} \end{aligned} \tag{7.7}$$

$P_2(\mu)$ in Equation (7.7) is the probability of reaction R_μ to fire at any point in time and the same as Equation (7.4). On the other hand, $P_1(\tau)$ represents the probability density for the time to any next reaction firing being τ. Separating the PDFs illustrates that τ has an exponential distribution with rate parameter λ being a_0. The probability that any reaction will fire by said time point can be calculated through the area under the curve from 0 to that time point (Figure 7.2).

Utilising two random numbers between 0 and 1 allows sampling from P_1 and P_2 resulting at each step in a randomly selected time step τ and reaction R_μ to fire within it. In order to stochastically determine the time step τ until the next reaction fires, τ is calculated such that the CDF for τ is equal to r_1. This mean that the area under the curve from 0 to τ in P_1 should be equal to r_1:

$$\int_0^\tau a_0 \exp(-a_0 t)\mathrm{d}t = r_1 \tag{7.8}$$

Integration of the left side gives

$$a_0(-1/a_0)\exp(-a_0 t)\Big|_0^\tau = r_1 \ , \tag{7.9}$$

which can be simplified to

$$-\exp(-a_0 t)\Big|_0^\tau = r_1 \ . \tag{7.10}$$

Evaluating the integral results in

$$-\exp(-a_0\tau) - (-\exp(-a_0 0)) = r_1 \ , \tag{7.11}$$

which, given that $\exp(0) = 1$, simplifies to

$$-\exp(-a_0\tau) + 1 = r_1 \ . \tag{7.12}$$

The left hand side is the CDF of this exponential distribution. This can be rearranged to

$$1 - r_1 = \exp(-a_0\tau) \ . \tag{7.13}$$

Since r_1 is a random number between 0 and 1, $1 - r_1$ is also a random number between 0 and 1. Equation (7.13) can therefore be simplified to:

$$r_1 = \exp(-a_0\tau) \ . \tag{7.14}$$

Taking the natural logarithm results in

$$\ln(r_1) = -a_0\tau \ . \tag{7.15}$$

Dividing by $-a_0$ solves for τ:

$$\tau = -(1/a_0)\ln(r_1) \ . \tag{7.16}$$

Since $-\log(x) = \log(1/x)$, this can be finally simplified to

$$\tau = (1/a_0)\ln(1/r_1) \ , \tag{7.17}$$

which is the formula through which the time step τ is calculated from the random number r_1 in the Gillespie algorithm implementation (Gillespie, 1977).

The CDF for the discrete variable μ would be the sum of $\frac{a_\mu}{a_0}$ for all values of μ (Equation (7.5)). Therefore, μ can be calculated from r_2, which is also a random number between 0 and 1, by applying:

$$\sum_{\nu=1}^{\mu} \frac{a_\nu}{a_0} \geq r_2 \ . \tag{7.18}$$

Multiplying both sides by a_0 gives

$$\sum_{\nu=1}^{\mu} a_\nu \geq a_0 r_2 \ . \tag{7.19}$$

Theoretically this means that μ is determined in such a way that r_2 multiplied by the total reaction probability a_0 lies between the sum of the reaction probabilities up until reaction $\mu - 1$ and up until reaction μ (Gillespie, 1977):

$$\sum_{\nu=1}^{\mu-1} a_\nu < r_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu \ . \tag{7.20}$$

Since, as opposed to τ with Equation (7.17), no direct calculation is possible,

practically, ν is incremented by 1 until the condition in Equation (7.19) is met.

The Gillespie algorithm can then be implemented by repeating three steps after an initial set-up. This set-up includes initialising the numbers X_i of each of the N chemical species and the random number generator. Then, step (1) is to calculate the reaction rates a_μ for all M different reactions. Since, as seen in Equation (7.2), these depend on the numbers of molecules that can react together it has to be calculated in every round. The same is true for the total reaction propensity a_0, which is the sum of all individual rates. In step (2) the random numbers r_1 and r_2 are utilised to calculate τ and μ as showed in Equations (7.17) and (7.19), respectively. In step (3) the total time t of the simulation is increased by τ and the numbers of chemical species X_i are updated to reflect the firing of reaction R_μ. These steps are repeated until t reaches a predefined maximum time value for the simulation (Gillespie, 1977).

7.2 Application to Virus Assembly

The binding of viral CPs to packaging signals (PSs) and assembly into a full capsid can be understood as a series of chemical reactions. In the simplest form, first, a CP binds a PS forming a reversible complex CP:PS. This step depends on the CP concentration and PS affinity for CP and is a second–order reaction. Next, two CPs with bound PS bind to each other or a new CP binds to a growing complex. These reactions depend on the CP:CP binding energies only and are first-order reactions. For the purpose of modelling they can be understood as follows:

$$\begin{aligned}
\mathrm{R}_1 &: \mathrm{CP} + \mathrm{PS} \rightarrow \mathrm{CP{:}PS} \\
\mathrm{R}_2 &: \mathrm{CP{:}PS} \rightarrow \mathrm{CP} + \mathrm{PS} \\
\mathrm{R}_3 &: \mathrm{CP{:}PS} + \mathrm{CP{:}PS} \rightarrow \mathrm{CP}_2 \\
\mathrm{R}_4 &: \mathrm{CP}_2 \rightarrow \mathrm{CP{:}PS} + \mathrm{CP{:}PS} \\
\mathrm{R}_5 &: \mathrm{CP{:}PS} + \mathrm{CP}_x \rightarrow \mathrm{CP}_{x+1} \\
\mathrm{R}_6 &: \mathrm{CP}_{x+1} \rightarrow \mathrm{CP{:}PS} + \mathrm{CP}_x
\end{aligned} \tag{7.21}$$

Here, R_1 and R_2, R_3 and R_4, and R_5 and R_6 are respective forward and reverse reactions and x is the size of the CP complex ranging from 2 to the number of CPs needed to form a complete capsid. R_5 and R_6 are thus describing a whole set of reactions. This becomes even more complex when each unique configuration of the RNA path within the CP complex and respective intermediated are to be considered (Dykeman et al., 2013a). Before, when the assembly process had been regarded with respect to CP only, capsid assembly could easily be modelled deterministically by solving the respective set of differential equations (Zlotnick, 1994; Endres and Zlotnick, 2002). However, when PSs are included, these become more complex and difficult to solve (Dykeman et al., 2013a). Therefore, the reactions (Equations (7.21)) and consequently the assembly of viral capsids from CPs and RNA (PSs) are modelled using the Gillespie algorithm where capsid configurations are enumerated on a graph (Tarjan, 1971). The implementations by Dr Eric Dykeman for different systems and my modifications of them will be described in the following sections.

7.2.1 The Dodec Model

The simplest implementation of the Gillespie algorithm for viral assembly is the dodec model. It is has been modified from models developed by Zlotnick (1994), Jamalyaria et al. (2005), and Sweeney et al. (2008) and was described in detail in Dykeman et al. (2013a). The viral capsid is assumed to take the shape of a dodecahedron, a regular shape made up of twelve pentagons (Figure 7.3a). Each

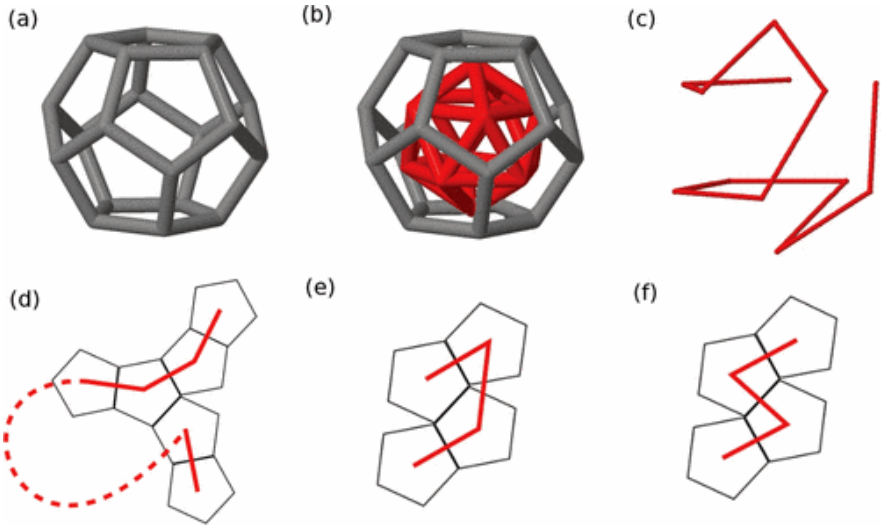

FIGURE 7.3: **Dodecahedron model and associated path rules.** (a) A dodecahedron consists of twelve pentagons and is utilised as a model for a viral capsid. (b) A shape that touches the middle of every pentagon in a dodecahedron with its vertices, is an icosahedron. (c) Going along the edges of an icosahedron in a Hamiltonian path describes one of the possible configurations of the RNA as modelled. (d) A disconnected path that is not allowed in the dodec model. (e) and (f) Two possible arrangements of four CPs bound to RNA forming a complex. Reprinted figure with permission from Dykeman, E. C., Stockley, P. G., & Twarock, R., Physical Review E, 87(2):022717–1–12, 2013a, `http://dx.doi.org/10.1103/PhysRevE.87.022717`. Copyright 2019 by the American Physical Society, license number RNP/19/APR/013930.

pentagon represents a pentamer of CPs, and PSs are assumed to bind in the middle of these pentamers (see Figure 1.3 in Chapter 1) (Dykeman et al., 2013a). Note that formation of the pentamers from single CP proteins is not included in the model. Complete assembly in this model thus requires twelve CP units (pentamers) to bind twelve PSs and then complex together into a capsid. Most viral capsids have an icosahedral, i.e. a regular shape made up of 20 triangles, rather than dodecahedral capsid. Nevertheless this simple model follows *Picornavirales* assembly. These viruses have pseudo T=3 geometry (Lin et al., 1999; Tuthill et al., 2009) with pentamers around the 5-fold symmetry axes. Imagining these twelve pentamers as flat results in a dodecahedral symmetry. Still this model represents a simplification for the purpose of modelling the basic properties of viral assembly. Whilst it is appropriate for modelling *Picornavirales* assembly, it may not be directly translatable to other virus families.

To implement the Gillespie algorithm for the model of PS co-assembly a few specifics of the system have to be considered. While the association of CPs to PSs and potential dissociation are straight forward to model, the next step needs to consider that PSs are not free units but occur successively in the genomic RNA. The model works under the assumption that only CP bound to PS local to the growing complex can be incorporated. It is considered too unlikely that parts of the RNA distant on the sequence would be in close enough proximity in tertiary structure to interact (Dykeman et al., 2013a). Therefore, only CPs bound to neighbouring PSs are allowed to interact with each other and form a complex (Figure 7.3d–f). Correspondingly, a complex can only grow from the one or two free RNA ends and disassemble from the outer most CPs with respect to PS position on the RNA incorporated in the growing capsid. This ensures that the assembly process follows a Hamiltonian path, i.e. a path in a graph that visits every vertex exactly once (Figure 7.3b–c). Here this means that the RNA contacts each CP in the finished capsid exactly once without crossing (Hamilton, 1858).

In the original Gillespie algorithm the next reaction to implement is determined by looping through all possible reactions until the condition in Equation (7.19) is satisfied. However, the implementation used here employs a more dynamic programming kind of approach. Instead of looping through all possible reactions and different species of capsid intermediates, it keeps track of the actual configuration for each capsid/RNA and the possible reactions are determined on the fly. This reduces the number of reactions one has to sum over to obtain a_0 (see Equation (7.2)) and subsequently μ (see Equation (7.20)). It is achieved through an object-oriented programming approach, where each partially formed capsid is an object with a set of reactions. The program keeps track of the three states of the PSs on the RNA (not bound by CP, bound but not incorporated, incorporated in complex), the complex intermediate (which CP positions in the complex are occupied), and the PS affinities. This formalism allows for all possi-

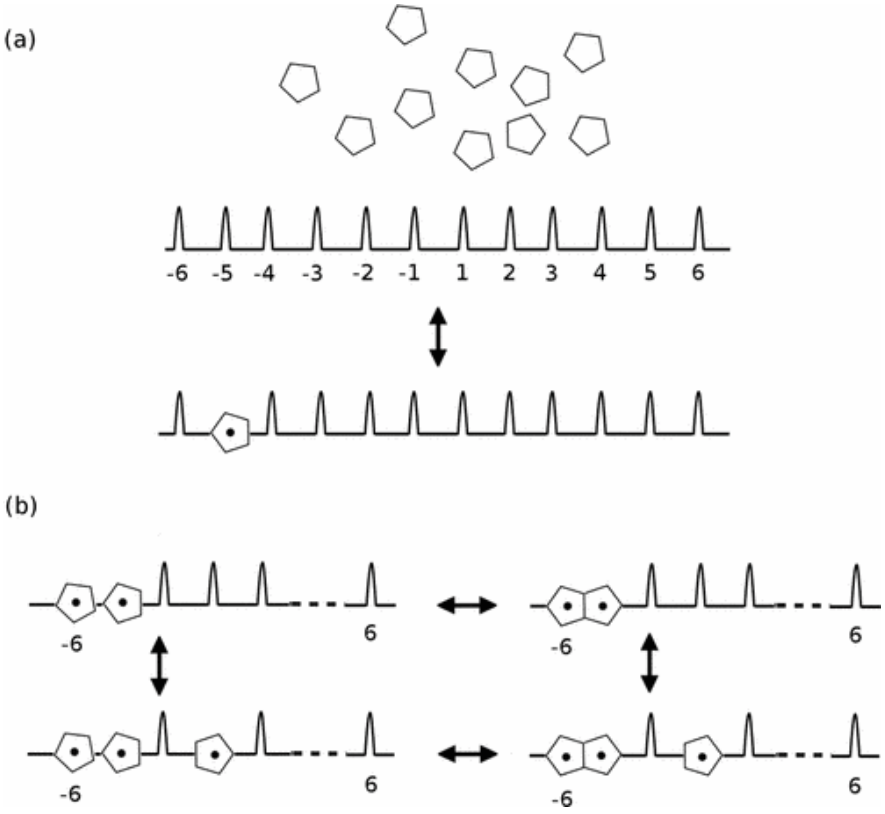

FIGURE 7.4: **Possible reactions in dodec model.** (a) First one CP unit needs to bind to a PS. This reaction is reversible. (b) Once another CP has bound to a neighbouring PS (top left), these can bind to each other nucleating capsid assembly by forming a two-CP complex (top right). Afterwards another CP can bind to a PS elsewhere on the RNA (bottom right). Alternatively, the other CP can bind to the RNA first (bottom left) and then the CPs on neighbouring PSs form a comp75lex (bottom right). The reverse is possible for each reaction. Reprinted figure with permission from Dykeman, E. C., Stockley, P. G., & Twarock, R., Physical Review E, 87(2):022717–1–12, 2013a, `http://dx.doi.org/10.1103/PhysRevE.87.022717`. Copyright 2019 by the American Physical Society, license number RNP/19/APR/013930.

ble reactions to be easily computed from the stored information (Dykeman et al., 2013a).

A set of parameters is required for the dodec and other viral assembly models. This includes the reactions probability rates, the inter-protein binding energies, and the PS affinities and distributions. As described by Gillespie (1977) there is a straightforward connection between the kinetic reaction rate k_μ and the probabilistic rate c_μ. For the simplest reactions k_μ can be calculated from c_μ simply by division through the volume V of the system given that it is spaciously homogenous (Gillespie, 1977). This is sufficient for the reactions modelled for this system and allows the use of experimentally derived kinetic rates (Dykeman et al., 2013a). c_μ is thus calculated as:

$$c_\mu = k_\mu / V \tag{7.22}$$

The on-off rates are calculated from the probabilistic rates for the forward divided by the backward reaction. Dykeman et al. (2013a) describe them as follows for their PS-mediated model:

$$\frac{c_{\mathrm{R}}^1(i)}{c_{\mathrm{R}}^2(i)} = \mathrm{e}^{-\beta \Delta G_{\mathrm{R}}(i)} \tag{7.23}$$

$$\frac{c_{\mathrm{P}}^1}{c_{\mathrm{P}}^2} = \mathrm{e}^{-\beta \Delta G_{\mathrm{P}}} \tag{7.24}$$

Here, the superscript numbers represent the forward (1) or backward (2) reaction, while the subscript letters R and P represent the PS-CP binding and the complex formation, respectively. Index i refers to the specific PS to account for differences in affinities between PS sites on the RNA. ΔG are the free binding energies of either PS to CP (R) or CPs to each other (P). The specific values for these parameters are taken from the literature. $c_{\mathrm{R}}^1(i)$ is set to 0.0024 s^{-1}, which is based on a diffusion kinetic rate of 10^6 M^{-1} s^{-1} and the approximate volume of a small bacterial cell, 0.7 $\mu\mathrm{m}^3$ (Endres and Zlotnick, 2002; Dykeman et al., 2013a). The lowest value for $\Delta G_{\mathrm{R}}(i)$, corresponding to the strongest binding, is based

on experimental evidence from TR in MS2 and thus -12 kcal/mol (Lago et al., 2001), variation towards higher binding energies, i.e. weaker affinities, is allowed and needs to be optimised (Dykeman et al., 2013a). ΔG_{P} is set to -2.5 kcal/mol, which is similar to the estimated value *in vivo*. c_{P}^{1} is variable and depends on the system to study.

A later modification to the model was the inclusion of a protein ramp. The idea is that *in vivo* CP does not suddenly appear at stoichiometric concentrations but is synthesised during infection: concentrations start out low and increase over time. The gradual increase in CP has two important effects in the model: (1) assembly efficiency almost doubles for a viral RNA with mixed PS affinities and (2) a higher degree of specificity for viral RNA (mixture of PS affinities) over cellular RNA (only low affinity "PSs") is observed. When competing with an excess of cellular RNA, the modelled viral RNA is almost exclusively packaged into capsids under ramp conditions, whereas when all protein is present at the start of the simulation more cellular RNA is packaged (Dykeman et al., 2014).

7.2.2 Modification of Dodec Model for MS2 Assembly

While the dodec model is easy to use and provides some insights into the mechanisms of PS-mediated assembly of viral capsids, it is still very simplified and can only be cautiously applied to the assembly process of specific viruses. When considering model virus MS2, it quickly becomes apparent that the dodec model is lacking some important features of this system, which may result in inaccurate conclusions. In Section 6.1.2 PS-mediated assembly in MS2 is described in detail. The relevant points are that MS2 assembles from 89 CP dimers and one maturation protein rather than twelve pentamers. Moreover, not all dimers are the same. There are 60 heterodimers (AB) and 29 homodimers (CC) (Figure 7.5A). In the lattice the maturation protein takes the position of one CC dimer. While A, B, and C are all the same protein, they differ by their conformations. Initially, all dimers are of the CC form. It is the binding to PSs that facilitates the

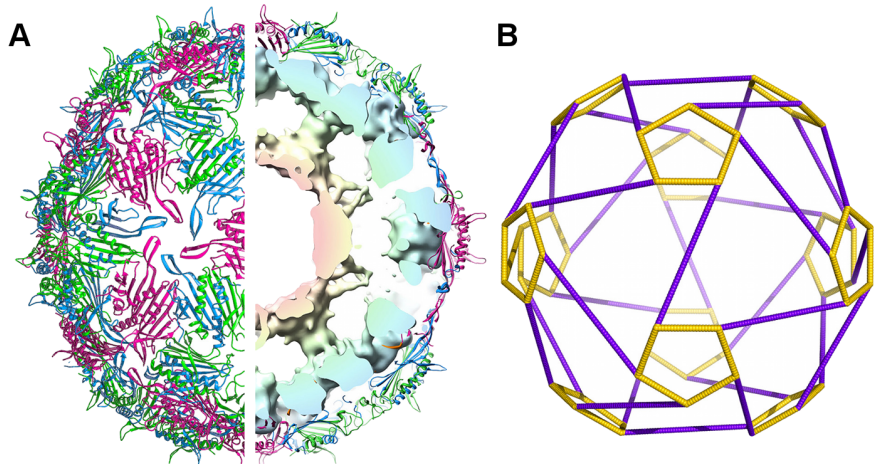

FIGURE 7.5: **MS2 capsid organisation and RNA contacts.** (A) The icosahedrally averaged crystal structure of the MS2 capsid at 2.8 $\mathring{A}$ resolution with CC dimers in pink and AB dimers in blue/green (left). AB dimers are situated around the 5-fold axes, while CC dimers are between them on the 2-fold axes. Maturation protein is not visible in averaged structures but would take the place of one CC dimer. The structure shows two layers of density within the capsid corresponding to the genomic RNA (right). The outer one (light blue) corresponds to contacts between the protein shell and the RNA. (B) The RNA within the MS2 capsid takes on the shape of this cage when averaged. Since RNA is expected to be in contact with each AB dimer but only cross under CC dimers, there are short distances between PSs around the 5-fold axes (yellow pentagons) and longer ones crossing the 2-fold axes (purple lines). Figure cropped from Fig. 1 in Geraets et al. (2015). Copyright under Creative Commons Attribution license.

conformational switch to AB dimers. There are thus estimated to be 60 PS-CP contacts in an MS2 capsid: one for each AB dimer. In crystal structures two layers of inner density were observed, which corroborates this idea (Figure 7.5A) (Geraets et al., 2015). The outer layer of an icosahedrally averaged structure is ordered in a regular cage form (Figure 7.5B). The RNA forms rings around the 5-fold axes, where the AB dimers are situated with connections between them (see also Figure 6.1C and D).

Due to the reasons mentioned above the assembly model requires a number of tweaks to be applicable to MS2 capsid assembly specifically.

The first difference between the dodec and MS2 models is that the number of PSs used in the MS2 model is 60 while this number is 12 in the dodec model. This is achieved by simply expanding the relevant arrays in the capsid class. The

more complicated change is the inclusion of 30 CPs that are not in complex with PSs. Note that maturation protein is modelled as simply another CC dimer here. Previously, any CP unit added to the growing capsid was first complexed with a PS. This means that an additional second-order reaction, which also depends on the concentration of CP, has to be added:

$$\mathrm{R}_7 : \mathrm{CP} \ + \ \mathrm{CPcomplex}_x \rightarrow \ \mathrm{CPcomplex}_{x+1} \tag{7.25}$$

$$\mathrm{R}_8 : \mathrm{CPcomplex}_{x+1} \ \rightarrow \ \mathrm{CP} \ + \ \mathrm{CPcomplex}_x \tag{7.26}$$

These reactions have their own rates based on the respective interdimer energies. Two different interdimer energies have to be considered now: AB:AB around the 5-fold symmetry axes and AB:CC across the 2-fold symmetry axes. In the original model AB:AB energies were set to -4.5 kcal/mol and AB:CC to -4.0 kcal/mol. With this in mind it becomes clear that not all potential moves along a Hamiltonian path are equal. In the dodec model there were up to five different directions in which the path could move from a pentagon. All these moves are theoretically equal because CP-CP binding energies are the same all over. Differences arise from the number of inter-pentamer bonds that can be formed from neighbouring CP units. This is not the case for MS2. There are up to three types of possible moves from any position: (1) clockwise around the 5-fold axis, (2) across the 2-fold axis, and (3) counter-clockwise around the 5-fold axis (Figure 7.6). While moves 1 and 3 are equivalent, move 2 requires that a CC dimer has been added first to avoid "holes" in the growing capsid and involves different interdimer energies. Therefore, when deciding whether an AB can be added to the capsid, the presence of a CC is checked where applicable.

7.2.3 Expansion of MS2 Model

While theoretically the MS2 model described in the previous section provides an appropriate model for this virus, it fails to produce assembled capsids to

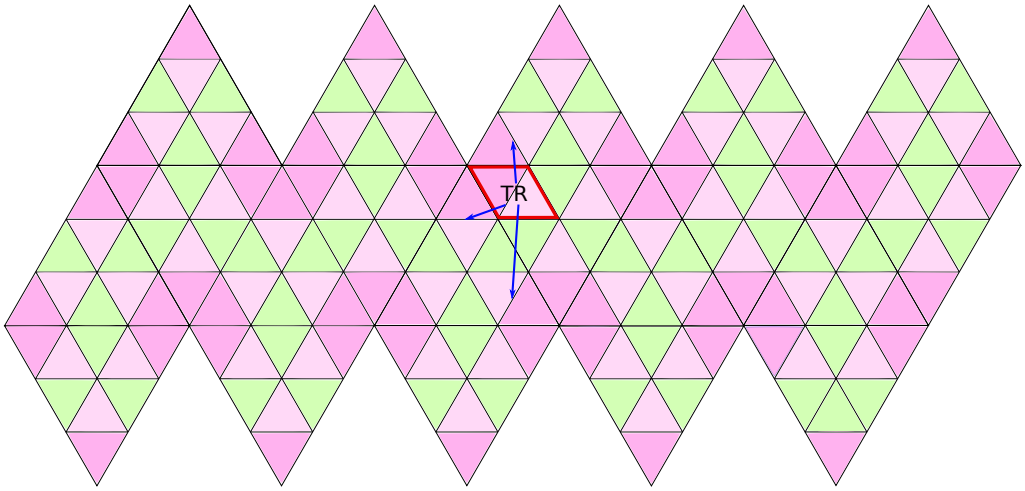

FIGURE 7.6: **Minimal nucleus of MS2 assembly model with TR.** MS2 capsid assembly is initiated at a high affinity PS, the model equivalent of TR. From there, there are three possible moves: The CP bound to a neighbouring PS can either be incorporated clockwise, counter-clockwise around the 5-fold axis of symmetry or across the 2-fold axis involving a unbound CC dimer in-between (arrows). The same moves are possible for each following addition with the exception of already incorporated CP. AB dimers are shown in pink and CC dimers are shown in green in the lattice.

a reasonable extent. Only a very small number of modelled RNAs get fully encapsidated. We therefore attempted to further improve on it by adding more features based on the finding from Chapter 6 regarding the nucleus and conserved PSs.

Like the dodec, the MS2 model assumes the capsid is built from two ends and the RNA follows a complete Hamiltonian path. Recent findings by Dai et al. (2017) suggest that the RNA path may in fact be "interrupted". The RNA can and does connect through the inside of the capsid and does not necessarily follow a complete Hamiltonian path along the protein shell. This insight becomes useful for modelling a larger nucleation complex. Moreover, the genomic RNA is also bound to maturation protein at both the 5' and 3' end (Shiba and Suzuki, 1981; Rumnieks and Tars, 2017) and therefore needs to be incorporated. In the complete capsid, maturation protein takes the place of one homodimer breaking the icosahedral symmetry (Dent et al., 2013), so it is probable that capsid assembly

starts from there. In the models described above, however, assembly is assumed to start from one or two CP:PS complexes. In MS2 assembly is thought to nucleate from TR as binding of CP to this PS also inhibits further viral genome replication (Hung et al., 1969; Ling et al., 1970; Beckett and Uhlenbeck, 1988; Beckett et al., 1988). In Chapter 6 I found that only a small number of PSs confirmed by Dai et al. (2017) are conserved between *Leviviridae* MS2, BZ13, and Qβ. Three of these are in the middle of the RNA and include TR and its two flanking PSs. Additionally, there is one close to the 5' and two close to the 3' end. In the structure by Dai et al. (2017) the middle and 3' contacts are all in close proximity to maturation protein (see Figure 6.7 in Chapter 6). This led to the hypothesis that MS2 and related viruses employ a larger nucleation complex of capsid assembly, which involves PSs at both ends and in the middle of the RNA as well as maturation protein. Modelling a large nucleation complex like this requires the capsid to be buildable from more than two ends. For simplicity, it was decided to model the 5' and 3' contacts from the maturation protein rather than the mapped positions of the conserved PSs. The conserved PS at the 3' end is close to the 3' maturation protein contact (MPC) allowing for only a few PSs between them. The RNA was therefore modelled to contact the maturation protein on the right from the 3' end. The 5' MPC is at genomic position 388–398 (Shiba and Suzuki, 1981) in MS2, resulting in PSs both upstream and downstream of it. Whilst the RNA positions of the MPCs are not mapped in Qβ, they are assumed to be similar. Since the lattice positions on the right of the maturation protein are occupied by other PSs, the 5' MPC was assumed to be on the left at the other 5-fold symmetry axis. The capsid is therefore grown from five ends: two at the 5' end, two in the middle from TR and neighbouring PSs and one from the 3' end. The proposed nucleation complex is shown in Figure 7.7.

Building from five ends posed the problem to ensure that the respective ends met to result in a full Hamiltonian path, e.g. the path starting at PS60 at the 3' end eventually becomes the path starting at the PS 3' of TR and vice versa. Once

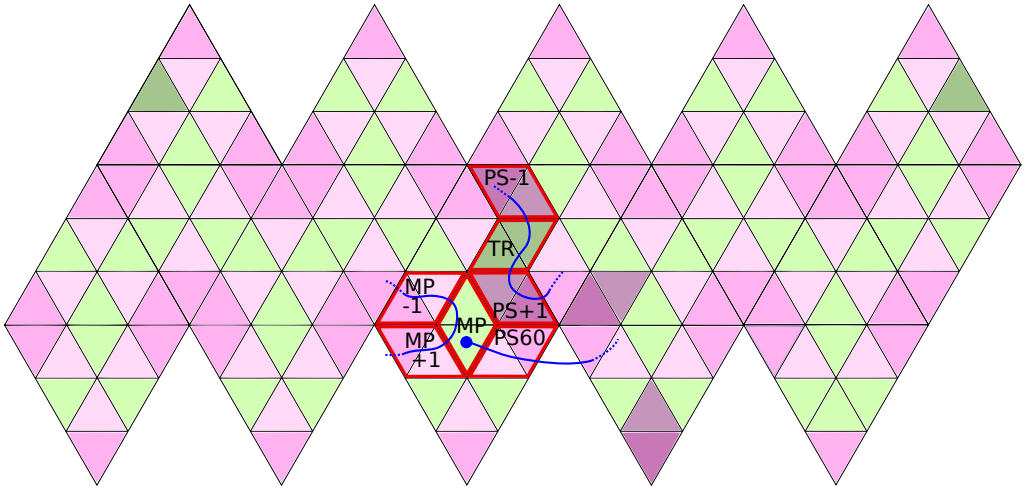

FIGURE 7.7: **Hypothesised large MS2 nucleus.** Nucleation of MS2 capsid assembly is hypothesised to involve six PSs and seven CP dimers. The nucleating CPs are marked in red and the RNA path is shown in blue. In addition to maturation protein, MP, the complex includes the neighbouring CPs as well as one across. This results in three pentamers being initiated. Contacts on one side of MP are facilitated through PSs directly up- and downstream of the 5' MP contact. On the other side one CP is bound to the most 3' PS, PS60, which is upstream of the 3' MP contact. The other CP neighbouring MP is bound to the PS just downstream of TR (PS+1). From there the RNA crosses the 2-fold axis incorporating two more CPs into the complex bound to TR and the PS just upstream of it (PS-1). Note that TR is bound to a CC dimer rather than an AB dimer, which is according to the position of this stem-loop (SL) in Dai et al. (2017).

the latter has reached for example PS46 and the former the neighbouring P47, the respective AB bound to them should be next to each other and connectable following the rules of the Hamiltonian path. In the initial test an additional parameter matrix was used, which specified the distance between two CP positions in the lattice in the minimum number of moves along a Hamiltonian path. Each step was then tested to ensure it would not result in the ends moving too far away from each other to meet in the end. Such a move was not permitted. This was called the "pull factor".

Previously, capsid assembly was allowed to nucleate from any neighbouring PSs. The location was mostly regulated by PS affinity. Since now several nucleation sites were implemented, nucleation location was instead forced by user-given PS positions for the 5' end and TR while the 3' end was always PS 60. Note that TR was observed under a CC dimer in Dai et al. (2017) rather than an AB dimer. The TR part of the nucleation complex therefore involved two ABs and one CC in a row and only two PS contacts. That means that for simplicity the contact with TR itself was not modelled.

Nucleation for this large nucleus is modelled in three steps: First the two PSs around TR make a small nucleus of two AB dimers bound to PSs and one CC dimer. The next step incorporates maturation protein and PS60. Finally, the two contacts at the 5' side are added on the other side of maturation protein. Each step can occur after the previous one given that a CP is bound to the respective PSs. The outermost CP can dissociate from any end as long as at least on remains to maintain all three parts of the original nucleus. This makes the nucleation reactions in this model irreversible, i.e. these nuclei were modelled to be unable to disassemble once formed.

Further restricting the way the capsid continues to build from the three-part nucleus can provide insight into the most favourable path along which the virus assembles. To force a certain continuation, the three nuclei were extended further to include the capsid positions that are least likely to result in a quick trapping

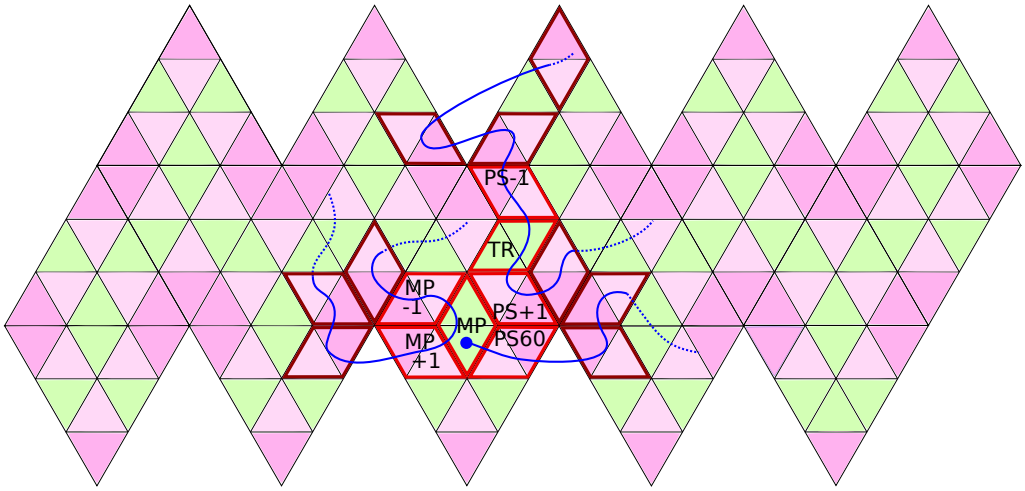

FIGURE 7.8: **Large extended MS2 nucleus.** The CPs of the small extended nucleation complex are marked in bright red and the added ones in the large extension in dark red. This ensures completion of the three neighbouring 5-folds. The RNA path is shown in blue with dotted lines indicating the likely continuation of each path. AB dimers are shown in pink and CC dimers and maturation protein (MP) in green.

of ends. For the 3' end this extension included the conserved PS. The positions of the extended nucleus in the lattice is shown in Figure 7.8. While *in vivo* the assembly paths would likely be regulated through PS affinities, in the model they were hard coded for simplicity while the affinities were modelled as uniform.

When testing the model, it was found that many times the paths ended up in a dead-end resulting in incomplete capsids. This was due to previously four dimers being added around the 5-fold axis before crossing the 2-fold axis, which left only one dimer to be added after which the path had no way of continuing. If this position was encountered at the first (PS1) or last (PS60) PS, it would not cause a problem. However, whenever a CP was added there that was not bound to PS1 or PS60 assembly would get stuck and could not complete. To avoid this pitfall, the final change to the model was to not allow for the path to cross the 2-fold axis if this resulted in four dimers around the 5-fold axis. Two, three or five dimers had to be incorporated before crossing. This was called the "no 4s" rule. The pseudocode for this test for one RNA end is shown in Appendix A

Algorithm A.14. When building from five ends, the same principle is applied five times to check that none of the ends results in four dimers around a 5-fold axis.

7.3 Discussion

The modelling of viral capsid assembly provides insights into many functional questions. Especially in the context of PS-mediated assembly, computational models have widened our understanding of many features of this process. Previously, these models were based on a highly simplified dodec model, which assumes twelve pentamers binding one PS each and assembling into a dodecahedral capsid. While much can be learned from this simple model it is not directly applicable to any virus. When wanting to study properties of a specific virus, an appropriate model needs to be designed. Here, I attempted to modify and improve upon a computational model for MS2 PS-mediated capsid assembly. These changes were based on novel insights from my own PS conservation analysis from Chapter 6 as well as PS and capsid imaging by Dai et al. (2017).

In Chapter 6 I found a small set of PSs conserved between MS2, GA, and Qβ. Due to their relative position in the asymmetric structure by Dai et al. (2017), some of these were hypothesised to be involved in assembly nucleation. Previously, this role has mostly been ascribed to one PS in MS2: TR. However, the close proximity of TR to maturation protein and other conserved PSs implies that it does not fulfil this role on its own. To test this idea, the MS2 model was modified to include three nucleation steps: First the AB dimers bound to the two PSs around TR and a CC dimer form a complex across a 2-fold axis as seen in the cryo-electron microscopy (cryo-EM) structure (Dai et al., 2017). Next, maturation protein is added as a CC together with a neighbouring AB bound to the 3' most PS, PS60. Finally, two more AB dimers bound to neighbouring PSs around the 5' end are recruited at the other side of maturation protein. At the end of all nucleation steps three pentamers have been started: the two left (5' contacts) and right (3' TR and 3' contact) of maturation protein and the one

above (5' TR). Once formed this structure could not disassemble to ensure the correct nucleation points throughout the run of the model, which are user given. This allows the testing of different nucleation PS positions along the RNA and their effects on efficient assembly.

The three part assembly nucleus required the implementation of building the capsid from five ends rather than two. Previously, when assembly began at one point in the RNA, the rules of the model allowed further build-up with CP bound to either PS directly neighbouring the ones already incorporated. With five nucleating RNA contact points, assembly can also continue from five ends. A distance matrix can be utilised to ensure that the ends that build towards each other such as 3' from TR and 5' direction form the 3' MPC, do not end up too far away from each other. Interestingly, Dai et al. (2017) found in their imaging work that not all the RNA is distributed directly in contact with the protein shell. Instead, the RNA sometimes dips into the centre of the capsid and emerges elsewhere. It may therefore not actually follow one complete Hamiltonian path as previously thought but a set of shorter, incomplete ones. The distance matrix may, thus, not be necessary.

Further insight into the continued path of assembly can be obtained through the extended nuclei models, in which either of the three nucleation steps is extended to include more AB positions, i.e. fixing of the continuation of the respective path. *In vivo* assembly is unlikely to randomly follow any path (Dykeman et al., 2014). Instead assembly follows a small set of paths or even just one path, which ensures efficient and complete assembly of most capsids. This can be regulated in a number of ways *in vivo* such as PS affinities and distances. Since these are difficult to model at this stage, extensions of the nuclei positions were implemented instead to test if this would result in increased numbers of completed capsids at the end of the simulation.

The final change was to prohibit a path from crossing from one 5-fold axis to another when this would result in exactly four dimers around the 5-fold axis.

When four dimers are present, it is impossible to complete the pentamer without the next path getting stuck. Even though the nature of the model allows for this to happen on a small number of occasions, too many such sets would quickly make completion of the capsid impossible. While no such rule directly exists *in vivo*, the virus likely employs a strategy to prevent such scenarios for the same reasons. Since moving along the 5-fold axis requires less spacing between neighbouring PSs than moving across the 3-fold axis, this may be achieved through the relative spacing of PSs along the RNA. If they mostly occur in sets of 2s, 3s or 5s in too close proximity to allow crossing, only two sets of 2s would result in four ABs around a 5-fold axis. Due to such distances not being included in the model, the rule had to be added to the model instead.

These four changes to the original MS2 capsid assembly model are meant to reflect the biology of the virus from what is known to date. While this updated model is still a simplification in many ways, it will hopefully improve upon the old model and increase the yield of completed capsids at the end of the simulation. The model with different sets of parameters will be tested in Chapter 8.

Chapter 8

Modelling the Effects of Nucleation on the Assembly of MS2

The goal in this chapter was to learn more about *Leviviridae* capsid assembly by improving upon the MS2 assembly model. The model and its background are described in detail in Chapter 7. At first the original model with provided parameters was tested for performance and assessed. This was used as a basis for trying out different parameters and other changes to achieve a higher number of completed capsids at the end of the simulation.

8.1 Initial Model Settings and Performance

The original model, which formed the basis of the work in this chapter, assumed nucleation of assembly at a single point between CPs bound to two neighbouring PSs. Two different nucleation options were tested at the start: Either the two CPs would assume two positions next to each other around the 5-fold axis of symmetry (ABAB nucleus) or they would be on different pentamers with a CC dimer between them (ABCCAB nucleus). From there the capsid was built up in

5' and 3' direction along the RNA. The points of nucleation on the RNA were directed through PS affinities, which had been evolved to maximise the number of assembled capsids at the end of the simulation. The affinities were converted to free energy of complex formation using the same formula as in Chapter 2 Section 2.4.1.5 Equation (2.7) and, thus, ranged from -4 to -12 kcal/mol. The number of RNA molecules was set to 2000 and the number of CP dimers to 18,000. This ensured that just enough CP was present to package all RNA. The formation of dimers was assumed to be instantaneous as observed experimentally and was not modelled. The reaction volume was $1\mu m^3$. Since CPs are present in two different states, AB or CC, their interactions with each other are modelled as 1st or 2nd order reactions for AB:AB or AB:CC interactions, respectively. The rates for these 1st and 2nd order reactions were set to 10^6 s^{-1} and $M^{-1}s^{-1}$, respectively, and represent kinetic diffusion rates as detailed in Section 7.2.1. The temperature of the system was given as 298.15 K, which corresponds to room temperature at 25°C. The simulation was modelled for 1001 seconds. The interdimer interaction energies were set to -4.5 and -4.0 kcal/mol for AB:AB and AB:CC interactions, respectively.

Neither of the nucleation options resulted in complete capsids by the end of the simulation. For nucleation on one 5-fold axis the highest number of incorporated CPs was 89 and for nucleation across two 5-fold axes 87, which means at least one and three CPs were missing in the capsids, respectively. Histograms of the number of RNAs with a certain number of CPs incorporated revealed distinct peaks in either nucleation condition (Figure 8.1). Interestingly, despite small differences in the height of the peaks they occurred at the same capsid intermediate sizes for both conditions. This strongly indicates that the reason no capsids are completing is that assembly gets trapped in intermediates.

To test the hypothesis that assembly gets trapped in these models the Hamiltonian path intermediates were outputted and analysed. By checking whether any neighbouring CP position was free to move to from either end, the capsid

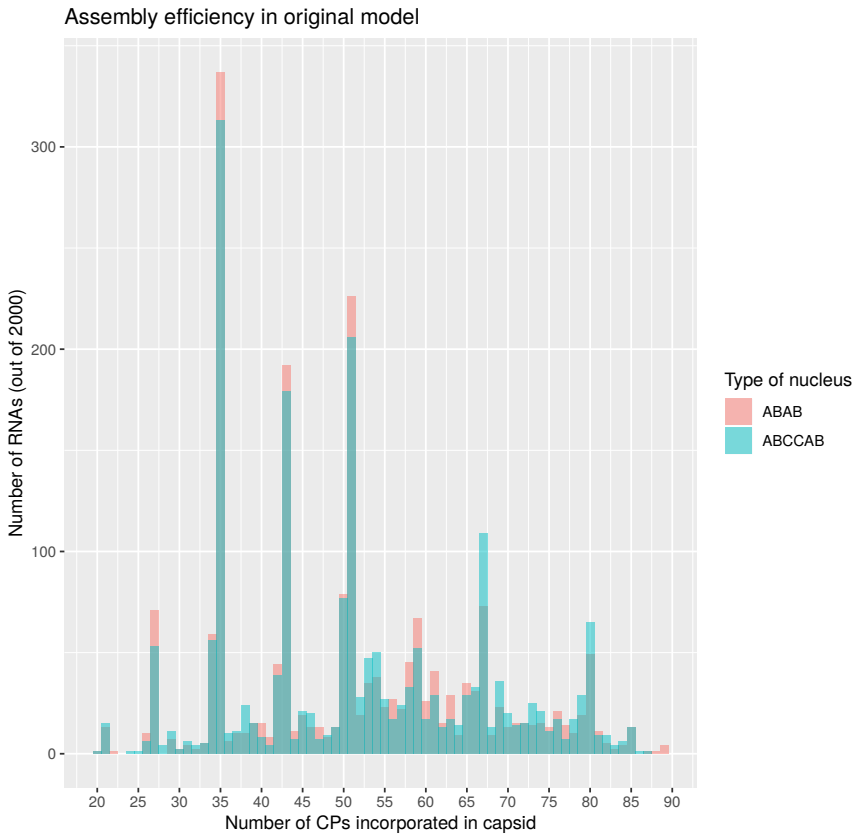

FIGURE 8.1: **Incorporated CPs per RNA in minimal nucleus model using an ABAB or ABCCAB nucleus.**

intermediates at the end of a simulation were split into dead-ends, i.e. no possible moves from either end, and continuable. Note that the model includes dissociation of CP; however, this can be made more difficult when the CP at the end of the trapped partial path is surrounded by other CPs in the growing capsid, which each stabilise it through interdimer binding energies. When only ABAB nuclei were allowed 1430 RNAs were stuck in assembly dead-ends versus 570 RNAs that could continue the assembly process. These numbers were similar for the ABCCAB nucleus with 1374 dead-ends versus 626 continuable. The length of the dead-end paths largely corresponded to the peaks observed in the histograms showing that assembly mostly gets trapped in a small number of intermediate sizes.

8.2 Varying of Important Parameters

With no fully assembled capsids at the end of the simulation, the original model failed to sufficiently recreate biology. In order to improve the assembly efficiency of the model a number of parameters were varied including the interdimer energies, PS affinities, and nucleating PS positions. This involved running the model many times with different combinations of parameters. To maximise efficiency at the testing state, the simulations were run in parallel using GNU parallel (Tange, 2011). All tests were performed with the "no 4s" rule, which disallows formation of 5-fold partial capsids, since no complete Hamiltonian path is possible with four CP (dimers) in a 5-fold.

8.2.1 Interdimer Energies

There are two types of capsid protein dimers in the MS2 capsid: the homodimer CC and the heterodimer AB. AB dimers interact with each other, AB:AB, and with CC dimers, AB:CC. The binding energies for these two interactions are not the same and may influence efficiency of assembly (ElSawy et al., 2010). In order to identify the optimal settings for AB:AB and AB:CC interdimer energies, the assembly model was run on a set of different energy combinations summarised in Table 8.1. They ranged from the lowest combination, -4.5 kcal/mol and -4.0 kcal/mol, as used originally, to the highest at -0.5 kcal/mol and -1.0 kcal/mol for AB:AB and AB:CC interactions, respectively. To remove other variables from this test, the affinities for all PSs were set to -12 kcal/mol, which corresponds to TR. The PS positions for assembly nucleation were then set to 29 and 30 out of 60, which would place them in the middle of the RNA. This is similar to the positioning of TR and its neighbouring PSs. To follow the positioning of these PSs in the Hong structure, a ABCCAB form of nucleation was also enforced.

Keeping the AB:CC interaction energies constant, the AB:AB energies were varied and compared. The first set of AB:CC energies tested was -4.0 kcal/mol

TABLE 8.1: Interdimer energy combinations. The energies are given in kcal/mol and tested combinations are marked "x".

AB:AB \ AB:CC	-1.0	-2.0	-3.0	-4.0
-0.5	x	x	x	x
-1.0	x	x	x	x
-1.5	x	x	x	x
-2.0		x	x	x
-2.5		x	x	x
-3.0			x	x
-3.5			x	x
-4.0				x
-4.5				x

as used in the original model. As seen in Figure 8.2, if combined with AB:AB energies of -4.5 kcal/mol, the peaks in the histogram are visible as before (see Figure 8.1). This indicates that the trapped intermediates were not the primarily the result of the growing capsid adding exactly four CP at a 5-fold axis before crossing to the next. Instead, trapping appeared to be facilitated by low AB:AB interaction energies. The peaks were also visible for AB:AB up until -1.5 kcal/mol (moss green). For the higher energies -1.0 and -0.5 kcal/mol, the histogram flattened and the distribution generally shifted to the right towards a higher degree of completion. Complete capsids were observed for AB:AB -1.5 kcal/mol (2), -1.0 kcal/mol (4), and -0.5 kcal/mol (3).

The next set of energy combinations that was tested was AB:CC of -3.0 kcal/mol with AB:AB ranging from -3.5 to -0.5 kcal/mol. Similar to the previous test, defined peaks were observed for higher energy combinations (Figure 8.3). However, they already disappeared when AB:AB was as low as -2.0 kcal/mol (turquoise). From there on the distributions were unimodal and shifted further to the right with higher AB:AB energies. Successful assembly, however, occurred only for AB:AB -1.0 kcal/mol (1) and -0.5 kcal/mol (10).

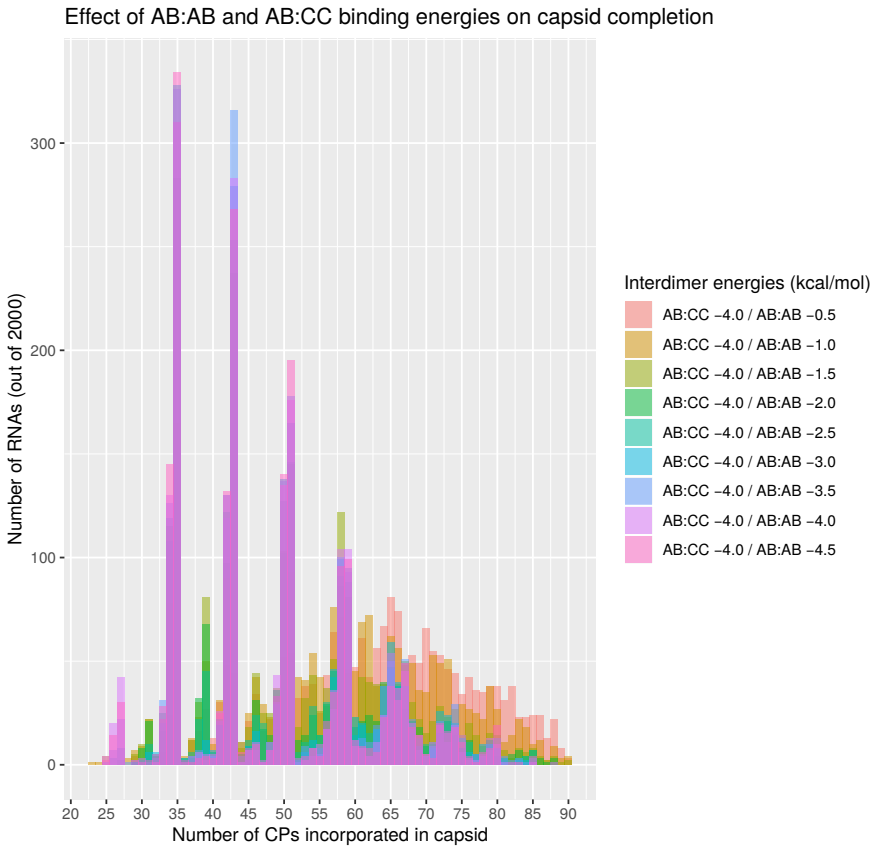

FIGURE 8.2: **Incorporated CPs per RNA for AB:CC energy -4.0 kcal/mol with minimal nucleus.**

The highest number of fully assembled capsids was thus far observed for AB:CC -3.0 kcal/mol and AB:AB -0.5 kcal/mol at 10 capsids. This indicates that in addition to AB:AB, AB:CC energies also play a role in assembly success with higher energies performing better. To test this idea, the AB:CC energies were increase further to -2.0 kcal/mol and tested with AB:AB energies ranging from -2.5 to -0.5 kcal/mol. Here the separation of combinations with trapped intermediates and those without became even more pronounced (Figure 8.4). Only AB:AB -2.5 kcal/mol and -2.0 kcal/mol still showed peaks in the histogram (pink and blue). From -1.5 kcal/mol onwards (turquoise) the distributions were unimodal and, as for AB:CC -3.0 kcal/mol, shifted increasingly right the higher the AB:AB energies were. These were also the combinations for which completed capsids were observed. There were 1, 3, and 15 completed capsids with AB:AB -1.5 kcal/mol, -1.0 kcal/mol, and -0.5 kcal/mol, respectively. Increasing

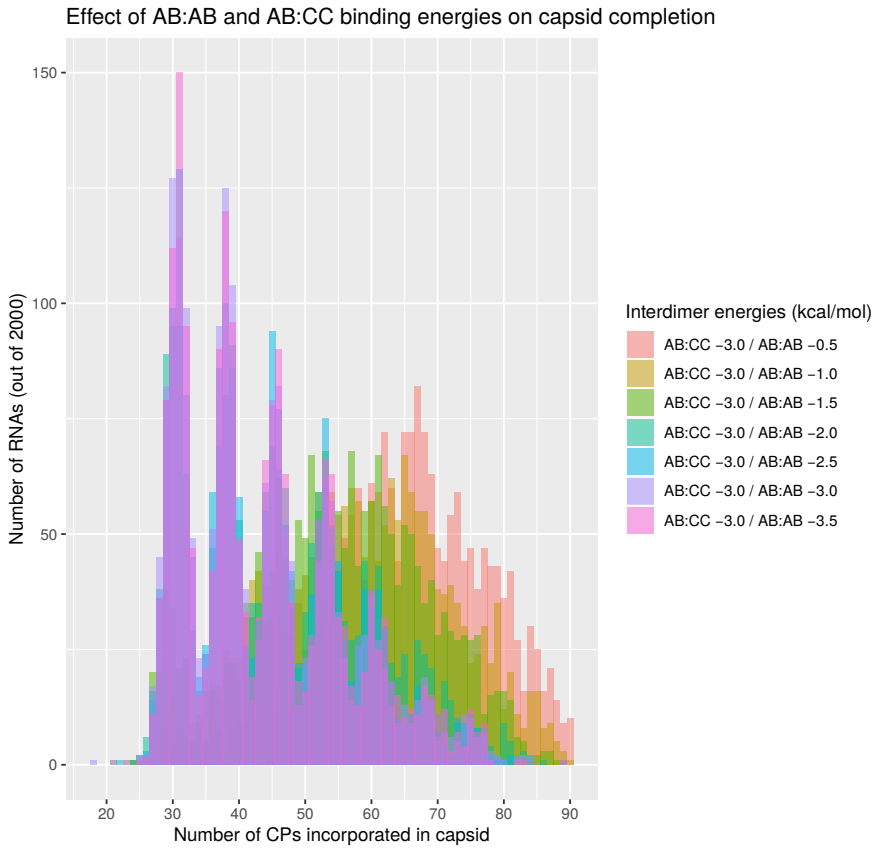

FIGURE 8.3: **Incorporated CPs per RNA for AB:CC energy -3.0 kcal/mol with minimal nucleus.**

the AB:CC energies had an unexpected side effect. Whilst there were only 15 and 3 complete capsids in the -0.5 and -1.0 kcal/mol AB:AB conditions, 36 and 8 capsid intermediates had all PSs bound by CP and incorporated, respectively. This means that when the AB:CC interactions get too weak, more CC dimers dissociate from the growing capsid resulting in incomplete assembly.

Using an AB:CC energy of -2.0 kcal/mol combined with AB:AB of -0.5 kcal/mol gave the highest number of completed capsids despite CC dimer dissociation. Thus, the performance when using AB:CC of -1.0 kcal/mol with AB:AB of either -1.5 kcal/mol or -1.0 kcal/mol (Figure 8.5) was tested. The condition -0.5 kcal/mol was not included due to the simulation taking an excessive amount of time to complete. Also here the lower energy condition displayed peaks of trapped intermediates. Interestingly, while AB:AB of -1.0 kcal/mol did not have these peaks and showed a unimodal distribution it was the worst performing condition of all

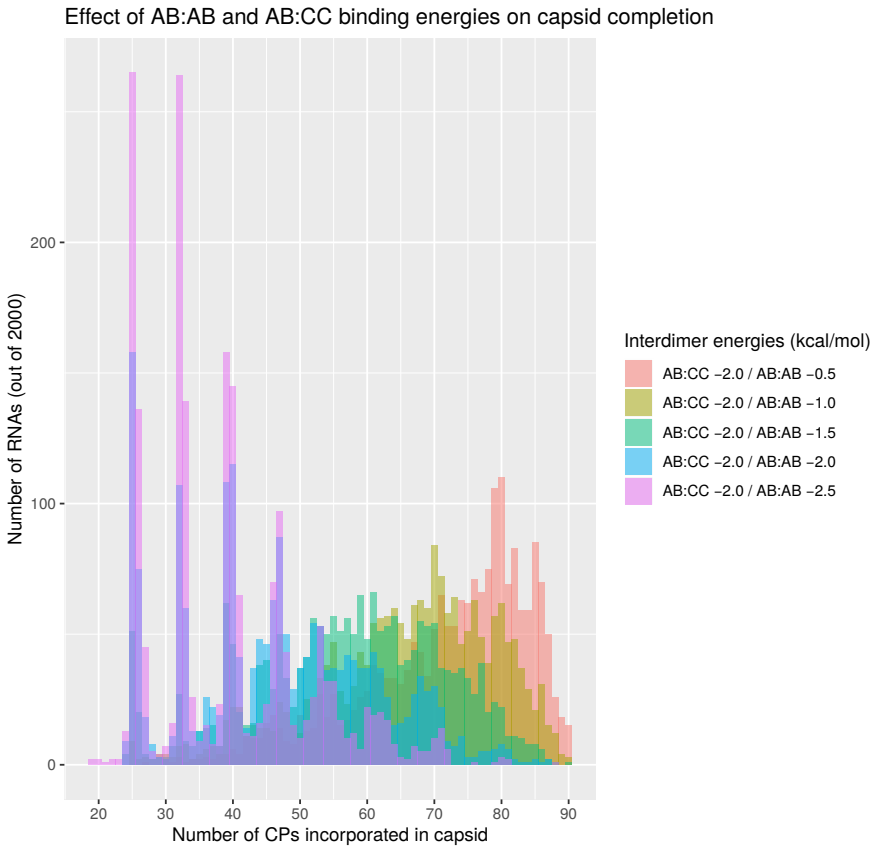

FIGURE 8.4: **Incorporated CPs per RNA for AB:CC energy -2.0 kcal/mol with minimal nucleus.**

tested with regards to the highest degree of assembly reached. At most 75 CP dimers were incorporated, while this number was at least in the 80s for any other energy combination. This is most likely due to CC dimers not staying associated to the capsid intermediates. Similar to the -2.0 kcal/mol AB:CC conditions, also here there were more capsid intermediates with all PS-bound CP dimers incorporated. With -1.0 kcal/mol there were 27 and with -1.5 kcal/mol AB:AB energies 4 such capsid intermediates.

Taken together these data show that the interdimer energies play a vital role in assembly efficiency. Despite the "no 4s" rule and fixing the point of nucleation, not much change was observed to the original model when interdimer energies of -4.0 and -4.5 kcal/mol were used for AB:CC and AB:AB, respectively. Only when the energies were increased, especially for AB:AB, more capsids were complete at the end of the simulation. However, increasing the interaction energies came

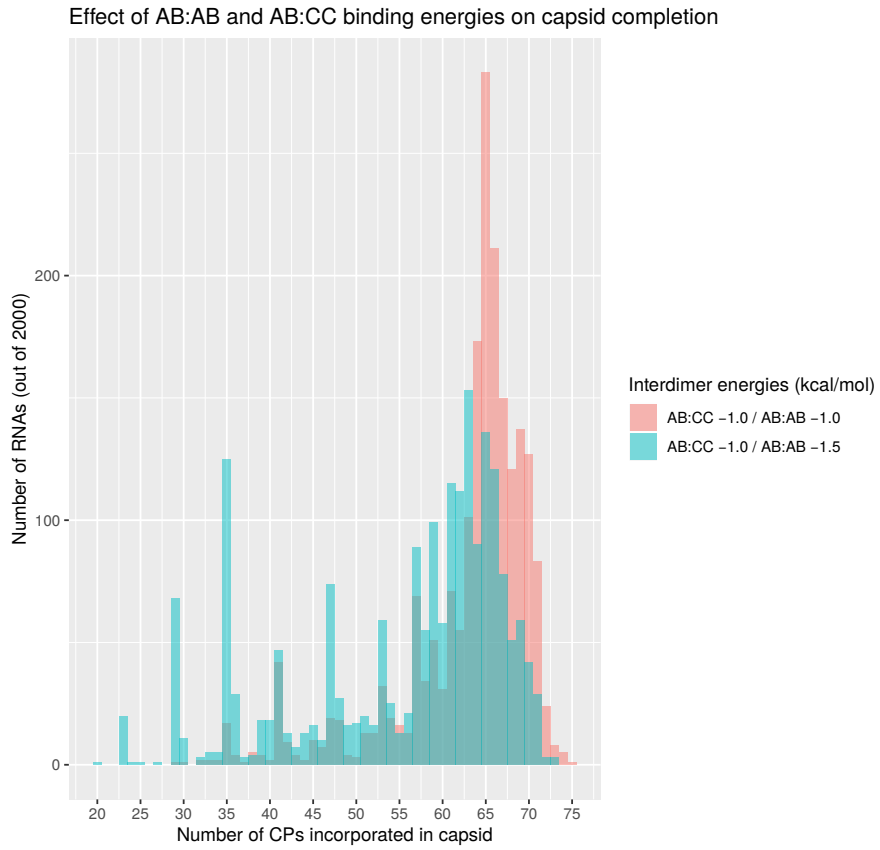

FIGURE 8.5: **Incorporated CPs per RNA for AB:CC energy -1.0 kcal/mol with minimal nucleus.**

with a trade-off of simulation running time. While the original model finished in less than an hour, the run time increased as the energies increased with the combination of -1.0 and -0.5 kcal/mol (not shown) running for several weeks without finishing. This is due to the large number of fast CP on/off reactions, which occur in the model when the interactions between CPs are less stable. Modelling more fast reactions results in modelling time incrementing more slowly and thus taking more real time to reach the simulation time threshold of 1001 seconds. It also results in semi-complete capsids, which have all AB dimers incorporated but some CC dimers are missing. The condition that resulted in the highest number of completed capsids, -2.0 and -0.5 kcal/mol, ran for approximately one day. Even under these conditions only 15 capsids were fully assembled, which represents merely 0.75% of modelled RNAs being encapsidated, a number that is too low to be viable *in vivo*. Therefore, other parameters were also tested.

8.2.2 Effect of Nucleus Size on Assembly

The nucleation of assembly in MS2 is thought to occur at TR. In the original model it was therefore implemented that assembly started with two CPs bound to neighbouring PSs. When Dai et al. (2017) resolved an asymmetric cryo-EM structure of MS2, they were able to fit 15 SLs into the inner density representing RNA. These are RNA structures that are in contact with the capsid so are thought to be PSs. This set of SLs is referred to as "Hong PSs" here after the group leader from that paper. These SLs were shown in Figure 6.4 in Chapter 6. PS conservation analysis performed in Chapter 6 identified a small set of Hong PSs that was conserved between MS2, Qβ and BZ13 phages (see Table 6.4 and Figures 6.5 and 6.6). Interestingly, a subset of these PSs were in close proximity to each other and the maturation protein in the cryo-EM structure (Dai et al., 2017) (see Figure 6.7 in Chapter 6). This led to the hypothesis that nucleation of assembly in these coliphages actually involved more PSs and protein than previously thought. The hypothesis was put to the test in the updated MS2 assembly model. As described in detail in Section 7.2.3, the nucleation was modelled in three steps: First CPs bound to TR and its two neighbouring PSs would be incorporated as two AB dimers and one CC dimer across a 3-fold axis. Next maturation protein, modelled as a CC dimer, and a CP bound to the last PS were added. Finally, two CPs bound to two neighbouring PSs upstream of TR close to the 5' end were incorporated next to maturation protein. To ensure that all nucleus parts were preserved later in the model not all CP could dissociate from either part. The MPCs were fully fixed, whilst the TR nucleus could dissociate and re-associate at a different position given that CPs that originated from that nucleus stayed associated. Once the nucleus had been formed assembly continued from each of these places in 5' and 3' direction except for the 3' end, which could only build up in 5' direction.

The interdimer energy test was repeated with the large nucleus. At first a pull factor was applied to facilitate the meeting of assembly ends. This was done

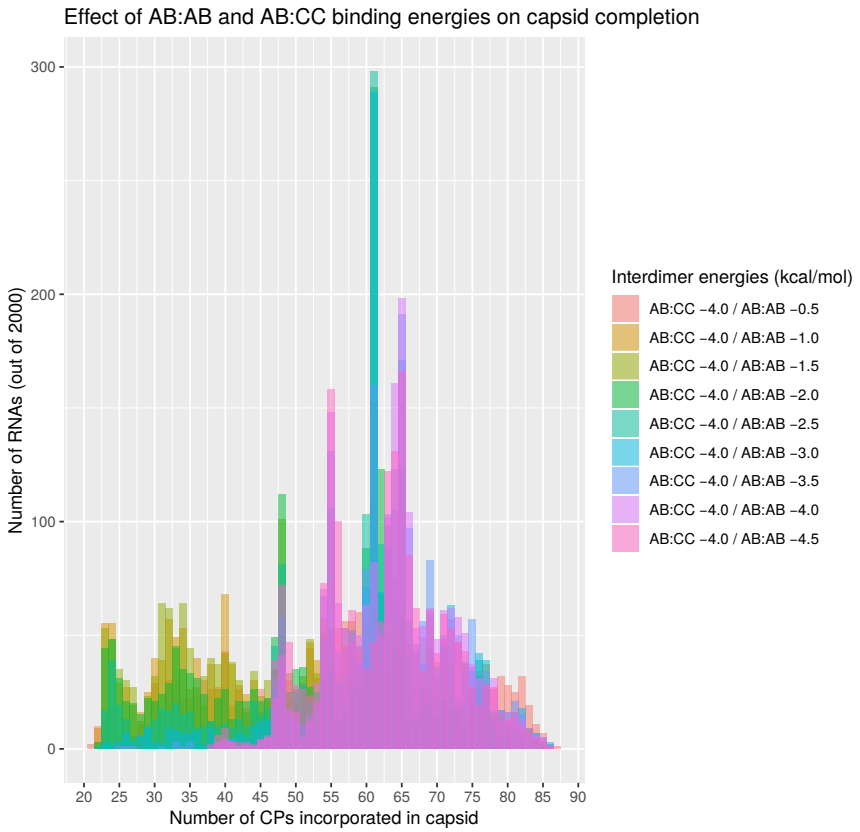

FIGURE 8.6: **Incorporated CPs per RNA for AB:CC energy -4.0 kcal/mol using the large nucleus.**

by calculating the minimum number of steps following a Hamiltonian path that would connect the respective ends. If adding the CP at a certain position would take the ends more steps away than there were PSs left between them, then the step was not allowed. The idea was that *in vivo* the RNA would be restricted and the ends would have to be next to each other once all PSs had been bound and the CPs incorporated. For this test the PS positions for the 5' MPC were set to 7 and 8 while the TR neighbouring contacts were set to 29 and 30 out of 60 total PSs. This was based on the positions of these SLs in Dykeman et al. (2013b).

Starting with AB:CC energies of -4.0 kcal/mol a very different histogram was observed compared to the minimal nucleus (Figure 8.6). While peaks of trapped intermediates were still observed for lower AB:AB energies, these were at higher CP numbers and were not as uniform between energy conditions. Interestingly,

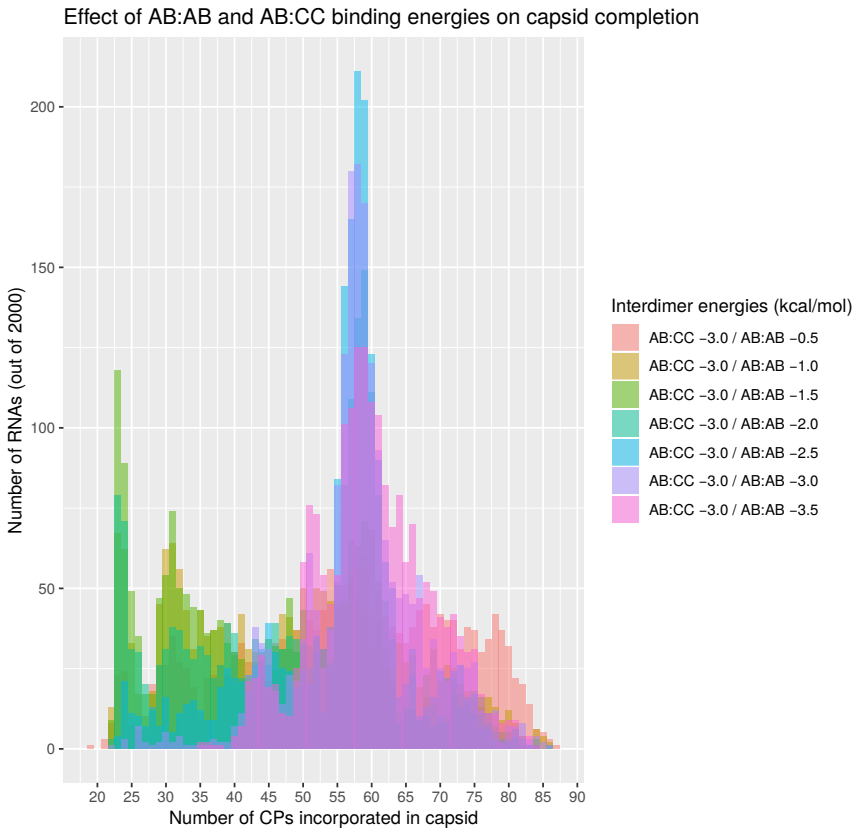

FIGURE 8.7: **Incorporated CPs per RNA for AB:CC energy -3.0 kcal/mol using the large nucleus.**

the middle energy values (shades of green) resulted in peaks at lower CP numbers, a general shift to the left. The highest energy caused a shift to the right again. However, no complete or semi-complete capsids were observed in any of the conditions.

Continuing with AB:CC energies of -3.0 kcal/mol the range of AB:AB energies was tested again (Figure 8.7). As opposed to the minimal nucleus condition, the lower AB:AB values resulted in an almost unimodal distribution with only a single main peak around 60 CP out of 90. At the middle values trapping occurred at lower CP numbers as observed for AB:CC -4.0 kcal/mol, while the highest AB:AB values resulted in a shift towards the right again. Also here there were no completed or semi-complete capsids.

A similar trend in the distribution of capsid intermediates was observed with AB:CC energies of -2.0 kcal/mol (Figure 8.8). More pronounced than in the

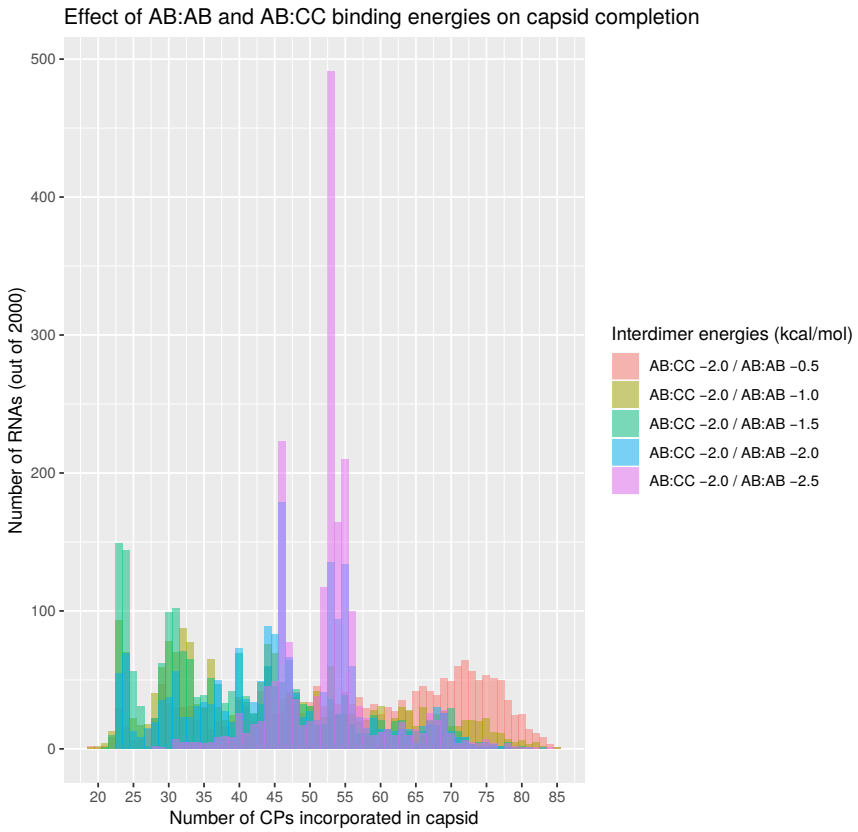

FIGURE 8.8: **Incorporated CPs per RNA for AB:CC energy -2.0 kcal/mol using the large nucleus.**

previous conditions, two high peaks were seen for AB:AB -2.5 and -2.0 kcal/mol around 45 and 55 CP dimers. Then there was a shift to the left for AB:AB of -1.5 and -1.0 kcal/mol with the highest peaks around 25 and 30 CP dimers. Finally at the highest energy -0.5 kcal/mol, the distribution appeared unimodal with no pronounced peaks and was shifted to the right. The highest number of CP in an intermediate were 85. For AB:AB -1.0 kcal/mol a single semi-complete capsid was observed.

Using an AB:CC energy of -1.0 kcal/mol performed expectedly badly with no intermediates with 75 or more incorporated CPs observed (Figure 8.9). Here both AB:AB energies used resulted in bimodal distributions with a number of peaks of trapped intermediates and one semi-complete capsid each.

The energy tests showed that using the large nucleus worsened the model performance. Under no conditions was even a single capsid fully assembled. The

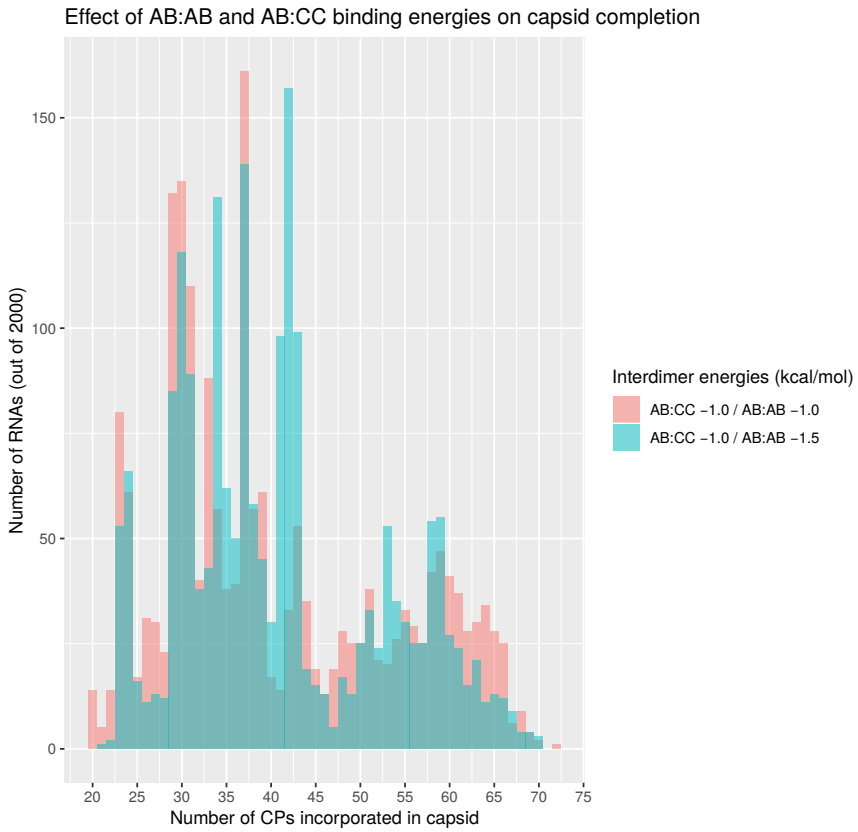

FIGURE 8.9: **Incorporated CPs per RNA for AB:CC energy -1.0 kcal/mol using the large nucleus.**

patters of intermediates as seen in the histograms changed drastically compared to the minimal nucleus. While lower AB:AB energies showed a few peaks around 50–60 CPs, medium energy values resulted in a left-ward shifted distribution, and only the highest values resulted in the right-ward shifted distributions as observed for the minimal nucleus. The use of several points of assembly and the pull between them appeared to lead to more trapping of intermediates as the paths could not connect and got stuck. Therefore, the large nucleus was tested without use of the pull factor. This is based on the observations in Dai et al. (2017) that the RNA does not always follow a path along the inside of the capsid but sometimes dips into the centre in one place and resurfaces at another. The path does therefore not have to be continuous.

When building from five ends without the need for these to meet, model performance was strongly improved. For no energy combinations were peaks of

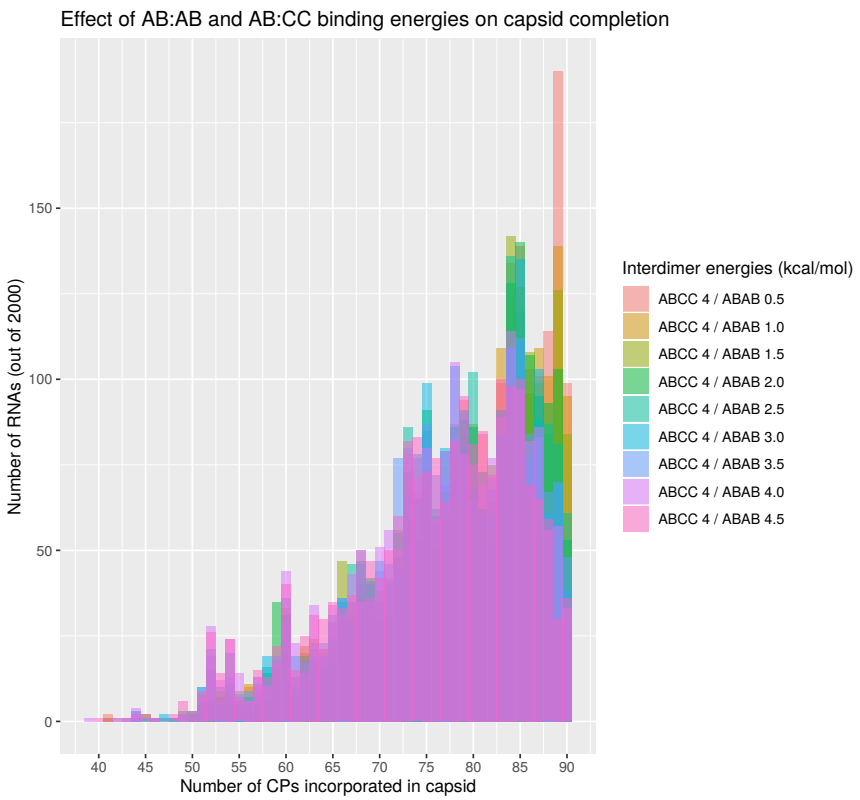

FIGURE 8.10: **Incorporated CPs per RNA for AB:CC energy -4.0 kcal/mol using the large nucleus without pull factor.**

trapped intermediates observed (Figures 8.10–8.13). Already for the lowest energy combination of -4.0 and -4.5 kcal/mol, which previously resulted in many trapped intermediates, 36 capsids were fully assembled. This number increased when AB:AB energies increased and were highest for -4.0 and -0.5 kcal/mol at 99 (Figure 8.10 and Figure 8.14 (red)). Increasing AB:CC energies improved the outcome further. The lowest number of assembled capsids with AB:CC -3.0 kcal/mol was seen with AB:AB -3.5 kcal/mol at 63. Also this increased further with increasing AB:AB energies and was highest with -0.5 kcal/mol at 225 assembled capsids at the end of the simulation (Figure 8.11 and Figure 8.14 (green)). This was the best performing combination of energies of all tested. When AB:CC was increased to -2.0 kcal/mol at most 112 capsids assembled when AB:AB of -1.0 kcal/mol was used (Figure 8.12 and Figure 8.14 (blue)). Interestingly, also here there was a large discrepancy between the number of fully assembled capsids

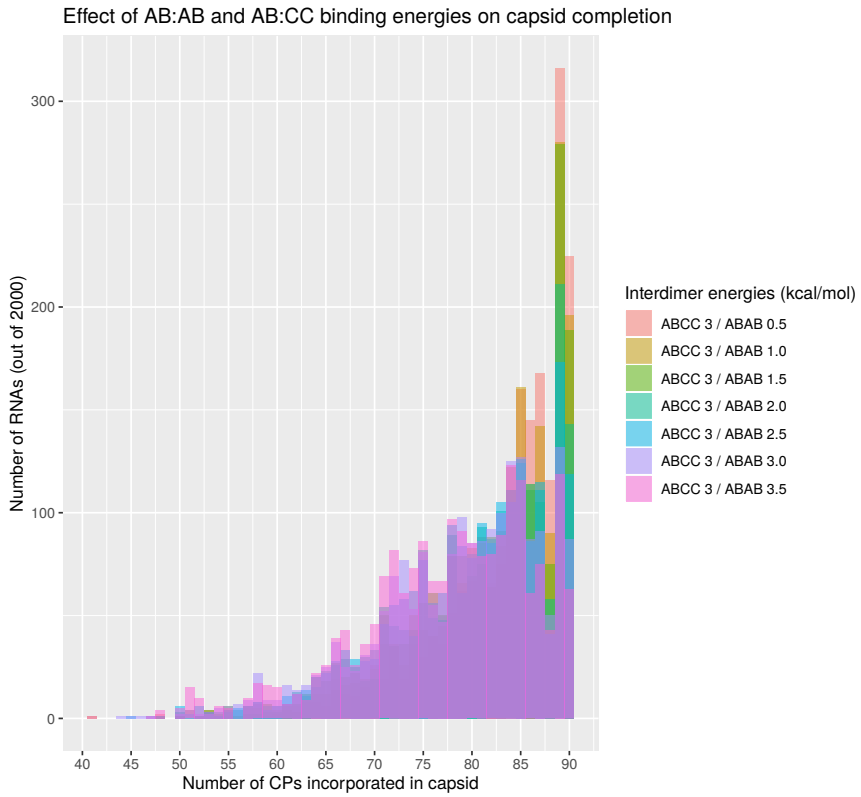

FIGURE 8.11: **Incorporated CPs per RNA for AB:CC energy -3.0 kcal/mol using the large nucleus without pull factor.**

and semi assembled for this AB:CC energy (Figure 8.15). When AB:AB energy of -0.5 kcal/mol was used almost half of the capsid intermediates were semi assembled whereas only 10% of those also contained all CC dimers and were fully assembled. This showed that an AB:CC energy of -2.0 kcal/mol was the best for building up capsids from AB dimers but too high to stably incorporate CC dimers as well. Again, a further increase all the way to -1.0 kcal/mol for AB:CC resulted in worst model performance (Figure 8.13 and Figure 8.14 (purple)). Under these conditions most capsids appeared to never assemble further than the nucleus with a sharp peak at seven CP dimers. The numbers of semi- and complete capsids for all energy conditions are listed in Appendix A Table A.9.

The conserved PS at the 3' end was observed two steps away from the 3' MPC site used for nucleation here and assembly was not forced to proceed that way. The TR part of the nucleus could dissociate later during the simulation if more

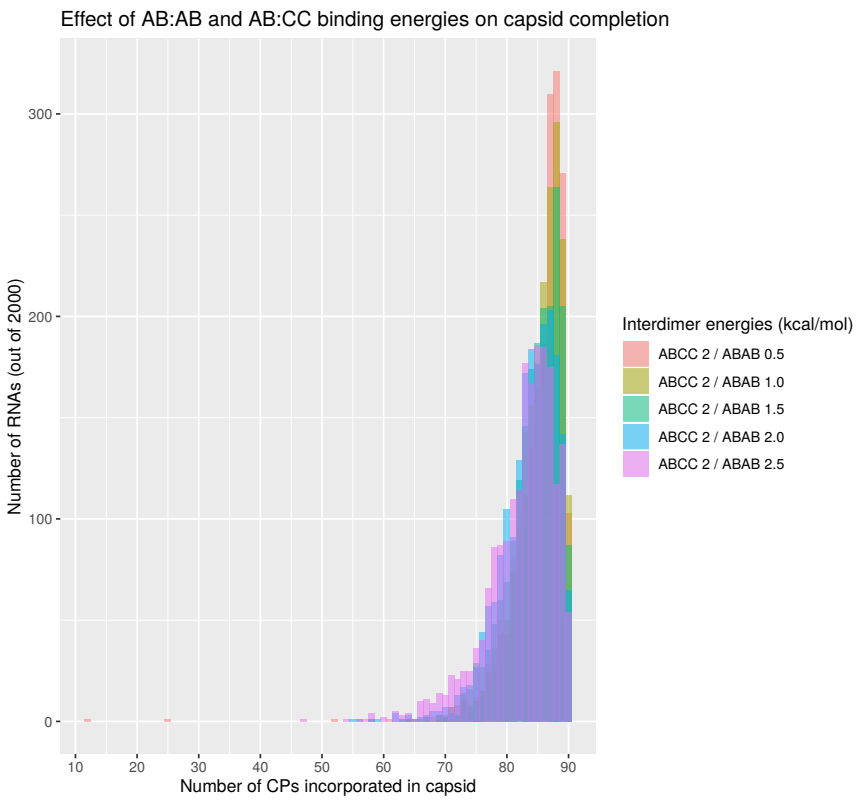

FIGURE 8.12: **Incorporated CPs per RNA for AB:CC energy -2.0 kcal/mol using the large nucleus without pull factor.**

capsid was built up on one side and those CPs were the outermost ones. It was therefore of interest to see, whether the conserved sites would be reached and preserved at the end of the simulation (see Table A.9 in Appendix A). For lower AB:CC energy conditions (-3 and -4 kcal/mol) the vast majority of complete capsids had preserved the TR nucleus part. This was not the case for AB:CC of -2 kcal/mol. The same was true for the conserved PS at the 3' end. Even when not forced, most complete capsids had assembled following that partial path in the AB:CC of -3 or -4 kcal/mol but not in the higher energy conditions.

8.2.3 Position of Nucleating Packing Signals

Thus far the positions for the nucleating PSs had been assumed to be 7, 8, 29, 30, and 60. However, these were just estimates based on the SL numbering in Dykeman et al. (2013b). Since the choice of nucleating PS position may affect

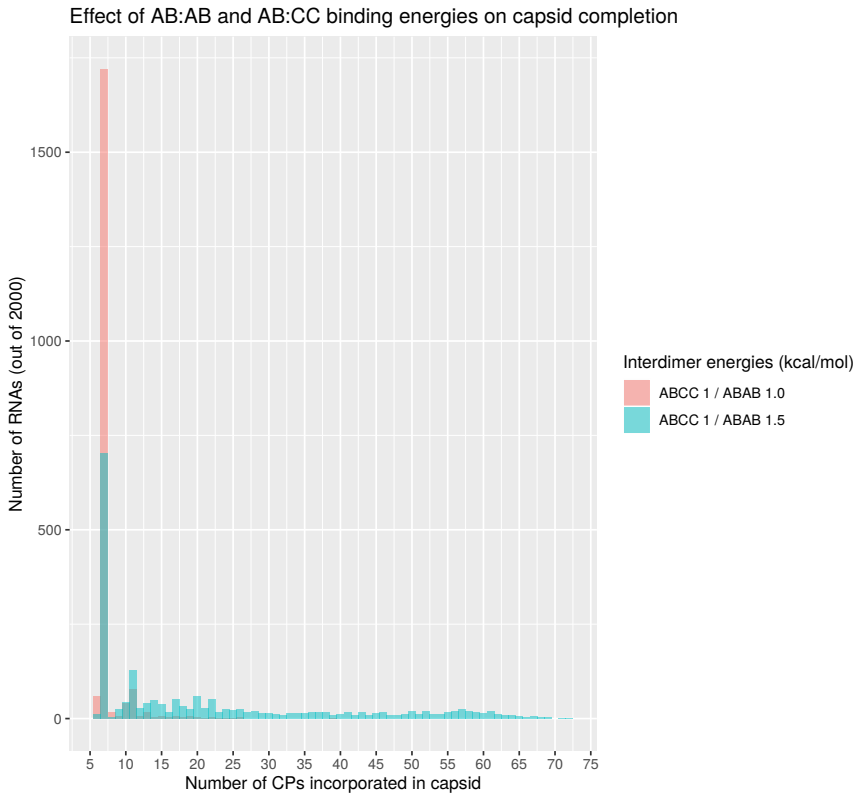

FIGURE 8.13: **Incorporated CPs per RNA for AB:CC energy -1.0 kcal/mol using the large nucleus without pull factor.**

assembly efficiency due to the number of PSs left to incorporate at each end, the PS positions of the 5' MPC and TR were varied. The 5' MPC PS was varied between 1 and 11, whereas TR was between 20 and 50. While a rather wide range of positions was included in this test, not all make sense given the nucleotide positions of the respective SLs. Since the 5' MPC is close to 400 nucleotides, it is unlikely to be between the first and second PS on the RNA. TR, on the other hand, is close to the middle of the RNA and is therefore expected to be closer to PS 30. Assembly efficiency was measured through the number of fully assembled capsids at the end of each simulation. In the energy tests it transpired that using AB:AB of -0.5 kcal/mol gave the best results with AB:CC of -4.0, -3.0 or -2.0 kcal/mol. The positions were therefore tested for all three energy combinations. Furthermore, the orientation of the 5' MPC was also tested both ways: 5' top (MP-1) and 3' bottom (MP+1) as well as swapped (see Figure 7.7

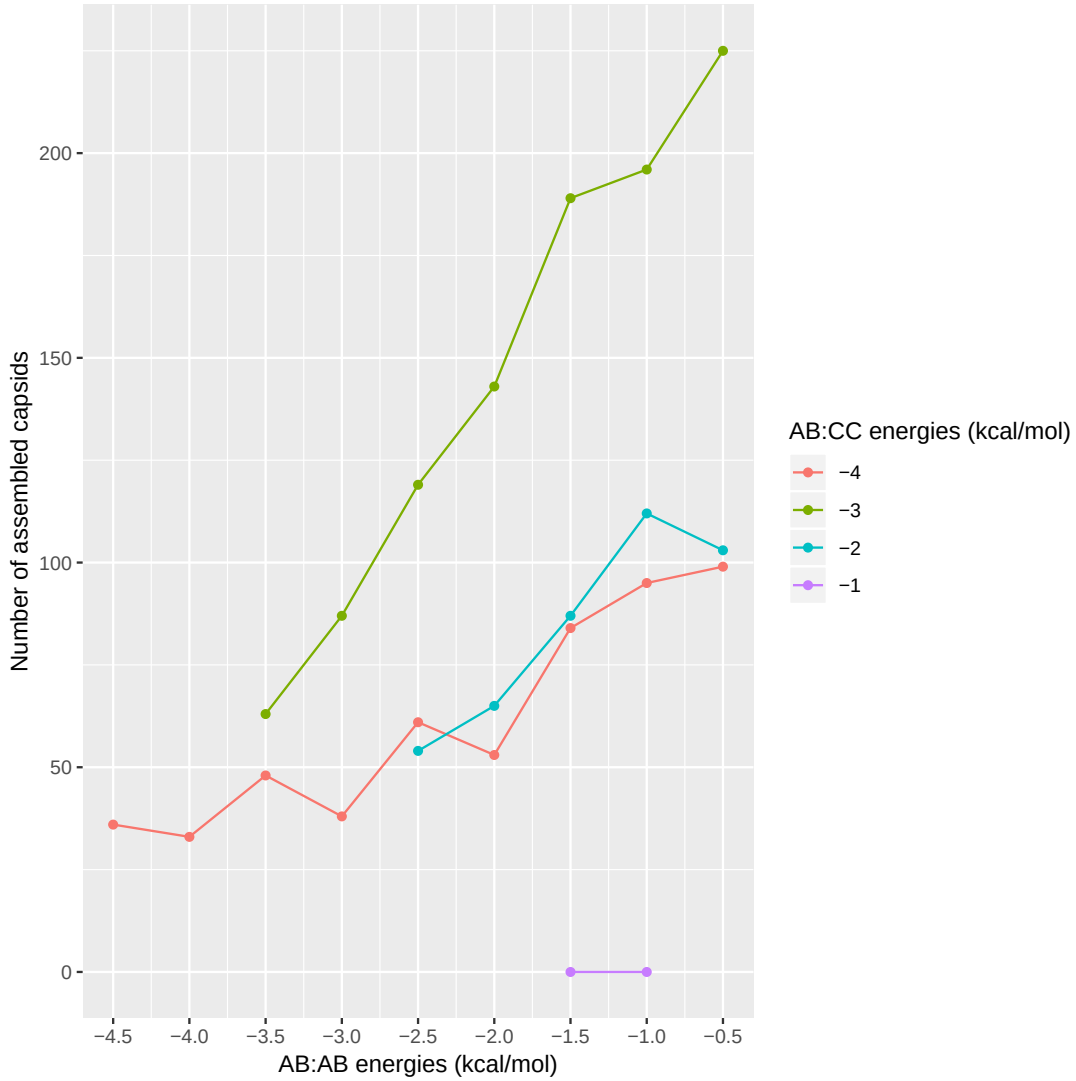

FIGURE 8.14: **Number of fully assembled capsids for different energy combinations.** All tests with one AB:CC energy are grouped: -4.0 in red, -3.0 in green, -2.0 in blue, and -1.0 kcal/mol in purple. The large nucleus was used with 5' MPC at PSs 7 and 8 and TR neighbours at PSs 29 and 30. No pull factor was included in these simulations.

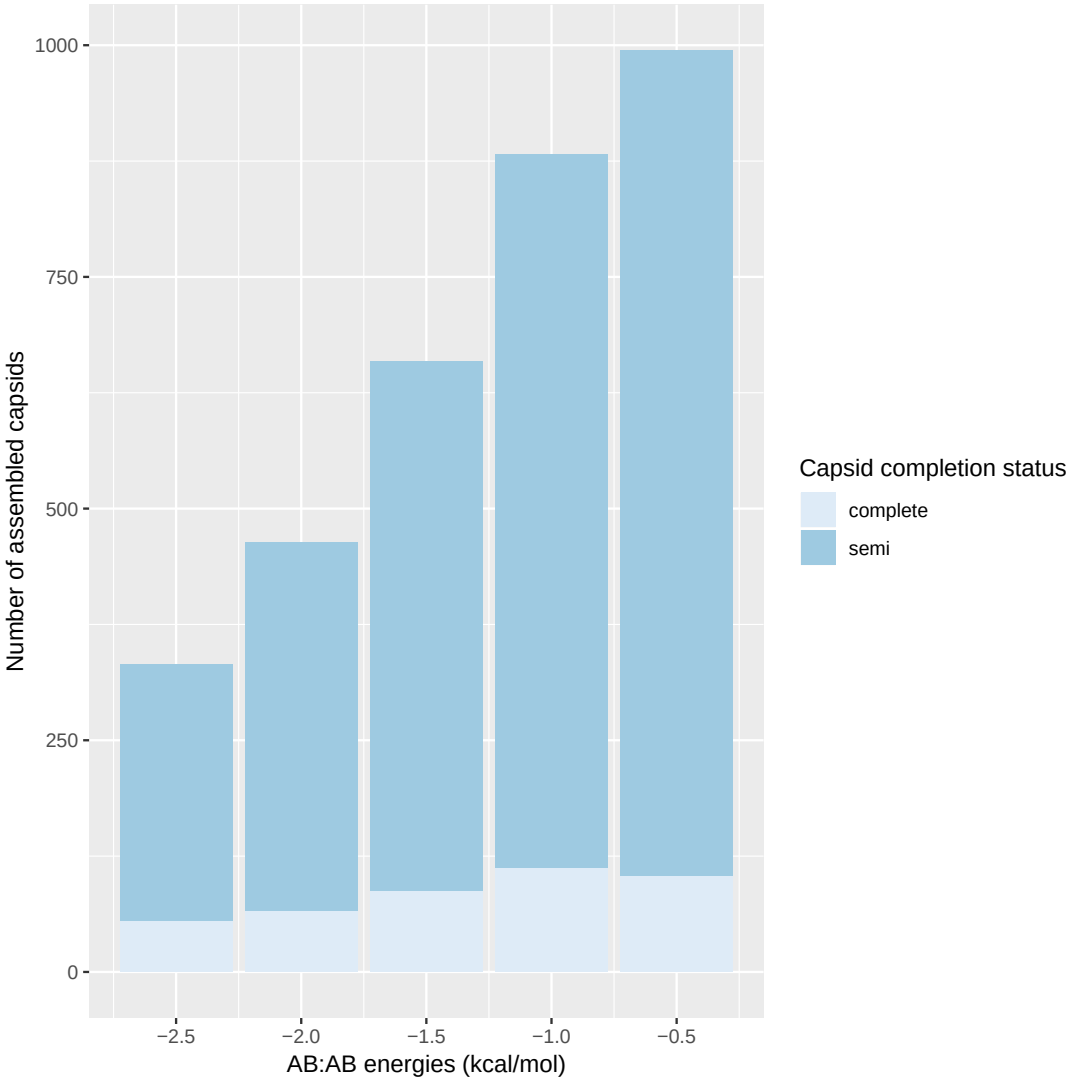

FIGURE 8.15: **Number of fully and semi assembled capsids for AB:CC -2.0 and different AB:AB.** When AB:CC energies of -2.0 kcal/mol were used, many capsids were semi complete, i.e. all PS-bound CP were incorporated but not all CC dimers. The large nucleus was used with 5' MPC at PSs 7 and 8 and TR neighbours at PSs 29 and 30. No pull factor was included in these simulations.

in chapter 7). Note that also here the TR nucleus part could dissociate later on in the simulation.

At first energy combination -4.0 and -0.5 kcal/mol were assessed for assembly performance with different nucleating PSs (Figures 8.16 and 8.17). The heatmaps showed pronounced differences in performance at different positions as well as between the original and the swapped orientation. For the original orientation having the 5' contact at PS 2 was generally better than PS 1 or PS 3 regardless of TR. Other positions were more dependent on TR, which gave the best model results when in the 20s, especially PS 21 and PS 23. The highest number of assembled capsids was observed for 5' contact PS 9 and TR PS 21 at 155 (Figure 8.16). When the 5' contact orientation was swapped, the heatmap landscape changed dramatically. Now 5' positions PS 1 and PS 3 produced better performing models. Again other positions depended more on TR, e.g. later positions generally performed worse except for TR PS 21 or PS 23, similar to the original orientation. The highest number of assembled capsids, however, was observed for 5' contact PS 5 and TR PS 47 at 177.

The next energy combination that was assessed for nucleation positions was -3.0 and -0.5 kcal/mol. Also here stark differences were observed between the original and swapped orientation (Figure 8.18 and 8.19). As expected from the energy analysis in Sections 8.2.1 and 8.2.2, using this AB:CC energy resulted in better model performance overall compared with using -4.0 kcal/mol above. Here, the worst performance still resulted in 93 and 113 assembled capsids in original and swapped orientations, respectively, whereas before it was as little as 27 and 37, respectively. General performance trends were similar. The original orientation performed better with 5' MPC PS 2 than PS 3. When the orientation was swapped, this trend was again also swapped and PS 1 or PS 3 performed better than PS 2. However, for these energies also other 5' MPC positions showed a clear improvement. Moreover, PS 7 and PS 9 performed well while PS 8 and PS 10 did comparatively poorly in the original orientation. In the swapped condition PS 6

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	77	112	69	105	116	96	132	126	108	76	87
21	66	123	95	142	106	114	108	144	155	82	105
22	54	90	85	103	98	97	96	90	103	58	112
23	44	131	81	106	128	111	124	135	131	102	91
24	57	106	74	102	109	109	95	107	114	83	108
25	52	125	71	87	87	82	105	83	103	77	102
26	52	117	81	95	115	109	118	100	109	95	68
27	60	101	74	72	89	89	93	68	108	87	79
28	56	110	74	98	90	87	120	89	129	95	92
29	50	96	64	91	89	67	99	84	101	74	69
30	63	115	62	94	102	91	107	105	103	93	86
31	56	99	88	76	81	85	70	57	81	84	73
32	47	108	60	83	75	80	72	80	103	77	70
33	55	116	81	97	89	88	101	84	98	95	64
34	44	88	58	69	88	85	87	79	71	83	67
35	61	101	58	79	90	78	103	75	75	75	82
36	51	109	60	76	82	87	78	66	76	94	72
37	31	105	64	75	86	88	87	61	104	65	76
38	67	116	81	74	68	89	94	105	85	80	75
39	52	114	81	66	71	78	104	71	96	86	71
40	48	87	71	94	86	85	96	59	88	90	56
41	36	120	75	70	115	100	108	81	106	76	92
42	52	93	59	76	80	81	85	85	83	81	77
43	50	112	68	67	106	86	104	62	117	90	86
44	51	120	37	88	94	95	92	69	91	84	61
45	40	97	87	89	94	88	124	67	101	98	62
46	66	117	75	82	116	87	115	70	77	80	62
47	73	109	73	90	102	88	86	79	136	95	72
48	53	87	73	73	81	83	72	75	103	75	81
49	61	96	84	83	96	72	96	64	79	39	68
50	38	116	79	81	93	107	101	80	94	92	85

FIGURE 8.16: **Effect of nucleation PS site combinations on assembly model performance with -4.0 AB:CC energies in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -4.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	156	91	134	121	120	93	58	76	37	115	27
21	122	113	152	106	127	100	121	89	140	81	63
22	154	104	112	107	96	106	78	79	52	90	28
23	150	98	139	111	93	110	95	106	104	86	101
24	130	83	129	93	111	122	90	83	74	106	39
25	95	97	145	97	95	92	82	79	86	106	66
26	163	88	139	120	100	106	65	105	77	127	65
27	139	102	132	77	118	79	76	79	69	97	70
28	114	76	127	96	109	115	72	99	81	123	78
29	128	77	136	120	84	97	96	105	56	90	83
30	131	101	127	96	106	114	68	77	73	111	52
31	114	90	116	81	102	105	82	86	65	102	53
32	122	81	104	102	95	82	98	77	67	99	66
33	158	87	153	97	101	95	92	89	87	104	74
34	119	82	112	88	98	89	80	68	59	87	74
35	99	84	141	100	82	99	72	71	83	95	70
36	100	99	109	121	92	102	80	96	72	98	58
37	108	85	123	87	102	84	81	73	81	78	69
38	109	100	135	106	101	126	88	111	55	112	63
39	122	97	125	124	99	90	73	83	94	91	86
40	133	94	116	104	107	124	71	85	74	98	60
41	162	111	147	92	116	105	97	102	92	107	89
42	118	77	99	90	129	115	89	94	72	103	76
43	156	109	127	122	132	111	64	87	97	77	92
44	98	91	105	116	126	94	76	85	57	106	92
45	121	96	97	137	91	112	81	97	88	81	96
46	107	87	90	116	118	118	81	77	110	104	84
47	140	97	123	112	177	146	93	105	107	97	87
48	102	98	90	130	108	100	78	86	70	100	85
49	69	68	126	103	118	109	48	84	80	86	70
50	113	123	106	124	106	133	104	126	104	116	76

FIGURE 8.17: **Effect of nucleation PS site combinations on assembly model performance with -4.0 AB:CC energies and swapped 5' MP contact site in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -4.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	224	245	151	214	232	176	227	169	204	144	175
21	199	247	169	224	223	221	241	219	252	154	214
22	203	204	178	214	168	191	224	180	221	176	188
23	230	290	167	219	211	200	262	228	255	211	235
24	206	231	154	218	219	190	217	233	236	180	228
25	216	263	172	210	184	208	256	204	243	190	229
26	218	245	171	222	191	227	268	199	263	187	218
27	225	240	151	222	184	197	228	216	254	193	219
28	256	292	179	226	194	203	250	231	283	223	226
29	211	232	150	193	166	196	225	224	212	220	215
30	244	298	170	223	213	225	259	247	275	221	244
31	229	283	179	200	193	224	242	210	273	215	227
32	233	246	145	190	199	208	248	213	266	219	245
33	218	282	152	218	211	223	257	221	272	206	241
34	224	199	135	170	158	185	212	174	265	213	219
35	252	261	157	234	200	203	262	208	289	168	208
36	202	234	163	187	179	201	243	190	245	198	193
37	234	247	129	212	186	219	277	182	244	207	235
38	213	250	159	182	179	185	239	188	260	196	220
39	198	238	137	174	190	201	244	169	242	196	225
40	226	231	126	201	189	195	250	203	259	194	227
41	184	227	147	171	193	167	224	163	243	203	196
42	180	195	121	196	158	192	263	158	253	188	179
43	194	253	124	173	193	182	222	159	226	161	168
44	186	219	123	175	214	148	193	142	219	177	172
45	212	221	126	161	209	181	233	177	278	177	208
46	173	212	111	153	209	175	237	172	192	168	162
47	193	197	107	186	186	175	222	150	223	183	196
48	170	217	135	120	187	152	185	133	196	160	150
49	141	160	93	139	150	152	150	122	184	121	166
50	136	176	97	144	144	139	198	134	213	143	161

FIGURE 8.18: **Effect of nucleation PS site combinations on assembly model performance with -3.0 AB:CC energies in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	234	172	260	177	152	193	160	141	130	210	124
21	252	150	237	193	156	219	205	144	180	182	140
22	256	171	246	192	152	236	190	141	127	193	113
23	275	170	282	210	164	242	196	172	188	208	189
24	268	197	253	212	150	252	206	165	155	196	122
25	297	163	236	191	175	217	197	148	185	206	160
26	322	175	234	189	157	233	162	190	180	227	166
27	266	159	241	195	177	186	191	170	150	190	198
28	289	173	245	204	161	240	207	195	169	247	195
29	252	198	275	195	168	242	178	209	202	206	171
30	283	176	287	199	178	242	204	161	198	276	193
31	263	169	240	191	156	233	199	208	182	240	173
32	224	168	263	204	188	263	168	186	191	227	174
33	308	214	254	204	169	247	183	207	193	252	187
34	264	160	254	202	207	232	170	162	168	210	193
35	267	191	243	192	169	262	205	191	217	236	179
36	261	170	245	222	194	240	182	181	195	234	179
37	291	155	226	210	173	215	189	177	221	218	185
38	272	185	255	225	160	273	156	186	190	189	168
39	250	187	265	206	175	231	175	167	192	206	188
40	272	155	235	238	171	270	201	181	213	223	182
41	232	180	239	197	210	221	180	165	185	187	202
42	218	157	231	234	185	255	154	197	199	217	157
43	205	163	193	191	183	263	181	181	236	205	180
44	224	150	217	178	190	203	178	185	157	200	172
45	245	157	205	164	165	275	168	211	204	191	203
46	188	130	177	176	165	222	154	148	217	175	162
47	224	143	190	222	209	267	190	186	178	185	212
48	205	168	157	167	188	187	162	155	179	172	163
49	148	118	171	144	169	194	145	164	167	153	144
50	180	147	154	192	178	198	159	189	190	156	181

FIGURE 8.19: **Effect of nucleation PS site combinations on assembly model performance with -3.0 AB:CC energies and swapped 5' MP contact site in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

and PS 10 performed well and the other positions poorly. The highest number of fully assembled capsids at 298 and 322 was observed for 5' MPC PS 2 with TR PS 30 and 5' MPC PS 1 with TR PS 26 for original and swapped, respectively.

Finally, also AB:CC -2.0 and AB:AB -0.5 kcal/mol was tested for performance using different nucleating PSs. Interestingly, for these energies the patterns observed in the histograms were different. Original orientation performed well with 5' MPC at PS 1 but neither version showed a strong pattern otherwise. The distribution of values looked more random (Figures 8.20 and 8.21). This may be due to this condition producing high numbers of semi-complete capsids and the number of associated CC is not influenced by the position of the nucleating PSs. Moreover, it was also shown that these energies do not preserve the nucleus well so the effect would be less pronounced as the positions often changed later.

As expected there was a difference in assembly efficiency between different nucleating PS positions. These depended also on the energy combinations used and the orientation of the 5' MPC. The box plots in Figure 8.22 show the distribution of values for completed capsids under each condition. It was similar for each energy condition between original and swapped 5' MPC orientation but differed considerably between the energy conditions. AB:CC of -3.0 kcal/mol produced a much larger variance in values than the other two but also the highest maxima. This showed that while best performance could be achieved using AB:CC energy of -3.0 kcal/mol, depending on the nucleus positions other energy combinations perform just as well or even better.

8.2.4 Extended Large Nucleus

The highest number of fully assembled capsids was observed with AB:CC and AB:AB energies of -3.0 and -0.5 kcal/mol, swapped orientation, and 5' MPC PS 1. As mentioned earlier, a 5' MPC at PSs 1 and 2 is unlikely due to the position of the maturation protein contacting SL around 400 nucleotides. If nothing else, two Hong PSs are situated upstream of this contact site at 102 and 179 nucleotides.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	117	106	97	102	94	100	103	113	144	115	112
21	122	93	87	113	94	109	125	108	114	107	118
22	123	97	92	95	88	120	112	126	142	119	110
23	131	82	74	117	95	110	130	113	114	98	114
24	115	82	100	113	100	86	111	97	111	138	122
25	93	82	102	107	107	116	116	115	129	79	106
26	117	96	81	103	116	120	120	115	121	112	115
27	126	104	85	111	100	100	102	100	124	112	100
28	116	91	101	89	98	90	111	108	142	103	113
29	127	96	99	114	90	80	103	105	100	100	113
30	136	87	99	96	105	116	100	133	113	102	119
31	111	103	91	107	100	107	110	111	103	102	142
32	125	79	100	94	110	99	111	103	124	118	127
33	122	87	107	119	108	130	104	119	126	104	124
34	112	72	101	112	99	97	120	95	106	106	105
35	123	82	86	107	107	114	116	123	116	101	107
36	127	88	101	97	114	116	105	101	126	113	103
37	100	90	93	111	97	138	98	124	109	119	108
38	119	89	76	113	101	96	124	113	122	109	115
39	101	83	95	123	114	117	96	96	120	107	125
40	119	71	92	105	113	115	105	120	123	108	115
41	105	93	92	103	114	112	141	119	110	108	111
42	111	96	92	91	97	101	108	108	115	108	115
43	111	72	88	113	101	101	117	103	130	115	108
44	123	83	112	95	120	100	122	127	93	97	133
45	129	99	95	98	95	103	132	109	113	102	136
46	141	103	105	91	92	116	120	101	107	89	113
47	155	94	88	125	112	94	110	110	105	118	117
48	97	85	85	102	108	106	101	112	110	109	107
49	109	88	94	95	89	108	98	117	121	114	102
50	114	89	103	85	102	101	105	101	109	113	109

FIGURE 8.20: **Effect of nucleation PS site combinations on assembly model performance with -2.0 AB:CC energies in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -2.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60										
	1	2	3	4	5	6	7	8	9	10	11
20	126	111	89	109	100	108	112	116	115	83	125
21	133	114	104	102	99	101	99	110	132	97	134
22	121	106	94	106	110	114	104	107	103	95	122
23	107	98	118	117	93	113	119	111	94	98	110
24	128	106	95	92	94	104	102	91	103	102	115
25	121	101	121	102	109	120	114	100	93	118	109
26	106	104	106	112	99	124	122	90	108	124	108
27	117	112	93	101	114	118	121	117	104	110	107
28	124	102	97	106	111	112	117	120	102	99	112
29	141	99	85	118	113	104	98	121	102	106	122
30	129	109	110	109	114	111	110	113	105	114	114
31	132	95	108	111	96	133	113	96	129	125	97
32	103	106	109	96	100	105	95	97	104	118	119
33	115	79	104	98	92	110	107	105	102	104	109
34	111	99	103	92	116	124	118	112	89	106	120
35	106	109	96	106	102	116	102	116	112	88	112
36	116	89	90	112	106	129	107	122	114	102	109
37	106	95	95	93	90	116	113	116	108	106	122
38	135	106	105	96	103	132	102	101	106	127	106
39	116	108	94	115	125	120	104	107	90	106	117
40	103	101	93	98	94	125	110	107	109	127	112
41	118	83	100	103	116	98	102	105	84	115	94
42	117	106	105	93	99	131	109	107	129	101	92
43	114	102	97	109	120	125	97	104	107	98	121
44	120	106	111	115	106	125	96	90	101	107	118
45	120	96	85	103	101	108	116	115	98	111	117
46	109	100	89	101	101	137	107	105	95	114	104
47	100	93	109	128	97	103	105	111	118	111	102
48	135	107	105	103	101	117	128	112	117	101	99
49	114	108	106	109	96	125	108	125	98	96	111
50	118	129	109	116	76	105	92	102	109	100	118

FIGURE 8.21: **Effect of nucleation PS site combinations on assembly model performance with -2.0 AB:CC energies and swapped 5' MP contact site in the large nucleus model.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 1 and 11 and 20 and 50, respectively. Interdimer energies of -2.0 AB:CC and -0.5 AB:AB were used.

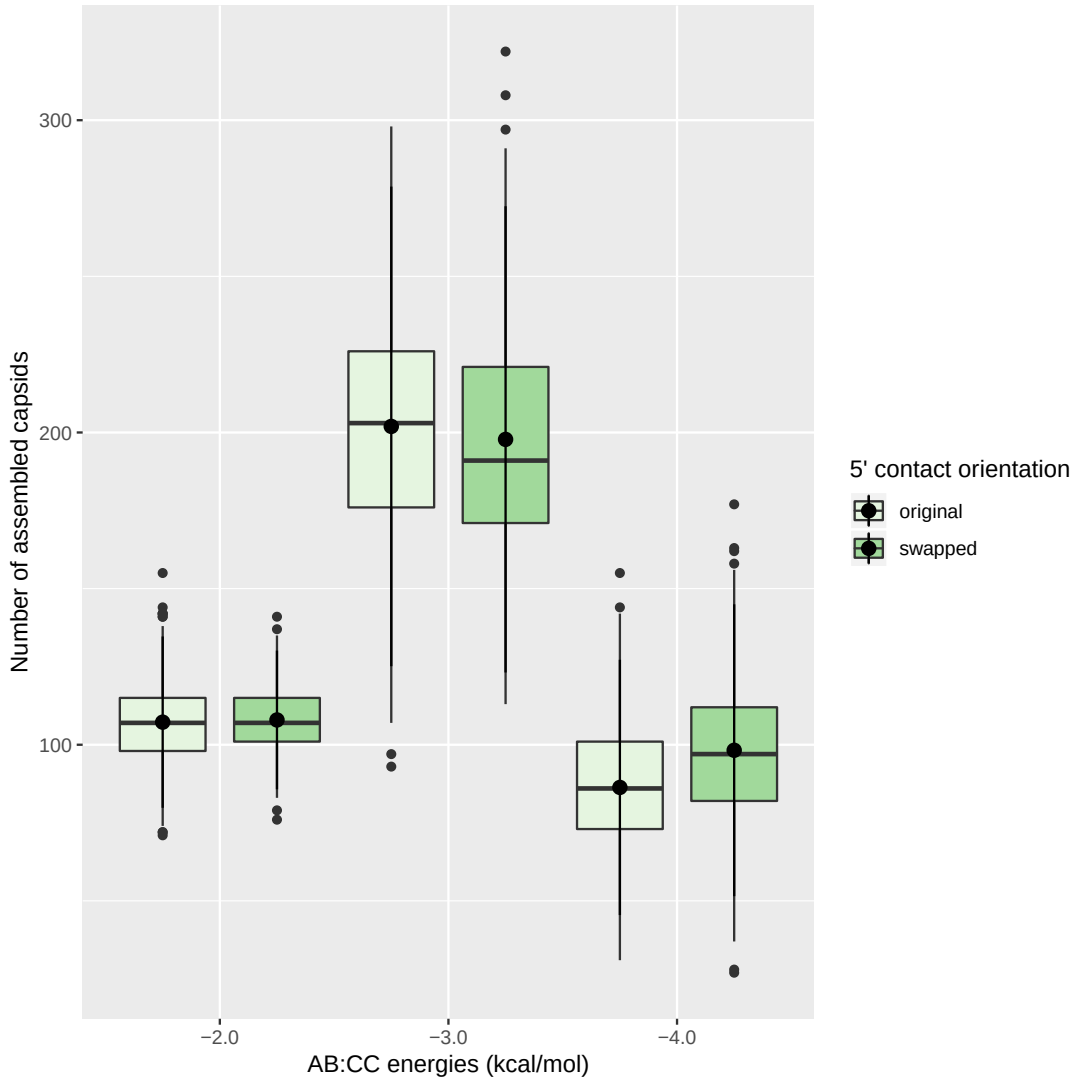

FIGURE 8.22: **Number of assembled capsids across nucleation positions for different energies in the large nucleus model.** Three interdimer energies were tested for different sets of nucleating PSs: AB:CC -4.0, -3.0, and -2.0 kcal/-mol with AB:AB -0.5 kcal/mol. Each was performed with the 5' nucleation site either in original orientation or swapped. The box plots also show mean, and standard deviation.

The latter may be MP-1 so the position for 5' MPCs is at least PS 2 but likely even higher. The next best outcome was observed in the original orientation with 5' MPC PS 2 and TR PS 30. Using this configuration an extended large nucleus was tested. The idea was that the virus would assemble through a specific path to avoid trapping the different ends. *In vivo* this would be achieved through PS affinities as well as interdimer energies favouring one path over others. To test if model performance could be improved by defining the next few steps from each nucleating position, the respective nucleation step was expanded. Instead of just including the 1–2 PSs and 2–3 CPs, more PS-bound CPs were incorporated and fixed so dissociation was not possible. The extended large nucleus is shown in Figure 7.8 in Chapter 7. In order to find which further extension provided the most if any improvement, each was tested separately and in combination with the others.

Surprisingly, any of the further extended nuclei performed worse than the original large nucleus model with these parameters (Figure 8.23). While the original resulted in 298 complete capsids, extension gave at most 284 when only the 3' nucleus was further extended, which corresponds to the partial path from the 3' MPC to the conserved PS. Other configurations performed even worse.

To test if this was affected by nucleating PS positions another configuration was tested. While the 5' MPC is theoretically possible to be at PS 2, it is more likely to be farther downstream. Therefore, 5' MPC at PS 9 and PS 10 and TR neighbours at PS 35 and PS 36 were used next. These were the positions that yielded the highest number of assembled capsids for downstream 5' MPC. In the original setting this position resulted in 289 fully assembled capsids. This was only improved upon when TR and the 5' nucleus were further extended (Figure 8.24). Then, 392 capsids were assembled. All other extensions performed worse than the original.

Considering the stark differences in extension performance between the two positions, the nucleus position test was repeated for all extensions, i.e. only 5',

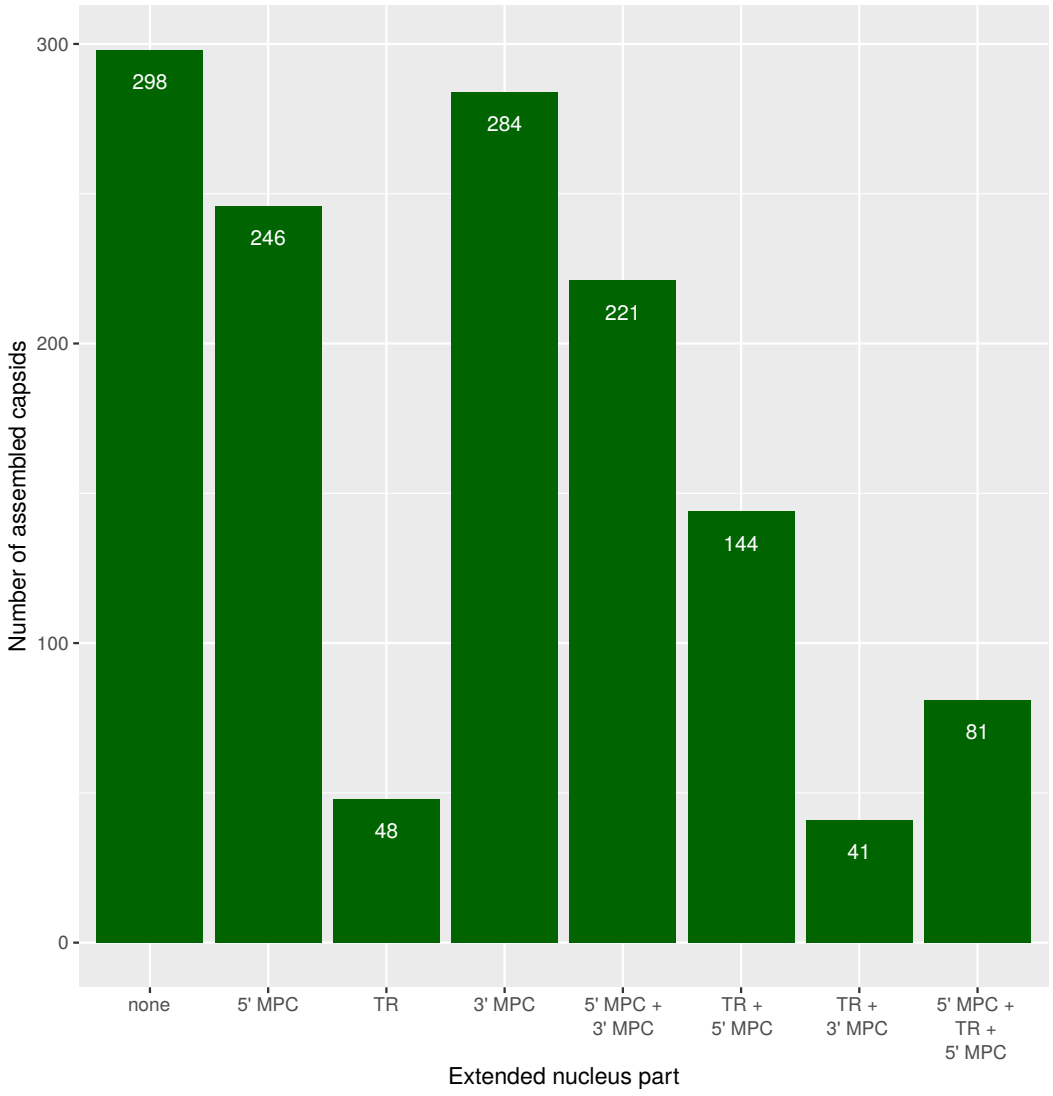

FIGURE 8.23: **Number of assembled capsids for each nucleus extension using 5' MPC 2 and 3 and TR 30 and 31.** Interdimer energies were AB:CC -3.0 kcal/mol and AB:AB -0.5 kcal/mol.

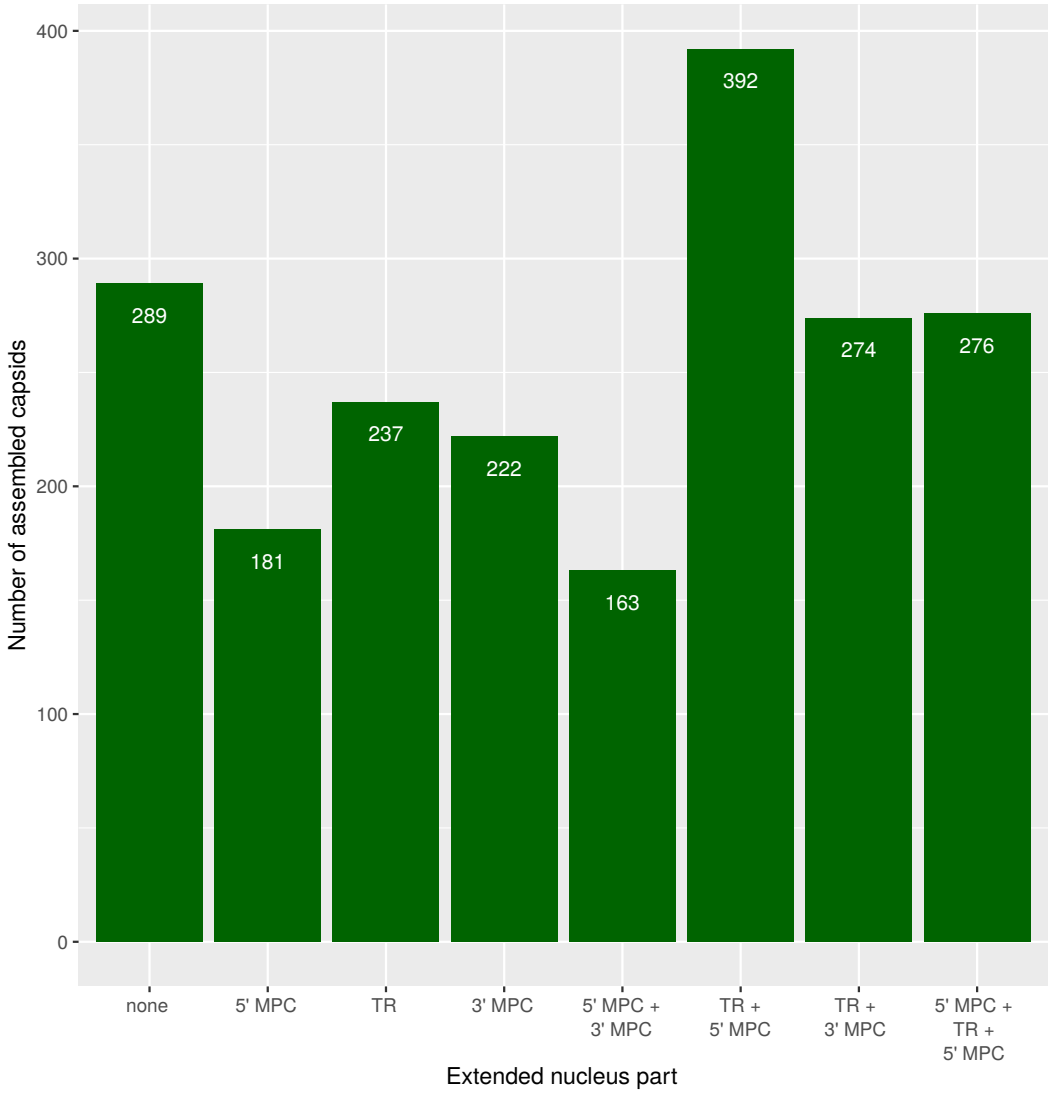

FIGURE 8.24: **Number of assembled capsids for each nucleus extension using 5' MPC 9 and 10 and TR 35 and 36.** Interdimer energies were AB:CC -3.0 kcal/mol and AB:AB -0.5 kcal/mol.

only TR, only 3', 5' and 3', TR and 5', TR and 3', and all. The heatmaps for all of these are shown in Figures A.3–A.9 in Appendix A. Differences in performance between positions were considerable. For example, the number of assembled capsids ranged from as little as six to as many as 409 when only TR was extended (Figure A.4). To better compare the distribution of values across positions all were plotted next to each other along with the original, unextended (Figure 8.25). The largest range of values was observed when TR was extended regardless of the other parts of the nucleus. The maximum values from these were higher than the highest in the unextended control and the minimal values lower without considering specific positions. The highest number of fully assembled capsids was 617, when both TR and 5' MPC were extended and positions PS 20 and PS 21 for TR and PS 10 and PS 11 for 5' MPC were used (Figure A.7).

Taken together these data show that while extending some nucleus ends can improve assembly, this effect largely depends on the nucleating PS positions. These introduce a wide range of variation especially when the TR part of the nucleus is extended. For most combinations of conditions the effect was in fact negative and can reduce the number down to fewer than 10 assembled capsids in the case of TR. Surprisingly, 3' MPC extension had the least effect on assembly success despite reflecting the position of the conserved PS.

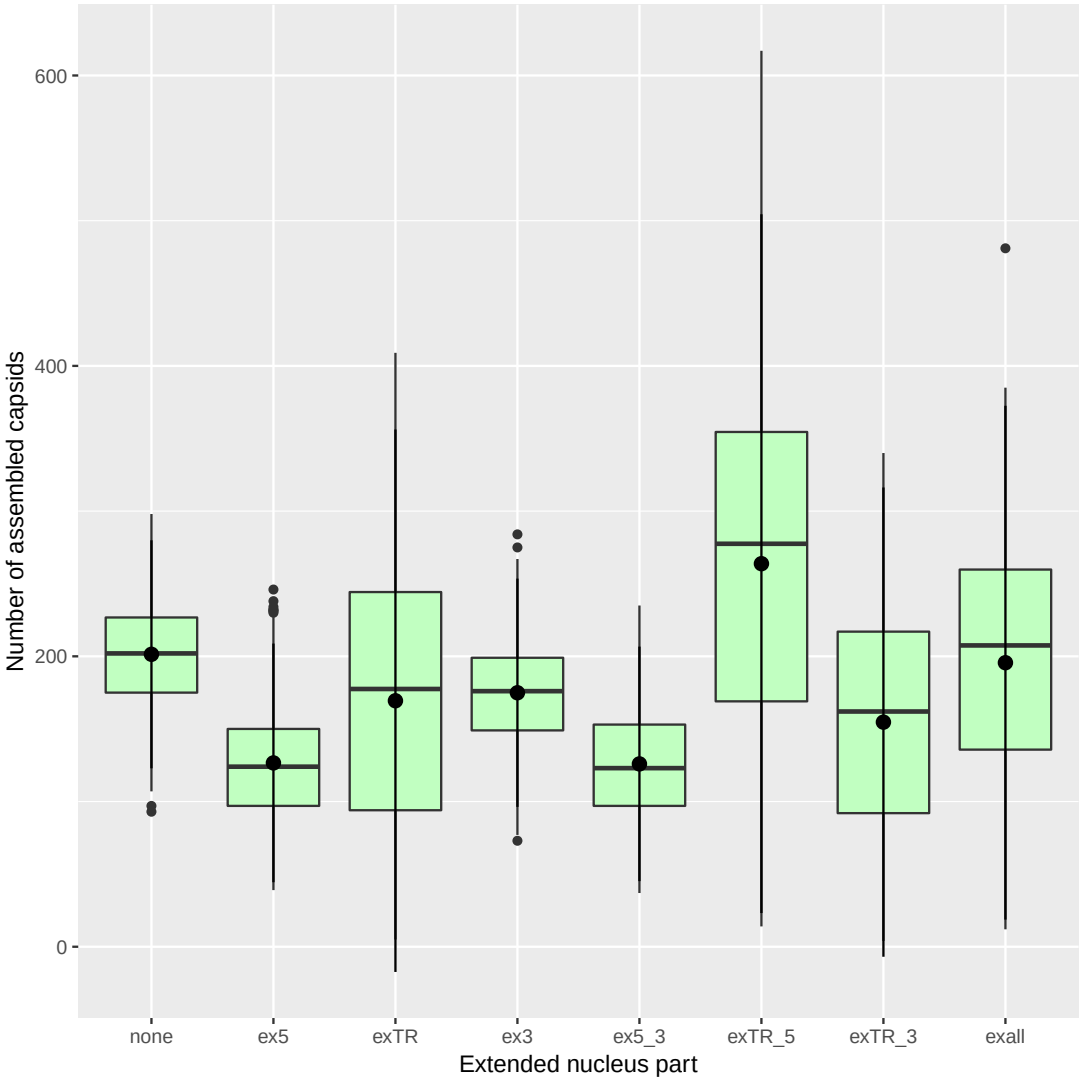

FIGURE 8.25: **Number of assembled capsids across nucleation positions for each nucleus extension.** All seven possible combinations of nucleus extensions were tested for different sets of nucleating PS positions. Interdimer energies were AB:CC -3.0 kcal/mol and AB:AB -0.5 kcal/mol. The box plots also show mean, and standard deviation.

8.3 Discussion

The original model, which used a minimal nucleus of only TR and a neighbour, suffered from trapping of capsid intermediates. At first the model was run using AB:CC interaction energies of -4.0 kcal/mol and AB:AB interaction energies of -4.5 kcal/mol. These energies are quite low and result in very stable capsid intermediates. This could be seen in the histograms, which revealed a large number of intermediates of a small number of sizes (Figure 8.1). It turned out that a large part of these were in fact trapped since there were no further steps possible along a Hamiltonian path. Theoretically, error correction is possible in the model: The outermost bound CPs can dissociate from an intermediate and reattach elsewhere at another point. However, with the energies being that low the already formed interactions are too stable especially when surrounded by several CPs as is the case in a dead-end where multiples of these energies would come into play. Therefore, practically little error correction can take place. Trying out different interdimer energy settings revealed that higher energies, especially between AB dimers drastically improved model outcomes. Fewer trapped low CP intermediates were generated at higher interdimer energies. The best outcome was observed for AB:AB -0.5 kcal/mol and AB:CC -3.0 kcal/mol. At this point a good balance was reached between error correction and capsid stability. When AB:CC energies were increase further to -2.0 kcal/mol, more capsids were semi-assembled, i.e. all AB dimers were incorporated but some CC dimers were missing. The use of higher AB:AB and lower AB:CC energies is justified by previous research. ElSawy et al. (2010) showed that while AB:CC interactions get stabilised through TR binding to AB, the opposite is the case for AB:AB. Binding of a PS to a CP reduces repulsion between this CP and another, unbound one. However, two PSs would conversely introduce too much negative charge and repel each other, which results in less stable AB:AB interactions (ElSawy et al., 2010).

Whilst adjusting the interdimer energies improved the overall landscape of intermediates, this was still too little to produce meaningful amounts of completed capsids. More successful was the inclusion of a larger nucleation complex. This nucleation complex was based on the PSs conserved among *Leviviridae* and their relative positions in the asymmetric capsid structure by Dai et al. (2017). Assembly was, therefore, modelled to nucleate around the maturation protein from TR and the 5' and 3' MPCs (Shiba and Suzuki, 1981; Rumnieks and Tars, 2017). This involved a contact at the very 3' end of the RNA, one at the 5' end but not before the first PS, and a pair of PSs around TR. Instead of being built from two ends, i.e. 5' and 3' of TR, the capsid was built from five ends, i.e. 5' and 3' from the 5' MPC, 5' and 3' from TR, and 5' from the 3' MPC. When a condition was included to try to ensure that the respective ends met, e.g. 3' from 5' MPC with 5' from TR, more trapping of intermediates occurred and no capsids were successfully assembled irrespective of interdimer energies used. Only when the ends were assumed free and independent from each other was an improvement in assembly performance observed. This resulted in a set of partial Hamiltonian paths rather than one complete one. Thus far MS2 assembly had been assumed to follow a Hamiltonian path (Toropova et al., 2008; Dykeman et al., 2011, 2013b; Stockley et al., 2013b; Twarock et al., 2018) meaning that whilst the RNA dips into the interior of the capsid, it re-emerges close by so that consecutive PSs are organised next to each other under the protein shell. However, in their cryo-EM structure Dai et al. (2017) observed that whilst the RNA is seen directly under and in contact with the capsid protein layer, it sometimes resurfaces at other places than it dips into the interior of the capsid from resulting in a broken up path. It is therefore not necessary to limit the capsid assembly model to a complete Hamiltonian path. The five independent ends, from which the capsid is built up, do not need to meet as they can be assumed to connect through the interior of the capsid. Of course, not every pair of successive PSs has enough sequence between them to allow for such a connection. *In vivo* the assembly process

would be more restricted by the actual distance between the specific PSs. This not only affects whether or not a connection through the interior is possible but even whether the two PSs need to be next to each other around the 5-fold axis of symmetry or can be further apart across the 3-fold axis (Dykeman et al., 2013b). In the model these distances are not included and all successive PSs are treated equally. In the future distances between PSs and subsequently possible moves from one to another may be included in the model. This could be evolved to find the optimal set of distances/moves to enable efficient assembly, which could be compared with predicted PSs in the respective virus.

When modelling PS-mediated capsid assembly nucleating from pre-defined positions, where those positions are, can have a significant effect on model performance. Since MS2 is thought to have 60 PSs, one for each AB dimer, the RNA is modelled as having 60 successive PS positions, e.g. from PS 4 there are three PSs upstream and 56 downstream. While some specific SLs have been identified as PSs experimentally (Valegård et al., 1994, 1997; Dai et al., 2017), most are predicted based on an assumed PS motif in MS2 (Dykeman et al., 2013b). It is therefore currently impossible to fully and reliably predict the position of a particular PS among the 60. Testing variations of nucleating PS positions out of 60 illustrated the importance of this consideration. Even a shift of one in PS position could severely change the performance of the model. Interestingly, which positions performed better was also somewhat influenced by the AB:CC energies applied. This may be due to certain energies favouring particular moves and paths over others making some start positions more favourable than others. The largest variation of and best performance was observed for AB:CC -3.0 kcal/mol and AB:AB -0.5 kcal/mol.

In vivo further restrictions for the assembly paths may be occurring to minimise trapping of growing ends. Therefore, it was tested whether an extension to either nucleating end would improve model performance. Using the best outcomes from the previous position test, two combinations were tested with AB:AB

energies of -3.0 kcal/mol. Surprisingly, most extensions resulted in fewer assembled capsids meaning that more flexibility at those positions performed better. When all combinations of 5' MPC and TR positions were tested, a large variation from almost no assembled capsids to more than double the maximum assembled without extension were observed. Since the best performance without extensions (298) was not obtained with the same combination of positions (PS 2 and PS 30 versus PS 10 and PS 20) as the best performance with extensions (617), it is difficult to pinpoint the optimal combination of positions at this point. More biological data about whether certain paths are always taken *in vivo* when assembly continues from the nucleus positions would be needed to resolve this problem.

Assembly and packaging are also influenced by PS affinities (Dykeman et al., 2013a; Stockley et al., 2013b; Dykeman et al., 2014). The tests performed for this chapter utilised a set of PSs that all had the same high affinity of 1.5 nM or -12 kcal/mol when converted to free energy of formation, corresponding to TR's affinity. Using uniform affinities meant that other attributes could be tested without fear of overlapping effects. This was especially important for the nucleation position test. When the nucleating PSs were varied, it was essential that the surrounding PSs did not differ in affinities. The point of this test was to determine which sites would be most favourable for assembly purely from a geometry point of view. PS affinities would have added another layer of complexity, since assembly would have then also been affected by high or low PS affinities in certain positions. In the future it will be important to investigate PS affinities as well. Using the now determined optimal nucleation positions and interdimer energies, affinities could be evolved to find the optimal set. This will further increase the number of assembled capsids and provide further insights into the function of this virus.

The work on this MS2 assembly model highlighted the advantage of a large assembly nucleus when building such a complex capsid. It also further validated the observations by ElSawy et al. (2010) that AB:AB interdimer energies are

weaker than AB:CC energies and the importance of this phenomenon for efficient assembly. Only when this was incorporated did assembly begin to occur even in the minimal nucleus model. The largest effect, however, had the addition of the maturation protein nuclei. Whilst the exact relative PS positions could not be reliably determined, the importance of a large nucleus, as predicted from conservation, was seen. This showed how conservation of PSs could be used to gain a better understanding of the assembly process of a virus.

Chapter 9

General Discussion

The aim of this project was to investigate whether conservation of PSs could be used to understand viral evolution and to identify PSs that have additional important functions particularly in assembly nucleation. The thesis followed a two-strand approach to what can be learned from PS conservation: Whilst variable sites could be used for phylogenetic analysis and thus provide insights into how PSs evolve, the highly conserved sites are indicators of essential function, which was thought to transcend a mere role in genome packaging and capsid assembly.

9.1 Variable PSs Inform Phylogenies

In single-stranded RNA (ssRNA) viruses, PSs—small sequence/structure motifs that bind viral CP—are crucial for ensuring specific packaging of viral RNA and efficient assembly of capsids (Stockley et al., 2007; Dykeman et al., 2011; Bunka et al., 2011; Stockley et al., 2013b; Ford et al., 2013; Dykeman et al., 2013b, 2014; Rolfsson et al., 2016; Twarock and Stockley, 2019). PSs bind to CP units and facilitate interaction between these units via virus-specific mechanisms, e.g. in bacteriophage MS2 PS-CP interaction results in a conformational change in the CP dimer (Stockley et al., 2007; Dykeman and Twarock, 2010; Dykeman

et al., 2010, 2011; Stockley et al., 2013b; Twarock and Stockley, 2019). However, whether a similar mechanism is used by DNA viruses that replicate through an RNA intermediate, i.e. in retro- and pararetroviruses, is poorly understood. In Chapter 3, I showed in collaboration with experimentalists that the pararetrovirus hepatitis B virus (HBV) contains SLs in its pre-genomic RNA (pgRNA) with a shared sequence motif in the apical loop, which bind to HBV capsid protein and trigger re-assembly of viral capsids *in vitro* (Patel et al., 2017). These results stand in contrast to previous work in HBV, which assumed and concluded the existence of only a single PS ε (Junker-Niepmann et al., 1990; Bartenschlager et al., 1990; Bartenschlager and Schaller, 1992; Pollack and Ganem, 1993; Fallows and Goff, 1995; Hu and Boyer, 2006). However, these studies failed to consider alternative explanations for their observations, such as that ε functions most importantly to switch off translation freeing up the pgRNA to form other SL structures presenting PSs. Without this initial step PSs cannot form, especially in eukaryotic cells, which employ a ribosome with helicase function. In prokaryotic expression systems HBV capsids are usually found containing RNA even in the absence of ε (Birnbaum and Nassal, 1990; Crowther et al., 1994). Furthermore, foreign sequences utilised in the above experiments were found here to also contain PS-like structures. These results point to the importance of ε in translation suppression and thus packaging initiation, whilst additional PSs dispersed in the pregenome play additional roles in assembly. Analysis of the content of HBV capsids expressed in prokaryotic cells for PS-like motifs in comparison with general cell messenger RNAs (mRNAs) would provide further evidence for this interpretation.

Initially, a single PS was often identified for a viruses such as TR in MS2 (Carey et al., 1983a,b; Beckett and Uhlenbeck, 1988; Rolfsson et al., 2008) or ε in HBV (Junker-Niepmann et al., 1990; Pollack and Ganem, 1993) that is essential in assembly nucleation and highly conserved. The concept of PS-mediated assembly implies the existence of several dispersed PSs (Dykeman et al., 2011; Stockley

et al., 2013b; Dykeman et al., 2013a, 2014). In Chapter 3, I provided evidence for the existence of additional PSs in HBV (Patel et al., 2017), and recently for the first time a cryo-EM structure of MS2 visualised several PSs in contact with CP dimers (Dai et al., 2017), proving evidence for presence of several PSs. However, little is known about whether PSs, particularly secondary dispersed ones, are conserved between strains or even species or how they evolve. The first strand of this thesis tackled this lack in knowledge through the development of a phylogenetic method specifically for PSs, which was applied to HBV and *Leviviruses* MS2 and BZ13.

Phylogenetic relationships between species were originally inferred from morphological features or characters. These have become less used with the advent of sequencing methods and the subsequent widespread use of molecular characters, but are not obsolete today (Suárez-Díaz and Anaya-Muñoz, 2008; Wright et al., 2016). Viral phylogenies are usually based on genomic sequences or parts thereof. They can be used on a large range of time scales and help to understand transmission events in recent outbreaks (Kenah et al., 2016) or the evolutionary history and origin of viruses (Bollyky et al., 1998; MacDonald et al., 2000; Zehender et al., 2014). Bamford and Stuart utilised the fold of viral capsid proteins to reconstruct phylogenies by deriving characters from the relative distances of the structures in 3D (reviewed in Bamford et al. (2005)). Viruses that infect hosts from different domains of life were clustered on the resulting phylogenetic tree, providing information on structural relationships between viral families. Their work illustrates that different types of characters are useful in understanding viral evolution, particularly when they can grant access to different evolutionary time scales. In Chapter 2, I developed a method to process PSs into characters for phylogenetic reconstruction. Since PSs are SLs, the method also involved secondary structure prediction. Despite many existing RNA folding methods, which take into account SL formation energies, suboptimal structures, kinetics, co-transcriptional folding, or conservation, most still struggle to accurately predict the structure

of larger RNA molecules (Zuker, 1989; Morgan and Higgs, 1996; Mathews et al., 1999; Wuchty et al., 1999; Ding and Lawrence, 2003; Zuker, 2003; Mathews et al., 2004; Ding et al., 2005; Wiese et al., 2008; Reuter and Mathews, 2010; Lorenz et al., 2011; Proctor and Meyer, 2013; Liu et al., 2016; Zuber et al., 2017; Shi et al., 2019) or to take RNA-protein interactions into account. Both of these are essential for correctly predicting the structures of several 1000 nucleotides long viral RNA genomes that present PSs. My structure prediction approach also minimises overall energy. However, it differs from classic minimum free energy (MFE) structure prediction programs by disregarding kinetically unstable SLs, adding CP-binding energies to the energies of each putative PS, and combining small local structures into a global one, thereby mimicking co-translational folding. In direct comparison, it correctly predicted more SLs than Mfold and notably outperformed it by predicting all experimentally confirmed PSs in MS2. It thus presents a powerful secondary structure prediction method for viral RNA genomes but could also be adapted to take into account other RNA-protein interactions. In the future, it will be interesting to benchmark the method against other programs and on other viral genomes.

Phylogenetic trees based on PS distributions were expected to reveal a different time scale for viral evolution. Instead, they showed that PS evolution is restricted, reflecting the limited options for viable PS configurations and the loose relationship between sequence changes and structure/function changes for PSs, i.e. many mutations can have no effect at all whilst few, targeted mutations can delete a PS completely. In Chapter 4 I looked closely at related viral strains in HBV and showed that there are differences in PSs for sequences classed as the same genotype, whilst PS can be similar for sequences of distinct genotypes. Particularly, strains from one patient over ten years were found fluctuating and not simply clustering together as they would based on nucleotide sequence (Osiowy et al., 2010). A large shift in PS profiles was observed post antiviral treatment, which does not affect PS function, but exerts evolutionary pressure on the virus

to find escape mutants, and highlights how PSs still can be indirectly affected by anti-virus strategies that do not target PSs directly. Interestingly, in HBV PSs were also found to diversify more in early infection than later after the immune clearance phase. This is opposite to findings for genomic sequences, which tend to evolve more when the virus adapts to the new host (Osiowy et al., 2006; Sede et al., 2014). The results are in line with the colonisation-adaptation trade-off (CAT) model of viral evolution (Lin et al., 2015), since PSs are more important for fast and efficient replication, and thus the initial colonisation step of infection. Note that while the pace of evolution of the PSs differed from that of the genomic sequence itself, they are still based on that sequence. The phenomenon is akin to the study of functional traits in ecology, i.e. certain functions can be the result of a combination of underlying traits, just as PSs combine sequence and structure, but these functions may not evolve in the same way as the underlying traits (Díaz et al., 2013).

Whilst HBV provided the opportunity to examine PS evolution on relatively short time scales with different genotypes and longitudinal studies, working with the two *Leviviruses* MS2 and BZ13 in Chapter 6 made it possible to compare PS profiles between related viral species. Moreover, as opposed to HBV, in which PSs were thought not to exist before this study, PSs in *Leviviruses* are well studied, including their affinity tiers and motifs. Secondary structure prediction identified only a small number of high affinity PSs in any of the *Levivirus* strains, but an overabundance of low affinity PSs. These data confirmed previous *in silico* model studies that showed a selective advantage for RNAs with a mixture of affinities over RNAs with only high affinity PSs (Dykeman et al., 2013a). The overabundance of low affinity PSs points towards robustness of the PS-mediated assembly process against mutation of some less important PSs and the collective action of many PSs, including low affinity PSs, results in efficient capsid assembly and genome packaging (Dykeman et al., 2014). Work by Dykeman et al. (2013b) who compared Hamiltonian paths for the RNA organisation within the MS2 and GA,

a BZ13 strain, capsid and identified a highly constrained assembly pathway consistent with their assumptions and data (Twarock et al., 2018). Reconstructing phylogenetic trees proved to be challenging between species. The high observed number of PS blocks indicates either low PS conservation between these viruses or unsuccessful matching of corresponding sites. The latter highlights an important limitation of the current phylogenetic method namely that it still depends on genomic sequence alignments, which are used to shift the PS profiles and match corresponding regions. Abstraction has therefore been utilised throughout the process to minimise artefacts from irrelevant sequence/structure changes due to misalignment. For instance, PSs only need to be in the same region but the exact sequence or fold of that SL is not taken into account by generating first pseudosequences and then PS blocks. In order to be able to compare PS block profiles from different viral species, the abstraction would need to be taken further to make it independent of multiple sequence alignment (MSA). In the future, this may involve abstracting PS profiles into some encoding of relative distances, so that general patterns are discernable without superfluous detail. Further limitations arise from the use of Hamming distances for calculating distance matrices. More sophisticated substitution models for DNA exist (Jukes and Cantor, 1969; Kimura, 1980, 1981; Felsenstein, 1981; Lanave et al., 1984; Hasegawa et al., 1985; Tavaré, 1986; Tamura, 1992; Tamura and Nei, 1993; Felsenstein and Churchill, 1996; Waddell and Steel, 1997); however, since they are specifically developed for evolution of genomic sequence, they are not suitable for PS comparisons. Even newer models developed specifically for sequences with known RNA secondary structures would not be helpful, because they assume lower and paired substitution rates in the helix portions and standard ones in the single-stranded portions of the structures (Schöniger and Von Haeseler, 1994; Tillier and Collins, 1998). PSs are assumed to evolve in a more complex manner in this context, because not only are the functionally important parts of the SL usually in the single-stranded parts, but also the exact SL is less important than having any SL with

the respective motif in the correct region. Therefore, a novel substitution model would have to be developed specifically for this application. It would also need to take into account the different affinity tiers. At the moment, Hamming distances would treat a transition from a medium to low affinity PS the same as going from a high affinity PS to the absence of any PS. Naturally, these transitions are not in fact equivalent, and a more refined model would need to reflect that. Once these limitations have been addressed, PSs-based phylogenies may be examined further on additional viruses, or be used similarly to morphological characters in providing a framework for molecular phylogeny to reconstruct older evolutionary history (Scotland et al., 2003).

9.2 Conserved PSs Indicate Additional Functions

Essential functional features of a species tend to be conserved. For instance, ε (Junker-Niepmann et al., 1990; Ostrow and Loeb, 2002), DR1, and DR2 (Molnar-Kimber et al., 1984; Lien et al., 1986; Bartenschlager and Schaller, 1988; Loeb et al., 1996) are conserved between distant members of the *Hepadnaviridae* family despite relatively low sequence similarity, resulting in duck hepatitis B virus (DHBV) often being used as a model virus for studying HBV (Beck and Nassal, 1998; Ostrow and Loeb, 2002; Beck and Nassal, 2007). Note that, especially for ε, the sequence—and to a small extent also the structure—varies between species, but its function in reverse transcription and packaging is conserved. This is similar to the translation repressor (TR) PS in *Leviviridae* (Hung et al., 1969; Ling et al., 1970). The sequence varies, e.g. for MS2 it is AUUA in the apical loop compared with Qβ, which has UAA, but the function is conserved (Horn et al., 2006). An example outside of the viral realm is the cloverleaf secondary, and L-shaped tertiary, structure of tRNA, which due to its functional importance is conserved between all organisms from budding yeast to humans (Goodenbour

and Pan, 2006). Since in general important functions are conserved, in the second strand of the thesis conservation was utilised to infer important additional functions of PSs. One such function is assembly nucleation, which was examined in Chapters 5 and 6. As mentioned above, ε and TR are known to be highly conserved within the respective viral family and are considered to initiate the assembly process. However, larger groups of PSs forming a nucleation complex are often not considered, and their collaborative action is thus poorly understood.

By combining insights from the structure of a re-assembled HBV capsid with aligned PS profiles of different HBV genotypes, I identified a highly conserved PS pair as candidate for the nucleation complex in Chapter 5. Whilst to date there is no direct experimental validation for it, there is strong evidence for the functional importance of each PS in this pair. The asymmetric cryo-EM structure from re-assembled capsids indicated nucleation by 2–4 PSs and a highly conserved pair of PSs was identified. ε and DNA polymerase (Pol) at the 5'-end may act in conjunction with this pair, which is located at the 3'-end of the pgRNA, and together account for the RNA density seen in the cryo-EM structure. Moreover, this cooperation would also bring both ends of the pgRNAs into close proximity and thereby facilitate the post-assembly interaction between ε and ϕ and subsequent reverse transcription (Tang and McLachlan, 2002; Abraham and Loeb, 2006). Furthermore, the first PS of the pair showed the highest similarity to experimental sequences with strong affinity for HBV capsid protein (Patel et al., 2017), whilst the other had a major overlap with *cis*-acting element ϕ, which is essential for reverse transcription (Tang and McLachlan, 2002; Abraham and Loeb, 2006). The latter led to the hypothesis that the overlap is due to a double function of that region and the PS sequesters the *cis*-acting element into a SL ensuring that reverse transcription only takes place within the viral capsid when the PS dissociates from capsid protein and melts exposing the sequence. Such a regulatory function would play an important role in escaping detection by the host immune system. The pgRNA contains a 5' cap and a poly-adenylation (poly-A)

tail, because it is transcribed by the same RNA polymerase II that also produces host mRNAs (Rall et al., 1983; Tiollais et al., 1985), which would make it indistinguishable from mRNAs and thus non-suspicious. DNA in the cytoplasm, on the other hand, is recognised by pattern recognition receptors (PRRs) and triggers an innate immune response (Hemmi et al., 2000; Kawai and Akira, 2011). Recent work by Verrier et al. (2018) supports the idea that reverse transcription needs to be regulated to avoid activation of the immune system. Exposing liver cells to naked viral DNA activated an interferon response via cyclic GMP-AMP synthase (cGAS), which is not detected with viral pgRNA (Verrier et al., 2018).

Thus far, conservation was used to identify functionally especially important PSs. Conversely, the significance of the proposed ϕ double function was hypothesised to result in conservation of this PS in related viruses. I could, therefore, use the information about the *cis*-acting element to predict PSs in other *Hepadnaviridae* without relying on experiments first. In *Avihepadnaviruses* this required identification of ϕ itself first based on the assumptions that it is located close to DR1 and base-pairs with ε. A location for ϕ in DHBV was originally proposed by Tang and McLachlan (2002) that was later shown to be incorrect (Maguire and Loeb, 2010). Maguire and Loeb (2010) concluded that no such *cis*-acting element existed in DHBV. Being less locationally restricted in my search for ϕ, I identified a putative region in DHBV that is also conserved in other *Avihepadnaviruses*. As opposed to HBV the putative ϕ consists of two parts separated by a small stretch of sequence including DR2, with a small overlap between DR2 and the latter part of ϕ. Interestingly, deletion of DR2 was found to negatively affect minus-strand DNA synthesis in DHBV unlike HBV, where it had no effect on that stage of reverse transcription (Maguire and Loeb, 2010), further supporting the location of ϕ slightly overlapping with DR2.

To date, there is not direct experimental evidence for the proposed nucleation complex, the hypothesised ϕ double function, or the predicted ϕ and PSs in DHBV. Re-assembly assays of DHBV capsid protein with the predicted nucleation

complex PSs could indirectly confirm all of these hypotheses as they build on each other. It would also be of interest to specifically test the role of ϕ's double function in evading the immune system by mutation of the SL whilst preserving base-pairing with ε and ω through compensatory mutations, and measuring cGAS activity similarly to Verrier et al. (2018). A positive outcome would further support this nucleation complex as a promising HBV drug target.

In MS2, the most important PS, TR, named for its additional function in repressing translation, was thought to nucleate assembly (Hung et al., 1969; Ling et al., 1970; Beckett and Uhlenbeck, 1988; Beckett et al., 1988). Conservation was expected to reveal additional important PS contacts. By aligning the secondary structures of three representatives of *Leviviridae* species and annotating PSs I was able to identify a small set of PSs that were conserved between all three species in Chapter 6. Six of these were part of a group of PSs recently identified by Dai et al. (2017) to be in contact with MS2 CP. Whilst TR was known to be conserved (Hung et al., 1969; Ling et al., 1970), the conservation of the additional five PSs was new. Mapping their positions inside the capsid showed four of the conserved PSs in close proximity to each other and the maturation protein, which takes the place of one CP dimer in the capsid. Maturation protein plays an essential role in the viral life cycle by attaching to the F-pilus and facilitating entry into a new host cell (Crawford and Gesteland, 1964; Lodish et al., 1965), and is in contact with the viral RNA's 5'- and 3'-ends (Shiba and Suzuki, 1981; Rumnieks and Tars, 2017). Nucleating assembly from this protein would ensure that it is correctly packaged in each progeny virus, avoiding the generation of non-infectious particles. From this information it was hypothesised that capsid assembly in *Leviviridae* nucleates from five points: upstream and downstream of the 5' maturation protein contact, either side of TR, and upstream of the 3' maturation protein contact.

Dai et al. (2017) themselves noted that a large proportion of their resolved PSs were located close to TR and its three neighbours, which they propose to be

the nucleation site with no mention of involvement of maturation protein. This connection was made later with the observation that in the structure by Dai et al. (2017) one SL at the 3'-end interacts with both maturation protein and two CP dimers, leading to the proposition of a nucleation complex consisting of these components and TR and other PSs only coming in later (Tavares et al., 2018). Interestingly, Dai et al. (2017) identified several RNA contacts with maturation protein including at the 3'-end and towards the middle of the RNA but do not mention any 5' contact, which stands in contrast to previous experimental data (Shiba and Suzuki, 1981). My proposed large nucleation complex combines conservation with all the above insights to include 5'-end contacts according to older experiments (Shiba and Suzuki, 1981), 3'-end contacts and maturation protein as in Tavares et al. (2018), as well as TR and its neighbours indicated by Dai et al. (2017).

Recent assembly kinetics studies by Garmann et al. (2019) in MS2 further supported the presence of a small critical nucleation complex, which forms in the RNA. Whilst their data could not determine an exact size for this nucleus, they estimated that no more than six CP dimers would be involved (Garmann et al., 2019), whereas the large nucleus proposed here would require six CP dimers and maturation protein. Note, however, that their re-assembly experiments were performed in the presence of CPs and RNA only and thus lacked maturation protein, which I believe plays a crucial role in capsid assembly *in vivo*.

In silico models can provide insights into aspects of the process that would be difficult or tedious to test experimentally. Parameter can be varied quickly and easily giving access to a wide range of experimental conditions, which can be difficult to create in the laboratory. Furthermore, they can shed light on why slightly different experimental conditions result in very different outcomes. In the context of viral capsid assembly simple computational models involving only twelve CP units and PSs have shown the importance of a mixture of PS affinities on the RNA and a gradual increase of CP concentration for selective

and efficient packaging (Dykeman et al., 2013a, 2014). However, modelling MS2 assembly computationally had thus far not resulted in the expected yields when nucleation was assumed to occur at TR only, indicating that the process may be more complex than previously thought. The model is based on the stochastic simulation algorithm or Gillespie algorithm (Gillespie, 1977). Stochastic models are considered superior to deterministic models when considering reactions that are heavily impacted by few molecules such as nucleation (de Levie, 2009; McAdams and Arkin, 1999; Martinez-Urreaga et al., 2003; Freeman, 1984), which is of particular interest here. Incorporating the idea of a large nucleus into the computational model in Chapter 8, together with optimising the inter-dimer energies resulted in capsid yields comparable to those found in experiments. Having a larger nucleus and more genomic RNA ends to build from reduced the complexity of the assembly process even more (Dykeman et al., 2014). Thus far varied PS affinities have not been taken into account and only high affinity PSs were used, which had previously been found to reduce capsid yield by almost 40% (Dykeman et al., 2013a). Through parameter optimisation I found the optimal inter-dimer energies, which confirmed the higher stability of interactions between homo- and heterodimers proposed by ElSawy et al. (2010). This is thought to be due to several PSs, which interact with heterodimers, repelling each other, whereas a single one can stabilise the interactions between the two types of dimers (ElSawy et al., 2010). Previously, it was shown that MS2 assembly is highly constrained geometrically (Dykeman et al., 2013b; Stockley et al., 2013a; Geraets et al., 2015; Twarock et al., 2018; Twarock and Stockley, 2019). As the large nucleus builds up the capsid from different ends connections through the inside of the capsid occur, implying that RNA organisation in the capsid consists of several complementary Hamiltonian path fragments in line with work by Dai et al. (2017) and Twarock et al. (2018). However, naturally not any two points of the RNA can be connected over any distance through the capsid. Therefore, it requires that there is sufficient sequence between the points. These actual distances between

PSs have not been taken into account so far. Dykeman et al. (2013b) showed how the relative distance between two PSs can be utilised to predict how they connect in the capsid and thus which Hamiltonian path is followed. A similar analysis allowing for several partial paths and the constraint of the large nucleus could be carried out in the future. Moreover, it would also be interesting to apply the assembly model to other viruses such as HBV to investigate the impact of other constraints on the RNA.

Chapter 10

Conclusion

Viral diseases pose a large burden on society. HBV alone infects millions of people worldwide. With viruses ever-evolving and escaping current anti-viral treatments through mutation, the search continues for more suitable drug targets. Conservation can reveal features that would be more difficult for the virus to mutate while preserving fitness. PSs have only relatively recently been considered as drug targets. Finding conserved PSs, especially those performing additional functions, opens up the path to specifically exploit these. Before this study, HBV was thought to not have PSs. Not only have I proven this assumption to be wrong in collaboration with experimentalists in Leeds but I have also identified a pair of PSs that are highly conserved and therefore thought to be part of the nucleation complex. One of these may even serve the double function in regulating reverse transcription and therefore be an especially interesting drug target. Further experimental research is needed to validate these predications and, given a positive outcome, would be the basis for drug discovery. Whilst *Leviviridae* are bacteriophages and thus do not have direct medical significance, insights gained from simple systems can often be translated to more complex ones. Previously, it was thought that assembly nucleated from TR only but this assumption was not compatible with computational models, which failed to produce capsids under those conditions. Through conservation I was able

to find other PSs in close proximity in the capsid, which are hypothesised to form a larger nucleation complex together with the maturation protein. Only when this was incorporated did the model yield fully assembled capsids. This significantly changes the way assembly nucleation is understood in *Leviviridae* and may in fact be translatable to other viruses. In HBV both ends of the pregenomic RNA need to come into close proximity inside the capsid for reverse transcription indicating a requirement for this to happen during assembly. The newly found nucleation complex pair is at the 3'-end of the pregenomic RNA whereas ε, which was thought to be the only SL involved in packaging, is located at the 5'-end. Similarly to *Leviviridae* those SLs may form a larger nucleation complex together ensuring spatial proximity of the RNA ends for the next step in the viral life cycle. Further research would have to elucidate how these larger nuclei work and how common they are. Nevertheless, in HBV it presents an intriguing drug target, especially given the potential role in suppression of an innate immune response.

Appendix A

Appendix

ALGORITHM A.1: Main program SL_extraction.

```
1  SET total number of folds, prev to 0
2  SET number of structures to 1
3  READ fold, fragment
4  WHILE not end of fold file
5  |  INCREMENT total number of folds
6  |  SET basepair counter, basepair stack to 0
7  |  SET first, bulge to TRUE
8  |  FOR each nucleotide positions in fragment
9  |  |  IF Vienna fold = "(" THEN
10 |  |  |  IF bifurcation THEN
11 |  |  |  |  truncate bifurcation
12 |  |  |  |  CALL MULTIS for truncated fold
13 |  |  |  END IF
14 |  |  |  INCREMENT basepair counter
15 |  |  |  SET stack of basepair counter to current position
16 |  |  |  IF first SL THEN
17 |  |  |  |  SET structure start to current postion + fragment-1
18 |  |  |  |  SET first to FALSE
19 |  |  |  END IF
20 |  |  ELSE IF Vienna fold = ')' AND basepair counter /=0 THEN
21 |  |  |  DECREMENT basepair count
22 |  |  |  IF no further basepairs in stack THEN
23 |  |  |  |  SET loop length to current position - last basepair position -1
24 |  |  |  |  SET loop position to last basepair position + fragment
25 |  |  |  |  INCREMENT helix1 length
26 |  |  |  |  SET helix1 start and end positions
```

```
27 | | | | SET prev to current position
28 | | | | SKIP to next position
29 | | | ELSE IF bulge THEN
30 | | | | IF current /= next basepair position -1 OR current position /=
       prev +1 THEN
31 | | | | | SET 5' bulge to next - current basepair position -1
32 | | | | | SET 3' bulge to current position - prev-1
33 | | | | | SET bulge to FALSE
34 | | | | | INCREMENT next helix length
35 | | | | | SET next helix start and end positions
36 | | | | ELSE
37 | | | | | INCREMENT helix1 length
38 | | | | | SET helix1 start and end positions
39 | | | | END IF
40 | | | ELSE
41 | | | | INCREMENT next helix length
42 | | | | SET next helix start and end positions
43 | | | END IF
44 | | | SET prev to current position
45 | | END IF
46 | | IF (basepair counter = 0 OR end of structure) AND NOT first THEN
47 | | | SET structure start to last basepair position + fragment -1
48 | | | SET structure end to current position + fragment -1
49 | | | RESET basepair count, basepair stack, prev to 0
50 | | | RESET first, bulge to TRUE
51 | | | CALL MULTIS
52 | | END IF
53 | END LOOP
54 | READ next fold
55 END LOOP
56 WRITE processed folds
57 WRITE helix2, 5'bulge, helix1, loop, 3'bulge lengths, start, end, loop pos,
       stab, total number of folds
58 WRITE helices length, start and end positions
```

ALGORITHM A.2: Subroutine MULTIS.

```
1 IF helix2 = single basepair THEN
2 | IF helix1 = single basepair THEN
3 | | DELETE fold
4 | | RETURN
5 | ELSE
6 | | truncate fold after helix1
7 | END IF
8 END IF
```

```
FOR each helix in structure
|  IF helix = single basepair THEN
|  |  truncate fold after previous helix
|  |  EXIT LOOP
|  END IF
END LOOP
FOR each saved structure
|  IF all attributes new = saved structure THEN
|  |  INCREMENT number of folds for saved structure
|  |  EXIT SUBROUTINE
|  ELSE IF last saved structure THEN
|  |  INCREMENT number of structures
|  |  SAVE attributes of new structure)
|  |  SAVE complete Vienne fold of new structure
|  END IF
END LOOP
```

ALGORITHM A.3: Main program SL_merge.

```
READ fasta file
READ structure attributes
READ first structure
SAVE structure
WHILE not end of structure file
|  READ next structure
|  IF not a fold THEN
|  |  SKIP LOOP
|  END IF
|  READ structure attributes
|  WHILE start position of saved structure >= new structure
|  |  IF all attributes new = saved structure THEN
|  |  |  IF number of folds new > saved structure THEN
|  |  |  |  SAVE number of folds of new structure
|  |  |  END IF
|  |  |  EXIT LOOP
|  |  END IF
|  END LOOP
|  IF start position of saved structure < new structure THEN
|  |  SAVE attributes of new structure
|  |  SAVE structure
|  END IF
END LOOP
WRITE saved structures
WRITE structure attributes
WRITE structure sequences
```

```
27 WRITE apical loop sequences
28 WRITE helices
```

ALGORITHM A.4: Recursive Weighted-Activity Selection.

```
1  n = total # stemloops
2  WAS_COMP(SL,n,Energies,CompatibleSLs,addedEnergies)
3  {
4    IF ( SL = 0 ) THEN
5    |  addedEnergies(0) = 0
6    ELSE IF ( addedEnergies(SL) = "empty" ) THEN
7    |  WAS_COMP(CompatibleSLs(SL),n,Energies,CompatibleSLs,addedEnergies)
8    |  WAS_COMP(SL-1,n,Energies,CompatibleSLs,addedEnergies)
9    |  addedEnergies(SL) = MAX{ Energies(SL) +
         addedEnergies(CompatibleSLs(SL)), addedEnergies(SL-1) }
10   END IF
11 }
12 SOL(SL,n,Energies,CompatibleSLs,addedEnergies,Selected)
13 {
14   IF (SL = 0 ) THEN
15   |  Selected(0) = 0
16   ELSE IF ( Energies(SL) + addedEnergies(CompatibleSLs(SL)) >
         addedEnergies(SL-1) ) THEN
17   |  Selected(SL) = 1
18   |  SOL(CompatibleSLs(SL),n,Energies,CompatibleSLs,addedEnergies,Selected)
19   ELSE
20   |  SOL(SL-1,n,Energies,CompatibleSLs,addedEnergies,Selected)
21   END IF
22 }
```

ALGORITHM A.5: Main program: PS_profiles.

```
1  SET end, c, g, u to 0
2  FOR each selected SL
3  |  SET PSprofile from end+1 to startSL-1 to 'A'
4  |  SET end to endSL
5  |  IF SL = high affinity PS THEN
6  |  |  SET PSprofile from startSL to end to 'C'
7  |  |  INCREMENT c by 1
8  |  ELSE IF SL = medium affinity PS THEN
9  |  |  SET PSprofile from startSL to end to 'G'
10 |  |  INCREMENT g by 1
11 |  ELSE IF SL = low affinity PS THEN
12 |  |  SET PSprofile from startSL to end to 'U'
13 |  |  INCREMENT u by 1
```

```
| ELSE
| | SET PS profile from startSL to end to A'
| END IF
END LOOP
```

Algorithm A.6: Main program: PS_align.

```
FOR each sequence
| SET k to 0
| FOR each nucleotide
| | INCREMENT k by 1
| | IF alignment of sequence at nucleotide = '-' THEN
| | | IF first nucleotide THEN
| | | | SET PSprofile_num at nucleotide to 0
| | | | SET PSprofile_lett at nucleotide to '-'
| | | ELSE
| | | | SET PSprofile_num at nucleotide to value at previous position
| | | | SET PSprofile_lett at nucleotide to value at previous position
| | | END IF
| | | DECREMENT k by 1
| | ELSE IF PSprofile at position k = 'A' THEN
| | | SET PSprofile_num at nucleotide to 0
| | | SET PSprofile_lett at nucleotide to 'A'
| | ELSE IF PSprofile at position k = 'C', 'G' or 'U' THEN
| | | SET PSprofile_num at nucleotide to 1
| | | SET PSprofile_lett at nucleotide to PSprofile at position k
| | ELSE
| | | PRINT "Something is wrong", PSprofile at position
| | END IF
| END LOOP
END LOOP
```

Algorithm A.7: Main program: PS_BLOCKS.

```
SET new to TRUE
SET BLOCKN to 0
FOR each nucleotide position
| IF SUM of PSprofile_num at nucleotide >= threshold THEN
| | IF new THEN
| | | IF BLOCKN not 0 THEN
| | | | CALCULATE block length
| | | | CALCULATE number of splits as block/SL length rounded to integer
| | | | IF number of splits < 1 THEN
| | | | | SET BLOCKS_N of BLOCKN to 0
| | | | | DECREMENT BLOCKN by 1
```

```
|  |  |  |  ELSE IF number of splits > 1 THEN
|  |  |  |  |  SET ENN to end of current block
|  |  |  |  |  SET end of current block to start + SL length - 1
|  |  |  |  |  WHILE end of block < enn
|  |  |  |  |  |  INCREMENT BLOCKN by 1
|  |  |  |  |  |  SET new block start to previous end + 1
|  |  |  |  |  |  SET new block end to start + SL length - 1
|  |  |  |  |  END LOOP
|  |  |  |  |  SET new block start to previous end + 1
|  |  |  |  |  SET new block end to ENN
|  |  |  |  END IF
|  |  |  END IF
|  |  |  INCREMENT BLOCKN by 1
|  |  |  SET current block start to nucleotide
|  |  |  SET new to FALSE
|  |  END IF
|  |  SET current block end to nucleotide
|  ELSE
|    SET new to TRUE
|  END IF
END LOOP
CALCULATE block length
CALCULATE number of splits as block/SL length rounded to integer
IF number of splits < 1 THEN
|  SET BLOCKS_N of BLOCKN to 0
|  DECREMENT BLOCKN by 1
ELSE IF number of splits > 1 THEN
|  SET ENN to end of current block
|  SET end of current block to start + SL length - 1
|  WHILE end of block < enn
|  |  INCREMENT BLOCKN by 1
|  |  SET new block start to previous end + 1
|  |  SET new block end to start + SL length - 1
|  END LOOP
|  SET new block start to previous end + 1
|  SET new block end to ENN
END IF
```

ALGORITHM A.8: Main program: BLOCK_MEM.

```
SET double to FALSE
FOR each sequence
|  FOR each block
|  |  IF SUM of PSprofile_num between block start and end > 0 THEN
|  |  |  IF double THEN
```

```
6  |  |  |  |  SET double to FALSE
7  |  |  |  |  SKIP to next block
8  |  |  |  END IF
9  |  |  |  IF PS continuous between current and next block THEN
10 |  |  |  |  IF gap in PS profile in next block THEN (3)
11 |  |  |  |  |  CALL AFF_ASIGN with start and end of current block
12 |  |  |  |  |  CALL AFF_ASIGN with start of next PS and end of next block
       for next block
13 |  |  |  |  |  SET double to FALSE
14 |  |  |  |  |  GOTO 1
15 |  |  |  |  END IF
16 |  |  |  |  IF gap in PS profile in current block THEN
17 |  |  |  |  |  CALL AFF_ASIGN with start of current block and end of PS for
       current block
18 |  |  |  |  |  CALL AFF_ASIGN with start and end of next block
19 |  |  |  |  |  SET double to FALSE
20 |  |  |  |  |  GOTO 1
21 |  |  |  |  END IF
22 |  |  |  |  IF more than 50% of block spanned by PS AND more than 50% of
       next block spanned by PS AND start current to end next block longer
       than SL length THEN
23 |  |  |  |  |  CALL AFF_ASIGN with start and end of current block
24 |  |  |  |  |  CALL AFF_ASIGN with start and end of next block
25 |  |  |  |  |  SET double to TRUE
26 |  |  |  |  ELSE IF more PS in current block than next block THEN (3)
27 |  |  |  |  |  CALL AFF_ASIGN with start and end of current block
28 |  |  |  |  |  SET next block to 'A'
29 |  |  |  |  |  SET double to TRUE
30 |  |  |  |  ELSE
31 |  |  |  |  |  CALL AFF_ASIGN with start and end of next block
32 |  |  |  |  |  SET current block to 'A'
33 |  |  |  |  |  SET double to TRUE
34 |  |  |  |  END IF
35 |  |  |  ELSE
36 |  |  |  |  CALL AFF_ASIGN with start and end of current block
37 |  |  |  |  SET double to FALSE
38 |1 |  |  END IF
39 |  |  ELSE
40 |  |  |  SET current block to 'A'
41 |  |  END IF
42 |  END LOOP
43 END LOOP
44 SET informative_characters to 0
45 FOR each block
```

```
|  FOR each sequence
|  |  IF block current sequence /= block next sequence THEN
|  |  |  INCREMENT informative_characters by one
|  |  |  EXIT LOOP
|  |  END IF
|  END LOOP
END LOOP
```

Algorithm A.9: Subroutine AFF_ASIGN.

```
SET max, num_C, num_G, and num_U to 0
FOR each position in block
|  IF pseudonucleotide at position = C THEN
|  |  INCREMENT num_C
|  |  IF num_C > max THEN
|  |  |  SET max to num_C
|  |  |  SET block to C
|  |  END IF
|  ELSE IF pseudonucleotide at position = G THEN
|  |  INCREMENT num_G
|  |  IF num_G > max THEN
|  |  |  SET max to num_G
|  |  |  SET block to G
|  |  END IF
|  ELSE IF pseudonucleotide at position = U THEN
|  |  INCREMENT num_U
|  |  IF num_U > max THEN
|  |  |  SET max to num_U
|  |  |  SET block to U
|  |  END IF
|  END IF
END LOOP
RETURN
```

TABLE A.1: MS2 stem-loops predicted by different algorithms. Ticks mark SLs predicted by the respective algorithm, whereas swung dashes represent similar structures with only minor changes, and crosses mean that no such SL was found by the algorithm.

SL	Mfold	win 30	win 60	win 90
A1	✓	✓	✓	✓
A2a	✗	✓	✓	✓
A2b	✗	✓	✓	✓
A2c	✗	✓	✓	✓
A3a	✓	✓	✓	✓
A3b	✓	✗	✓	✓
A3c	∼	∼	∼	∼
A3d	✗	✓	✗	✗
A4a	✓	✓	✓	✓
A4b	✓	✓	✓	✓
A4c	∼	∼	∼	∼
A5	✓	✓	✓	✓
A6	✓	✓	✓	✓
A7	✓	✓	✓	✓
A8a	✗	✓	✓	✓
A8b	✗	✓	✓	✓
A8c	✓	✓	✓	✓
A9a	✓	✓	✓	✓
A9b	✓	✗	✗	✗
A10	✓	✓	✓	✓
A11a	∼	∼	∼	∼
A11b	✗	✓	✓	✓
A12a	✓	✓	✓	✓
A12b	✓	✓	✓	✓
A13	✗	✓	✓	✓
C1	✓	✓	✓	✓
C2	✗	✓	✓	✓
C3a	✗	✓	✓	✓
C3b	✗	✓	✓	✓
C4	✓	✓	✓	✓
C5	✓	✓	✓	✓
C6	✓	✓	✓	✓
C7	✓	✓	✓	✓
C8	✓	✓	✓	✓

TABLE A.1: (continued)

R33	✓	✓	✓	✓
R32	✗	✓	✓	✓
R30/31	✓	✓	✓	✓
R29	✓	✓	✓	✓
R27/28	✓	✓	✓	✓
R26	✓	✓	✓	✓
R25	✓	✓	✓	✓
R24	✓	✓	✓	✓
R22/23	∼	∼	∼	∼
R20/21	✓	✓	✓	✓
R19	✓	✓	✓	✓
R18b	✗	✓	✓	✓
R18	✓	✓	✓	✓
R17	✓	✓	✓	✓
R15/16	✓	✓	✗*	✗*
R14	✓	✓	✓	✓
R13	✓	✓	✓	✓
R12	✓	✓	✓	✓
R11	✗	∼	∼	∼
R10c	✗	✗	✗	✗
R10b	✓	✓	✓	✓
R10a	✗	✓	✓	✓
R9	✓	✓	✓	✓
R8	✓	✓	✓	✓
R7	∼	∼	∼	∼
R6b	✗	✗	✓	✓
R6a	✓	✓	✓	✓
R5	✓	✓	✓	✓
R4	✓	✓	✓	✓
R3	∼	∼	∼	∼
R2	✓	✓	✓	✓
R1	✓	✓	✓	✓
U3	✓	✓	✓	✓
U2	✓	✓	✓	✓
U6	✓	✓	✓	✓
U5	✓	✓	✓	✓
U4	✓	✓	✓	✓

TABLE A.1: (continued)

V2	✗	✗	✗	✗
V1	∼	✓	✓	✓
U1	✓	✓	✓	✓

* makes a high affinity PS instead.

TABLE A.2: Summary of published Hepatitis B virus recombinant isolates.

Major type[1]	**Minor type**	**Break points**[2]	**Number of isolates**[3]	**Countries of isolation**	**References**	**Accession numbers**
A	A	1520, 2474	1	India	Ye et al. (2010)	AY161139
A	C	1730, 1934	1	South Africa	Shi et al. (2012a)	AY233277
A	D	1576, 2339	1	Senegal	Bowyer and Sim (2000)	X75664
A	D	2820–2895, 327–386	2	Africa (black)	Owiredu et al. (2001b); Simmonds and Midgley (2005); Yang et al. (2006)	AF297619-20

[1]The majority of the genome stems from this genotype (subgenotype).
[2]Beginning, end of inserted fragment. Ranges given when exact points not determined.
[3]Maximum number analysed in a single publication.

Table A.2: (continued)

A	D	1808, 2354	8	India	Simmonds and Midgley (2005); Yang et al. (2006); Ye et al. (2010); Shi et al. (2012a)	AF418674–75; AF418682–83; AY161140–41; AY161145–46
A	D	781, 1095; 2854, 90	4	India	Yang et al. (2006)	AF418690–92; AY161147
A	D	209, 526; 781, 1095; 3000, 3164	8	India	Yang et al. (2006)	AF418684–89; AY161148–49
A	D	741, 1472	1	Uzbekistan	Ye et al. (2010)	AB222708
A	E	882, 1060	1	Cameroon	Kurbanov et al. (2005); Simmonds and Midgley (2005); Yang et al. (2006)	AB194949

Table A.2: (continued)

A	E	unknown	2	Cameroon	Olinger et al. (2006)	unknown
A	E	1896–1906, 2419–2423	4	Guinea	Garmiri et al. (2009)	GQ161767; GQ161837–38; GQ161788
A	E	2120, 2419	1	Ghana	Garmiri et al. (2009)	GQ161753
A	F	2390, 85	1	Uruguay	Lopez et al. (2015)	KJ586810
A(2)	F(1b)	227, 1593	1	Uruguay	Lopez et al. (2015)	KJ586803
B	A	2014 ,2203	4	Japan	Bollyky et al. (1996)	D00329
B	A,C	1185, 1784 (A); 1784, 2401 (C)	1	Philippines	Yang et al. (2006)	AB219430

Table A.2: (continued)

B	C	1740–1838, 2443–2485	157	Cambodia, China, Hong Kong, Indonesia, Japan, Philippines, Switzerland, Thailand, Taiwan, USA [Asian ethnicity, not Japanese], Vietnam	Morozov et al. (2000); Bowyer and Sim (2000); Fares and Holmes (2002); Sugauchi et al. (2002, 2003); Luo et al. (2004); Ye et al. (2010); Simmonds and Midgley (2005); Yang et al. (2006); Shi et al. (2012a)	AB033554–55; AB031266–67; AB073821–37; AB073839–41; AB100695; AB115551; AB117759; AB205119–20; AB205122; AB219426–30; AB212625–26; AB246339–40; AF100308–09; AF121243–51; AF282917–18; AF479684; AF461360; AF461362; AJ31123; AY033072–73; AJ131133; AY163869–70; AY167089; AY167093–94; AY167097–102; AY206373; AY206375; AY206377; AY206380; AY206383; AY206390–91; AY217355–70; AY220697–98; AY220703–04; AY293309; AY518556; AY596102–12; AY766463; AY800389–92; D00330–31; DQ377158; DQ448620; DQ448623; DQ448625; DQ448627–28; DQ904357; DQ975271; M54923; X97850–51; X98072–75; X98077

TABLE A.2: (continued)

B	C	1880, 2260	1	Vietnam	Huy et al. (2003)	AB100695
B	C	3060–3191, unknown	1	Taiwan	Chen et al. (2004)	unknown
B	C	2910–2950, unknown	2	Taiwan	Chen et al. (2004)	unknown
B	C	1846, 2188	2	China	Luo et al. (2004)	AY217359–60
B	C	1792, 2599	1	China	Luo et al. (2004)	AY217365
B	C	1793, 2189	1	China	Luo et al. (2004)	AY217369
B	C		2	Philippines	Sakamoto et al. (2006)	AB241116–17
B	C	1859, 2294	1	Malaysia	Shi et al. (2012a)	GQ924624
B	C	1832, 2401	2	Indonesia	Shi et al. (2012a)	AB493827; AP011094
B	C	1661, 2267	1	China	Shi et al. (2012a)	EU158262
B	C	1229, 2274	1	China	Shi et al. (2012a)	EU939627

Table A.2: (continued)

B	C	1089, 1259; 1862, 2876	1	Taiwan	Shi et al. (2012a)	EU660227
B	C	1073, 1580; 1764, 2274	1	China	Shi et al. (2012a)	GQ377595
B	C	440, 658; 1657, 2272	1	China	Shi et al. (2012a)	FJ386648
B	C	493, 1068; 1873, 2188	1	China	Shi et al. (2012a)	FJ386674
B	C	225, 482; 1842, 2256	2	China, Indonesia	Shi et al. (2012a)	AB493831; EU939634
B	C	224, 635; 1842, 2266	1	China	Shi et al. (2012a)	HQ684848
B	C	282, 1051; 1565, 2253	1	China	Shi et al. (2012a)	GQ377592
B	C	1729, 3213	1	China	Shi et al. (2012a)	EU939629

Table A.2: (continued)

B	C,A	1740, 2443 (C); 2965, 3215 (A)	1	South Africa	Yang et al. (2006)	U87747
B	U,C	729, 1795 (U); 1795, 2443 (C)	1	Vietnam	Yang et al. (2006)	AB231909
C	A	2865, 1801	2	Vietnam	Hannoun et al. (2000)	unknown
C	A	661, 1831	1	Taiwan	Shi et al. (2012a)	EF494378
C	A	1881, 2775	1	Taiwan	Shi et al. (2012a)	EF494376
C	A, G	396, 666; 872, 1104 (G); 1, 396 (A)	4	Vietnam	Huy et al. (2008)	AB231908; AF241407–09

Table A.2: (continued)

C	B	1731–1838, 2444–2485	1	Japan	Morozov et al. (2000); Simmonds and Midgley (2005); Yang et al. (2006)	D16665
C	B	3129–3171, unknown	1	Taiwan	Chen et al. (2004)	unknown
C	B	1289, 1732	1	China	Luo et al. (2004)	AF233236
C	B	1244, 1799	1	China	Simmonds and Midgley (2005); Yang et al. (2006)	AF233236
C	B	1071, 1644	1	Vietnam	Ye et al. (2010)	AB031265

Table A.2: (continued)

C	B	2276, 224	16	China	Shi et al. (2012a)	EU939620–21; EU939623; EU939630; EU939632; FJ562247; FJ562328; GQ377549; GQ377564; GQ377573; GQ377596; GQ377604; GQ377613–14; GQ377626; GQ377634
C	B	2276, 3213	8	China	Shi et al. (2012a)	EU939622; FJ562229; GQ377539; GQ377556; GQ377565; GQ377590; GQ377594; GQ377602
C	B	255, 1741	4	China	Shi et al. (2012a)	EU939628; EU939631; GQ377630; GQ377635
C	B	126, 598; 1272, 1829	1	China	Shi et al. (2012a)	HQ684849
C	B	164, 388	1	China	Shi et al. (2012a)	GU357843
C	B	388, 886	1	Taiwan	Shi et al. (2012a)	EU882006
C	B	526, 833	1	Philippines	Shi et al. (2012a)	AB241109

Table A.2: (continued)

C	B	780, 1832	1	China	Shi et al. (2012a)	FJ386646
C	B	2200, 2681	1	China	Shi et al. (2012a)	FJ032343
C	B	2510, 2773	1	China	Shi et al. (2012a)	EU796069
C	B	2842, 3213	4	Taiwan	Shi et al. (2012a)	EU522070; EU660228; EU881995; EU919166
C	unknown	2865, 1801	3	Vietnam	Hannoun et al. (2000); Yang et al. (2006)	AB231908; AF241407–09
C(12)	G	2900, 3100	1	Indonesia	Mulyanto et al. (2012)	AB644285
C(13)	B(3)	2700, 2900	1	Indonesia	Mulyanto et al. (2012)	AB644282

TABLE A.2: (continued)

C(2)	D	0–10, 750–799	244	China	Simmonds and Midgley (2005); Zeng et al. (2005); Wang et al. (2005); Yang et al. (2006); Wang et al. (2007); Zhou et al. (2011); Shi et al. (2012a)	AF461043; AY817509–10; AY817512–15; AY800249; DQ478881–83; DQ478886–89; DQ478891–98; JF491447–56

Table A.2: (continued)

C(2)	D	10–50, 1450–1499	80	Tibet, China	Cui et al. (2002); Luo et al. (2004); Wang et al. (2005, 2007); Simmonds and Midgley (2005); Yang et al. (2006); Zhou et al. (2011); Shi et al. (2012a)	AY057948; AY817511; AY657948; DQ478890; HM750142-50
D	A	735, 2370	1	Italy	Bollyky et al. (1996); Simmonds and Midgley (2005)	X68292

TABLE A.2: (continued)

D	A	451–493, 735–779; 1605–1631, 1996–2017	1	Italy	Morozov et al. (2000); Bowyer and Sim (2000); Fares and Holmes (2002); Simmonds and Midgley (2005); Yang et al. (2006)	X65258
D	A	791–820, 1763–1996	1	Italy	Morozov et al. (2000); Bowyer and Sim (2000); Simmonds and Midgley (2005); Yang et al. (2006)	X65259

TABLE A.2: (continued)

D	A	657–735, 1167–1306; 1356,2096–2150; 2186–2190, 2359	1	Italy	Morozov et al. (2000); Bowyer and Sim (2000); Fares and Holmes (2002); Simmonds and Midgley (2005); Yang et al. (2006)	X68292
D	A	500, 1800	1	Italy	Simmonds and Midgley (2005); Yang et al. (2006)	AY236161
D	A	400, 700	1	India	Simmonds and Midgley (2005); Yang et al. (2006)	AF418681, AY161161
D	A	0, 595–618	1	India	Chauhan et al. (2008)	EF103284

Table A.2: (continued)

D	A	0, 639–659	1	India	Chauhan et al. (2008)	EF103283
D	A	319–359, 1170–1184	1	India	Chauhan et al. (2008)	EF103282
D	A	641, 926	1	Uzbekistan	Shi et al. (2012a)	AB188244
D	A	1730, 1944	1	Tunisia	Shi et al. (2012a)	FJ904409
D	C	194, 1702	3	China	Shi et al. (2012a)	FJ562223; GQ377532; GQ377627
D	C	3058, 3225	1	China	Shi et al. (2012a)	FJ562309
D	E	1600–1900, 2325–2360	1	Ireland	Laoi and Crowley (2008)	DQ991753
D	E	1651,2406; 2823,3081	1	Ghana	Garmiri et al. (2009)	GQ161754

Table A.2: (continued)

D	E	1640–1649, 2392–4; 2831–9, 3075–83	2	Niger	Chekaraou et al. (2010)	FN594769; FN594771
D	E	1932–63, 2385–2431; 2836–2864, 3083–3128	2	Niger	Chekaraou et al. (2010)	FN594768; FN594770
D(3)	F(1b)	2388, 2867	1	Uruguay	Lopez et al. (2015)	KJ586811
E	A	1287, 1896	1	Guinea	Garmiri et al. (2009)	GQ161775
E	A	1542, 2304	1	Guinea	Garmiri et al. (2009)	GQ161806
E	D	98, 438; 778, 1273	1	Niger	Chekaraou et al. (2010)	FN594767

Table A.2: (continued)

F	C	1558, 1844	1	Bolivia	Simmonds and Midgley (2005); Yang et al. (2006); Ye et al. (2010)	AB214516
F	G	1075, 1256	1	France	Fallot et al. (2012)	unknown
F(3)	A(1)	941 (incomplete)	1	Colombia (black)	Alvarado-Mora et al. (2012)	unknown
F(4)	G	1845, 2132	2	Brazil	Araujo et al. (2013)	HE981177 –78
F(4)	G	493, 1816	1	Brazil	Araujo et al. (2013)	HE981179
G	A	17, 217	2	USA	Kato et al. (2002); Simmonds and Midgley (2005); Yang et al. (2006)	AB056516

Table A.2: (continued)

G	A	unknown	2	Canada	Osiowy et al. (2008)	unknown
G	A	1, 392	1	Brazil	Shi et al. (2012a)	EF464099
G	C	1860, 2460	1	Thai	Suwannakarn (2005); Yang et al. (2006)	DQ078791
G	F(1b)	1824, 2154	1	Argentina	Araujo et al. (2013)	HE981188
G	F(4)	1817, 2442	1	Brazil	Araujo et al. (2013)	HE981180

TABLE A.3: Accession numbers for the HBV (sub)genotypes used.

Alias	Accession number	Origin
A1	AY233288	South Africa
A2	AY233286	South Africa
AD18082354_1	AY161140	India
AD18082354_2	AY161141	India
B1_2	AB010289	Japan
B1	D00329	Japan
B2	AF282918	China
B3	M54923	Indonesia
B4	AB117759	Cambodia
B5	DQ463801	Canada
BC17402485_58	AB246340	Japan
BC17402485_59	AY217356	China
C10	AB540583	Indonesia
C11	AB554020	Indonesia
C12	AB560662	Indonesia
C13	AB644280	Indonesia
C14	GQ377555	Indonesia
C15	AB644286	Indonesia
C16	AB644287	Indonesia
C1	AB112472	Thailand
C2	AY066028	China
C3	X75656	Polynesia
C4	AB048704	Australia
C5	AB241110	Phillippines
C6	AB493842	Indonesia
C7	EU670263	Phillippines
C8	AP011104	Indonesia
C9	AP011108	Indonesia
CD101499_3	DQ478890	Tibet
CD101499_4	AY057948	Tibet
CD10799_1	AF461043	China
CD10799_2	AY817509	China
D1	FJ899792	China
D2	GU456635	Iran
D3	EU594435	Estonia
D4	AB033559	Papua

Table A.3: (continued)

D5	AB033558	Japan
D6	FJ904433	Tunesia
E	X75657	Western Africa
F1	AY090459	Costa Rica
F2	AY090455	Nicaragua
F4	AF223965	Argentina
G_2	AB064310	Japan
G	AB056513	USA

TABLE A.4: Accession numbers for the HBV genomes used in Michitaka et al. (2006) study.

Alias	Accession number
Ehime_D_1	AB090268
Ehime_D_2	AB090269
Ehime_D_3	AB078031
Ehime_D_4	AB078032
Ehime_D_5	AB078033
Ehime_D_6	AB090270
Ehime_D_7	AB109475
Ehime_D_8	AB109476
Ehime_D_9	AB109477
Ehime_D_10	AB109478
Ehime_D_11	AB109479
Ehime_D_12	AB110075
Ehime_D_13	AB119251
Ehime_D_14	AB119252
Ehime_D_15	AB119253
Ehime_D_16	AB119254
Ehime_D_17	AB119255
Ehime_D_18	AB119256
Ehime_D_19	AB116266
Ehime_D_20	AB120308
Aa	AF297622
Ae	AB014370
Ba	AB073821
Bj	D00329
C	AB042283
D_China	AF280817
D_Egypt_1	AB104709
D_Egypt_2	AB104710
D_Egypt_3	AB104711
D_Egypt_4	AB104712
D_France	M32138
D_Germany_1	AF043593
D_Germany_2	AF043594
D_Germany_3	X72702
D_Germany_4	Y07587

TABLE A.4: (continued)

D_Germany_5	AJ131956
D_Germany_6	AF151735
D_Italy_1	AB188245
D_Italy_2	X65257
D_Italy_3	X65258
D_Italy_4	X85254
D_Japan	AB033558
D_Kamchatka_1	AB188241
D_Kamchatka_2	AB188242
D_Kamchatka_3	AB188243
D_Pap	AB033559
D_Poland	Z35716
D_Sweden_1	AF121239
D_Sweden_2	AF121240
D_Sweden_3	AF121242
D_Sweden_4	AY090453
D_UK_1	X80924
D_UK_2	X80925
D_UK_3	X97848
D_UK_4	X97849
D_USA	L27106
D_Uzbekistan	AB188244
E	X75657
F	X75658
G	AF160501
H	AY090454

TABLE A.5: Accession numbers for the HBV genomes used in Sede et al. (2014) study.

Alias	Accession number
Mother_a	KF584158
Mother_b	KF584159
Mother_c	KF584160
Daughter_a	KF584161
Daughter_b	KF584162
Son1_a	KF584163
Son1_b	KF584164
Son2_a	KF584165
Son2_b	KF584166
A_1	X02763
A_2	X51970
A_3	AF090842
A_4	AB241115
B_1	D00329
B_2	D00331
B_3	D23677
B_4	AF100309
C_1	X04615
C_2	AY123041
C_3	DQ089803
C_4	DQ478901
D1_1	AJ344116
D1_2	AB104710
D1_3	AB222709
D1_4	JF754590
D2_1	Z35716
D2_2	X80925
D2_3	AB109475
D3_1	X65257
D3_2	AY233291
D3_3	AY233295
D4_1	AB033559
D4_2	AB048702
D4_3	AJ627219
D5_1	AB033558

TABLE A.5: (continued)

D5_2	DQ315779
D7_1	FJ904406
D7_2	FJ904419
E_1	AB032431
E_2	AB205192
E_3	DQ060823
F_1	X69798
F_2	AB036906
F_3	AB036910
F_4	AF223965
G_1	AF160501
G_2	AB056514
G_3	AB064310
G_4	AF405706
H_1	AY090454
H_2	AY090460
H_3	AB266536
H_4	AB298362
I_1	AF241409
I_2	AB231908
J	AB486012

TABLE A.6: Accession numbers for the HBV genomes used in Osiowy et al. (2006) study.

Alias	Accession number
Patient1_1979	DQ463787
Patient2_1979	DQ463788
Patient3_1979	DQ463789
Patient4_1979	DQ463790
Patient5_1979	DQ463791
Patient6_1979	DQ463792
Patient7_1979	DQ463793
Patient8_1979	DQ463794
Patient1_2004	DQ463795
Patient2_2004	DQ463796
Patient3_2004	DQ463797
Patient4_2004	DQ463798
Patient5_2004	DQ463799
Patient6_2004	DQ463800
Patient7_2004	DQ463801
Patient8_2004	DQ463802
A	AB064314
B1_1	D23678
B1_2	AB010289
B2	AF121251
B3	D00331
C	AB033556
D	X02496
E	X75657
F	X75663
G	AF160501
H	AY090460

TABLE A.7: Accession numbers for the HBV genomes used in Osiowy et al. (2010) study.

Alias	Accession number
Isolate1 99	FJ882616
Isolate2 01	FJ882610
Isolate3 02	FJ882614
Isolate4 03	FJ882611
Isolate5 04	FJ882615
Isolate6 05	FJ882613
Isolate7 07	FJ882612
Isolate8 07	EU833891
Isolate9 08	FJ882618
Isolate10 08	FJ882617
I_1 Vietnam	AF241407
I_2 Vietnam	AF241408
I_3 Vietnam	AF241409
I_4 Vietnam	AB231908
I_1 Laos	FJ023660
I_2 Laos	FJ023661
I_3 Laos	FJ023662
I_4 Laos	FJ023663
I_5 Laos	FJ023664
I_6 Laos	FJ023665
I_7 Laos	FJ023666
I_8 Laos	FJ023667
I_9 Laos	FJ023668
I_10 Laos	FJ023669
I_11 Laos	FJ023670
I_12 Laos	FJ023671
I_13 Laos	FJ023672
I_14 Laos	FJ023673
I_15 Laos	FJ023674
I_16 Laos	FJ023675
I_17 Laos	FJ023676
A1_1	AY233279
A1_2	AB116087
A2_1	AY233286
A2_2	X02763

TABLE A.7: (continued)

A3_1	AB194949
A3_2	AB194951
B1	D23678
B2	D00330
C1	AB033556
C1_2	AF182802
C2_1	AF223954
C2_2	AB111946
C3	X75656
C4	AB048704
D1	AF151735
D2	X02496
E	X75664
F	AF223962
G_1	AF160501
G_2	DQ207798
H	AY090460

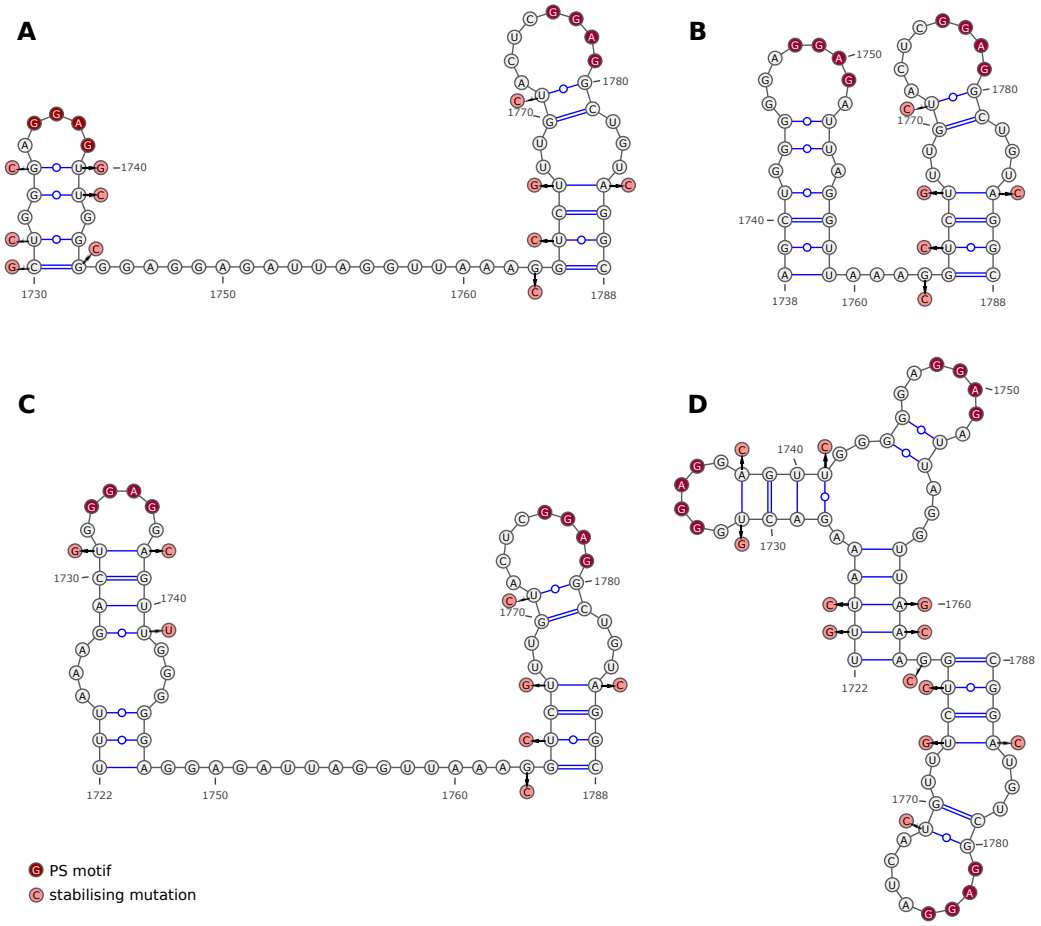

FIGURE A.1: **Nucleation complex structures.** The original RNA sequences in the laboratory strain (NC_003977.1) are shown with the RGAG PS motif highlighted in dark red. Mutations to stabilise the respective structure are suggested (light red) (A) The wildtype sequence folds into this structure with ΔG=-3.2 kcal/mol whilst the MFE structure for this fragment was at ΔG=-12.6 kcal/-mol. Stabilised with mutations the energy could be decreased to -15.1 kcal/mol. (B) This structure had a ΔG of -4.2 kcal/mol whilst fhe MFE structure for this fragment had a ΔG of -10 kcal/mol. Stabilising the structure with the suggested mutations decreased the ΔG to -11.6 kcal/mol. (C) The structure in wildtype sequence had a ΔG of -5.5 kcal/mol compared to the MFE structure at -9.6 kcal/-mol. The stabilised structure folded with ΔG of -14 kcal/mol. (D) The sequence is the same as in C and folds into this structure with ΔG of -4.5 kcal/mol. The suggested mutations decreased this to -15.6 kcal/mol.

Pseudocode for calculating the number of RGAGs in apical loop sequences:

ALGORITHM A.10: Main program RGAG_COUNT.

```
1  FOR each sequence
2  |  SET hit to 0
3  |  CALL FIND_MOTIF for window size 40
4  |  CALL FIND_MOTIF for window size 50
5  |  CALL FIND_MOTIF for window size 60
6  |  PRINT hit
7  END LOOP
```

ALGORITHM A.11: Subroutine FIND_MOTIF.

```
1  FOR each window
2  |  PRINT fragment to file
3  |  CALL Tfold for fragment with partition function
4  |  FOR each sampled structure
5  |  |  FOR each 4-tuple in apical loop
6  |  |  |  IF tuple = 'GGAG' OR 'AGAG' THEN
7  |  |  |  |  INCREMENT hit
8  |  |  |  |  GOTO 1
9  |  |  |  END IF
10 |  |  END LOOP
11 |1 END LOOP
12 END LOOP
```

Pseudocode for introducing random synonymous mutations:

Algorithm A.12: Main program MUTATE.

```
READ attributes with indexes
IF constraints file THEN
|  READ constraints
ELSE
|  SET constraints to 0
END IF
SORT attributes ascending
CONVERT T to U in sequence
READ converted sequence
SAVE sequences with sorted attributes 1-400
FOR i=1 to 1600
|  SELECT two saved sequences randomly
|  SELECT frag combination randomly
|  FOR each non-overlapping 18 nt fragment
|  |  CONCATENATE new seq + current frag of randomly selected saved sequence
|  END LOOP
|  CALL SYN_MUT for new seq
END LOOP
PRINT saved sequences
```

Algorithm A.13: Subroutine SYN_MUT.

```
SET cseq to list of codons
SET pseq to list of amino acids
IF translation empty THEN
|  TRANSLATE nucleotide seq to amino acids
END IF
FOR each amino acid
|  SET r to random(iseed)
|  IF r > 0.01 THEN
|  |  SET codon to original
|  |  SKIP to next amino acid
|  END IF
|  IF amino acid = 'F' THEN n = 2
|  IF amino acid = 'L' THEN n = 6
|  IF amino acid = 'I' THEN n = 3
|  IF amino acid = 'M' THEN n = 1
|  IF amino acid = 'V' THEN n = 4
|  IF amino acid = 'S' THEN n = 6
|  IF amino acid = 'P' THEN n = 4
|  IF amino acid = 'T' THEN n = 4
|  IF amino acid = 'A' THEN n = 4
```

```
21 |  IF amino acid = 'Y' THEN n = 2
22 |  IF amino acid = 'H' THEN n = 2
23 |  IF amino acid = 'Q' THEN n = 2
24 |  IF amino acid = 'N' THEN n = 2
25 |  IF amino acid = 'K' THEN n = 2
26 |  IF amino acid = 'D' THEN n = 2
27 |  IF amino acid = 'E' THEN n = 2
28 |  IF amino acid = 'C' THEN n = 2
29 |  IF amino acid = 'W' THEN n = 1
30 |  IF amino acid = 'R' THEN n = 6
31 |  IF amino acid = 'G' THEN n = 4
32 |  IF amino acid = '*' THEN n = 2
33 |  SET x to 0.0
34 |  SET r to n×random(iseed)
35 |  FOR each possible codon
36 |  |  IF pseq = amino acid THEN
37 |  |  |  INCREMENT x
38 |  |  END IF
39 |  |  IF x ≥ r THEN
40 |  |  |  IF constrained seq THEN
41 |  |  |  |  FOR each nucleotide in codon
42 |  |  |  |  |  IF constrained AND cseq ≠ original seq THEN
43 |  |  |  |  |  |  SET codon to original
44 |  |  |  |  |  |  GOTO 1
45 |  |  |  |  |  END IF
46 |  |  |  |  END LOOP
47 |  |  |  END IF
48 |  |  |  SET codon to cseq
49 |1 |  |  EXIT LOOP
50 |  |  END IF
51 |  END LOOP
52 END LOOP
```

TABLE A.8: Positions of ϕ and ω positions.

Species	ϕ	ω
HBV (human)	1773–1789	1832–1837
WHV (woodchuck)	1893–1908	1947–1953
GSHBV (Groundsquirrel)	3274–3289	17–23
WMHBV (woolly monkey)	1775–1791	1834–1839
HBHBV (horseshoe bat)	1619–1634	1673–1679
RBHBV (roundleaf bat)	1619–1634	1673–1680
TBHBV (tentmaking bat)	1584–1608	1648–1654
BtHBV ("bat")	1786–1801	1840–1848
DHBV (duck)	2471–2475 2492–2499	2554–2560
HHBV (heron)	2471–2475 2492–2499	2554–2560
CrHBV (crane)	2462–2467 2483–2490	2545–2551
StHBV (stork)	2477–2481 2498–2505	2560–2566
ShGHBV (sheldgoose)	2495–2500 2516–2523	2578–2584
SGHBV (snowgoose	2468–2472 2489–2496	2551–2557

TABLE A.8: (continued)

RHBV (Ross's goose)	2462–2467 2483–2490	2545–2551
PHBV (parrot)	2486–2490 2507–2514	2569–2575

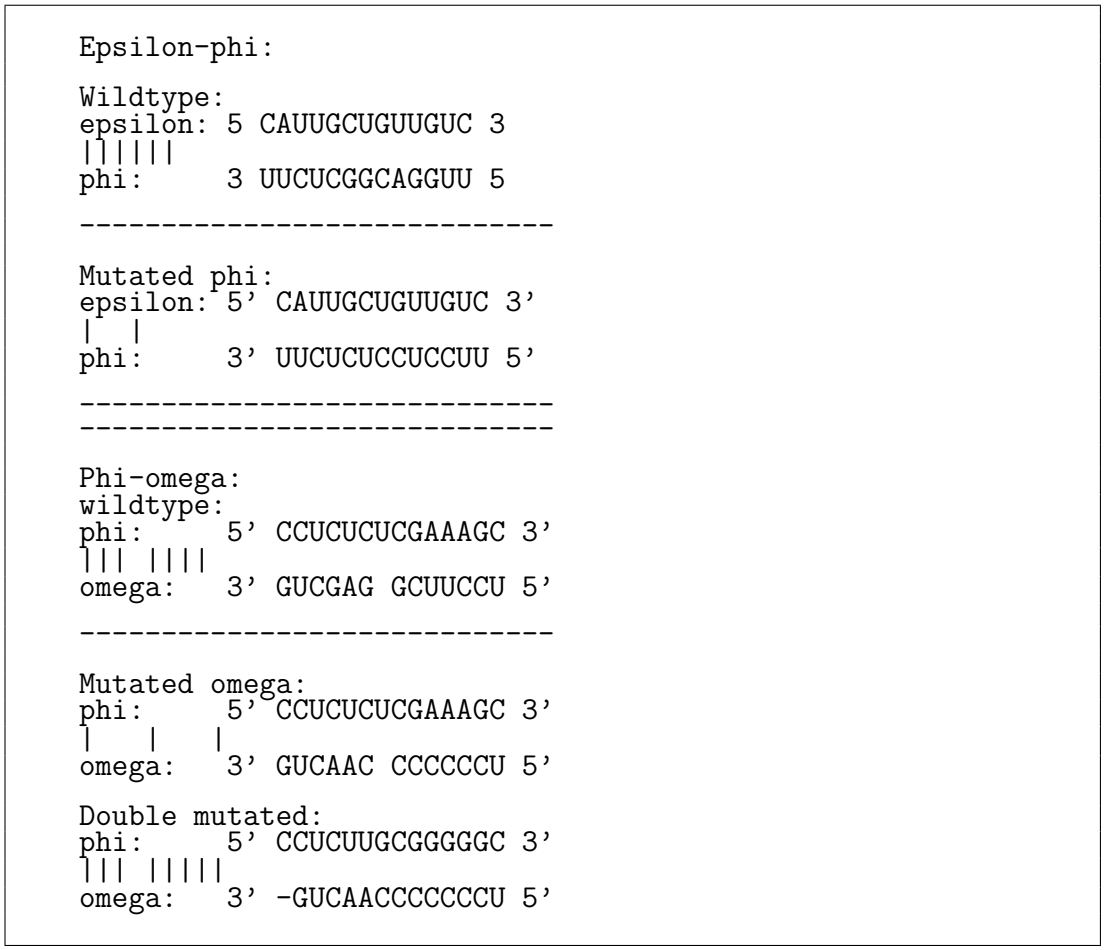

FIGURE A.2: **Suggested mutations in DHBV ϕ and ω.** The two identified parts of ϕ are shown together with their respective interaction partners ε or ω. Base-pairing is indicated by |. Suggested changes to the sequences are shown below. For ω compensatory mutations to restore base-pairing are also shown.

ALGORITHM A.14: No 4s rule.

```
1  SET currCP to CP position at PS most 5'
2  SET occupiedCP to 1
3  FOR 1 to 4
4  |  SET nextCP to CP position for clockwise move from currCP
5  |  IF nextCP occupied THEN
6  |  |  INCREMENT occupiedCP
7  |  END IF
8  |  SET currCP to nextCP
9  END LOOP
10 SET nextCP to CP position for across move from CP at PS most 5'
11 IF occupiedCP = 4 and nextCP unoccupied THEN
12 |  SET nextCP to unaddable
13 END IF
```

TABLE A.9: Number of RNAs part of semi- or complete capsids per energy condition in kcal/mol at the end of the simulation. The number of complete capsids with the 3' conserved PS or the three PSs around and including TR at the position mapped by Dai et al. (2017) as well as the entire predicted nucleus is shown.

AB:CC	AB:AB	semi	complete	3' conserved PS	TR + neighbours	whole nucleus
-1	-1	0	0	0	0	0
-1	-1.5	0	0	0	0	0
-2	-0.5	994	103	17	4	1
-2	-1	882	112	22	12	4
-2	-1.5	659	87	22	8	5
-2	-2	463	65	17	7	5
-2	-2.5	332	54	12	9	6
-3	-0.5	227	225	152	178	135
-3	-1	196	196	141	174	131
-3	-1.5	189	189	128	165	111
-3	-2	143	143	103	133	95
-3	-2.5	119	119	84	112	82
-3	-3	87	87	60	87	60
-3	-3.5	63	63	48	62	48
-4	-0.5	99	99	55	92	53
-4	-1	95	95	53	92	51
-4	-1.5	84	84	42	82	42
-4	-2	53	53	27	52	26
-4	-2.5	61	61	31	61	31
-4	-3	38	38	23	38	23
-4	-3.5	48	48	33	48	33
-4	-4	33	33	25	33	25
-4	-4.5	36	36	25	36	25

Packaging signal position of TR of 60 \ Packaging signal position of 5' maturation protein contact of 60	2	3	4	5	6	7	8	9	10	11
20	191	106	96	155	113	142	119	125	168	68
21	227	105	82	195	108	149	114	105	129	88
22	188	100	92	202	86	101	116	91	102	99
23	204	95	64	182	99	136	113	146	163	91
24	164	72	62	182	82	135	113	110	117	95
25	219	116	91	190	110	158	149	143	159	106
26	226	88	61	213	81	128	139	109	111	123
27	232	102	95	182	75	137	112	120	95	101
28	232	112	77	157	84	131	125	162	125	152
29	230	113	64	180	92	144	137	131	137	101
30	246	120	87	171	95	163	171	177	149	134
31	231	98	88	158	81	135	133	118	124	125
32	197	121	98	203	111	149	150	150	146	129
33	234	127	68	177	96	158	152	118	138	125
34	216	97	75	173	97	174	124	128	138	123
35	211	115	85	202	124	155	151	181	168	145
36	200	106	78	142	96	125	130	135	129	122
37	207	91	70	176	69	150	140	166	146	141
38	238	116	76	178	103	116	125	156	132	142
39	184	92	57	143	72	133	133	128	150	126
40	195	124	77	204	112	133	160	176	170	168
41	164	93	57	122	102	91	114	156	119	106
42	164	102	73	166	95	130	144	170	139	106
43	179	92	54	145	77	101	107	165	115	139
44	145	93	55	139	71	92	101	137	124	102
45	199	112	71	181	90	127	145	139	144	139
46	150	97	58	94	57	95	109	101	76	102
47	176	69	55	124	83	105	128	129	134	130
48	161	86	59	109	72	81	143	112	125	110
49	158	75	39	136	71	107	123	104	89	107
50	138	48	47	128	84	72	113	143	115	113

FIGURE A.3: **Effect of nucleation PS site combinations on assembly model performance with 5' MPC extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of extended 5' MPC, TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
20	36	64	226	147	199	277	257	251	160	273
21	56	94	248	150	190	329	251	291	266	270
22	96	92	270	170	200	317	254	247	239	256
23	36	69	288	146	164	353	208	245	192	303
24	49	82	204	151	222	306	231	278	233	286
25	39	66	225	169	249	343	196	245	222	305
26	38	58	173	152	152	362	246	261	221	335
27	53	64	218	137	195	290	205	278	182	255
28	52	55	207	118	199	351	214	325	224	321
29	56	69	182	136	158	270	229	253	184	304
30	48	72	284	134	180	341	191	251	186	334
31	46	54	250	137	135	343	213	278	185	270
32	24	46	192	125	126	285	166	241	139	304
33	28	39	283	152	191	354	242	248	203	339
34	23	38	203	94	129	272	141	232	165	257
35	31	37	234	164	152	295	204	237	179	291
36	17	35	201	106	166	299	195	226	175	260
37	25	44	247	119	138	287	159	240	150	312
38	20	28	182	110	153	283	191	219	168	284
39	12	31	191	94	107	277	135	191	117	237
40	29	49	268	137	201	409	195	321	176	312
41	21	18	180	79	97	255	136	164	97	200
42	11	28	277	120	122	285	147	213	152	245
43	14	26	181	81	110	236	125	197	130	186
44	12	23	211	88	128	235	101	182	115	210
45	5	19	263	126	96	316	147	221	122	231
46	8	15	163	73	82	207	71	152	75	154
47	12	33	267	97	117	276	125	222	133	238
48	6	15	166	57	82	218	87	146	83	128
49	6	17	157	58	88	235	96	179	96	204
50	7	12	227	68	65	255	88	226	131	184

FIGURE A.4: **Effect of nucleation PS site combinations on assembly model performance with TR extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, extended TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of 5' maturation protein contact of 60

Packaging signal position of TR of 60	2	3	4	5	6	7	8	9	10	11
20	214	123	204	185	165	185	135	143	89	142
21	228	151	218	217	165	182	185	173	156	145
22	176	139	161	142	182	159	163	158	100	153
23	263	146	235	207	190	236	200	236	132	204
24	176	122	174	207	160	173	170	158	127	198
25	246	125	168	181	172	219	138	228	129	182
26	222	129	182	197	179	181	182	186	172	185
27	225	124	180	198	159	174	188	208	138	196
28	267	143	206	200	173	250	181	248	184	196
29	231	107	159	170	167	177	162	195	168	188
30	284	134	203	231	198	226	190	247	163	248
31	220	131	161	172	193	205	143	197	179	204
32	232	90	163	193	169	195	140	245	140	216
33	266	149	197	214	195	205	199	236	175	236
34	189	108	180	168	162	176	168	212	152	214
35	275	111	161	177	189	216	218	222	178	197
36	255	113	147	162	168	197	148	207	191	184
37	241	106	179	181	176	228	149	209	146	200
38	214	92	137	166	160	178	145	178	135	191
39	235	125	176	204	199	245	174	237	174	201
40	202	82	128	160	133	181	134	173	163	174
41	257	119	181	199	139	223	159	210	190	223
42	192	89	148	175	138	182	127	172	165	153
43	235	114	141	192	168	187	107	185	182	174
44	239	119	152	181	146	187	118	196	167	163
45	204	126	158	195	154	211	141	179	169	162
46	222	79	165	178	171	215	124	214	158	136
47	178	109	166	183	153	200	121	190	169	168
48	219	103	159	157	143	195	118	188	182	138
49	168	73	151	148	114	145	77	119	87	123
50	174	85	177	162	150	255	123	217	147	181

FIGURE A.5: **Effect of nucleation PS site combinations on assembly model performance with 3' MPC extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, TR with neighbouring PSs, and extended 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60 \ Packaging signal position of 5' maturation protein contact of 60	2	3	4	5	6	7	8	9	10	11
20	235	119	121	208	125	194	189	161	193	115
21	196	106	63	119	72	156	105	113	138	84
22	178	111	70	152	91	105	109	115	133	81
23	200	118	58	180	83	157	105	93	162	87
24	166	107	61	173	89	134	130	119	139	82
25	170	117	81	162	88	175	101	161	154	122
26	181	97	85	178	106	155	130	119	161	85
27	225	159	76	169	97	158	137	147	130	120
28	196	111	47	153	87	146	145	106	159	110
29	169	131	62	176	87	134	124	118	128	112
30	221	166	67	189	122	179	141	166	129	114
31	201	113	76	164	98	163	139	104	128	123
32	221	160	92	192	110	126	135	129	141	102
33	206	137	67	166	90	161	145	152	143	127
34	204	163	74	184	110	167	155	159	144	153
35	195	150	86	235	101	143	149	163	167	137
36	227	162	84	155	97	125	123	116	142	127
37	222	141	72	165	106	135	140	125	145	124
38	225	136	82	175	80	125	117	110	129	124
39	218	147	57	153	86	116	109	107	130	110
40	164	117	83	211	96	138	133	139	139	129
41	183	127	64	124	64	100	92	87	87	81
42	164	119	84	191	99	151	99	132	119	92
43	175	125	49	109	65	109	90	108	84	73
44	145	123	72	124	88	112	102	105	130	118
45	135	132	78	194	72	131	121	109	100	116
46	117	86	37	125	56	110	58	80	66	69
47	190	114	74	188	98	125	119	124	118	133
48	108	62	80	145	47	100	68	94	89	94
49	156	130	54	124	79	105	105	83	102	80
50	207	105	89	207	83	151	104	122	125	96

FIGURE A.6: **Effect of nucleation PS site combinations on assembly model performance with extended 5' and 3' MPCs.** The MS2 Gillespie model was run with a complete nucleation complex consisting of extended 5' MPC, TR with neighbouring PSs, and extended 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of 5' maturation protein contact of 60

Packaging signal position of TR of 60	2	3	4	5	6	7	8	9	10	11
20	132	136	293	356	262	318	329	279	617	187
21	60	75	320	272	397	345	179	244	159	212
22	173	181	256	355	249	235	164	184	279	164
23	95	88	283	289	295	427	203	247	370	243
24	120	111	286	301	428	375	397	273	353	142
25	149	113	311	357	327	417	277	258	403	278
26	146	117	382	338	371	515	421	322	437	416
27	128	102	318	399	162	343	305	263	304	266
28	154	148	442	370	382	493	332	391	436	345
29	137	115	375	337	305	478	249	243	355	359
30	144	114	379	422	273	443	256	388	309	347
31	167	185	436	419	382	519	222	294	417	338
32	151	123	392	329	388	488	212	324	355	351
33	71	76	389	385	377	494	301	418	418	408
34	128	129	367	357	311	479	270	341	315	353
35	81	72	374	366	331	437	206	392	375	317
36	97	87	378	309	345	527	238	323	360	331
37	84	83	401	394	248	358	201	320	319	350
38	91	67	356	357	255	424	161	274	323	306
39	61	59	343	304	270	439	159	295	267	269
40	82	79	389	451	294	457	267	370	394	333
41	82	82	302	271	253	330	145	172	211	184
42	59	70	382	326	291	450	201	313	329	228
43	27	32	281	276	226	352	118	246	246	220
44	41	38	283	261	199	310	138	216	278	221
45	44	48	388	370	260	401	153	239	285	221
46	42	41	259	265	164	283	76	200	186	169
47	20	39	341	311	289	385	123	287	265	227
48	52	41	330	253	190	284	98	220	209	170
49	14	18	277	225	268	345	103	189	199	159
50	46	33	245	263	323	315	108	307	276	169

FIGURE A.7: **Effect of nucleation PS site combinations on assembly model performance with TR and 5' MPC extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of extended 5' MPC, extended TR with neighbouring PSs, and 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
20	45	66	224	138	149	266	187	163	171	214
21	50	82	282	132	133	282	214	218	179	208
22	76	97	217	186	167	263	210	201	184	221
23	38	55	249	155	154	248	199	196	172	212
24	62	77	210	156	183	261	212	225	191	249
25	51	69	219	153	197	286	190	254	208	246
26	30	55	174	151	160	292	215	238	205	296
27	37	48	202	119	168	254	223	299	171	232
28	60	53	190	133	131	283	183	295	182	267
29	60	82	172	115	141	234	196	217	170	230
30	41	61	242	120	158	282	198	290	163	284
31	27	56	214	115	130	266	172	232	170	233
32	26	51	217	141	139	256	166	254	138	266
33	24	29	207	111	171	247	203	248	161	233
34	16	36	201	109	129	251	142	239	140	224
35	27	51	226	133	133	302	189	274	138	284
36	27	29	208	87	147	225	162	207	124	228
37	24	49	256	129	114	213	145	223	117	267
38	29	42	177	101	144	237	165	212	120	246
39	26	43	183	92	116	197	113	193	125	236
40	29	46	236	111	155	284	182	267	167	278
41	17	28	176	72	99	169	121	171	101	181
42	15	30	296	98	121	312	160	246	185	261
43	9	26	174	78	86	200	98	168	98	182
44	17	51	245	81	122	234	124	192	102	177
45	22	20	340	105	102	251	140	226	92	236
46	16	30	171	63	74	170	74	155	74	148
47	19	36	282	116	116	268	157	213	143	240
48	4	19	182	71	89	148	93	168	76	151
49	6	12	153	37	71	193	95	119	84	162
50	12	8	249	85	103	292	114	291	137	224

FIGURE A.8: **Effect of nucleation PS site combinations on assembly model performance with TR and 3' MPC extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of a 5' MPC, extended TR with neighbouring PSs, and extended 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Packaging signal position of TR of 60	Packaging signal position of 5' maturation protein contact of 60									
	2	3	4	5	6	7	8	9	10	11
20	183	151	357	323	243	284	378	299	481	95
21	48	39	234	247	270	253	165	146	117	110
22	148	99	197	225	182	180	122	184	198	82
23	113	114	245	261	240	340	209	185	335	171
24	76	98	235	221	226	262	220	190	241	194
25	163	119	238	247	231	316	248	238	305	224
26	127	100	271	319	264	347	266	276	385	238
27	82	80	239	278	223	252	270	237	240	195
28	119	110	282	316	234	307	289	301	349	298
29	97	101	204	266	151	274	172	190	309	216
30	81	73	283	307	269	293	236	280	283	290
31	86	76	299	249	231	281	177	193	279	243
32	110	113	260	260	227	318	189	239	299	268
33	33	38	254	255	242	321	224	273	290	300
34	143	116	254	259	207	265	167	239	233	276
35	42	39	269	321	201	256	182	276	311	274
36	81	56	270	209	204	275	162	201	223	255
37	42	47	256	233	180	226	196	207	271	235
38	49	45	251	255	213	247	146	242	249	248
39	17	25	182	172	159	222	91	183	179	207
40	38	46	272	319	249	283	208	258	294	279
41	43	35	177	176	134	181	87	121	115	135
42	27	28	280	272	252	342	167	181	290	184
43	17	12	166	145	121	180	91	159	123	158
44	34	33	229	168	218	237	97	150	205	148
45	28	24	234	280	121	178	161	190	181	185
46	14	16	191	179	138	235	78	180	147	173
47	34	38	237	295	242	268	148	246	227	189
48	30	17	243	253	133	217	75	218	166	179
49	20	23	177	240	213	305	84	169	141	147
50	50	74	234	337	189	347	134	358	290	176

FIGURE A.9: **Effect of nucleation PS site combinations on assembly model performance with all nucleus parts extended.** The MS2 Gillespie model was run with a complete nucleation complex consisting of extended 5' MPC, extended TR with neighbouring PSs, and extended 3' MPC. The PS positions for 5' contact and TR were varied between 2 and 11 and 20 and 50, respectively. Inter-dimer energies of -3.0 AB:CC and -0.5 AB:AB were used.

Abbreviations

(+)DNA plus-strand DNA. 68

(-)DNA minus-strand DNA. 68, 70, 110

AS primer acceptor site. 68, 115

BtHV bat hepatitis virus. 63, 72

cccDNA covalently closed circular DNA. 65

CP capsid protein. 18, 19, 21, 47–49, 58, 61, 66, 73, 74, 76, 89, 103, 104, 108, 119–121, 124–126

cryo-EM cryo-electron microscopy. 68, 104, 105, 126, 128

DHBV duck hepatitis B virus. 63, 73, 110–112, 115, 119, 120

dsDNA double-stranded DNA. 61, 65

E. coli Escherichia coli. 96–100, 104, 119–121, 123, 126

eIF eukaryotic initiation factor. 36

fc fold change. 78

FM Fitch-Margoliash. 30

HBcAg core protein. 9, 63, 65, 66, 68, 71–74, 95, 98–100, 103, 106, 119, 121

HBeAg pre-core protein. 63, 65

HBV hepatitis B virus. 9, 21, 61–63, 65, 66, 68, 71–74, 79, 82, 83, 85, 91–97, 99, 101–105, 110–112, 115, 119–121, 126

HBxAg X protein. 63, 65

HHBV heron hepatitis B virus. 110, 112, 115

HIV human immunodeficiency virus. 21

LHBsAg long surface protein. 63, 65

MFE minimum free energy. 35, 36, 44, 76

MHBsAg middle surface protein. 63, 65

ML maximum likelihood. 30, 31

MP maximum parsimony. 30

MRCA most recent common ancestor. 28–30

mRNA messenger RNA. 18, 21, 35, 65, 96–100, 103, 111, 123, 124

MSA multiple sequence alignment. 27, 29, 50, 105

NC nucleation complex. 112

NJ neighbor-joining. 29, 30

nt nucleotide. 36, 50, 52, 57, 63, 66–68, 70, 74, 76, 91, 95, 96, 105, 108, 110–112, 115, 124

oligo oligonucleotide. 74, 89, 90

ORF open reading frame. 63, 65, 73, 98

PCR polymerase chain reaction. 73, 74

pgRNA pre-genomic RNA. 61, 64–66, 68–70, 91, 95–100, 102, 103, 105, 108, 110, 111, 114, 115

Pol DNA polymerase. 63–70, 73, 96, 98–100, 103, 104

poly-A poly-adenylation. 65, 92, 111

PS packaging signal. 18, 19, 21, 22, 34–36, 39, 44, 47–50, 52, 54, 55, 57–59, 61, 66, 68, 71, 73, 76, 79, 82, 85, 87, 89–92, 94–97, 99–101, 103–106, 108, 120, 124, 126, 133, 134

rcDNA relaxed circular DNA. 65

SD Shine-Dalgarno. 99

SDS-PAGE sodium dodecyl sulfate-polyacrylamide gel electrophoresis. 121

SELEX systematic evolution of ligands by exponential enrichment. 73, 74, 76, 79, 85

SHBsAg small surface protein. 63, 65, 71, 73

SL stem-loop. 18, 19, 35, 36, 38, 39, 42, 44, 46–49, 52, 55, 58, 59, 61, 65, 76, 79, 85, 89, 91, 92, 94–96, 98, 99, 101, 104–106, 108, 112, 124, 126

ssRNA single-stranded RNA. 18, 19, 21, 35, 36, 61, 65, 103, 104, 123, 126

UPGMA unweighted pair-group method using arithmetic averages. 29

VLP virus-like particle. 104, 120, 121

WAS weighted activity selection. 44, 46, 47, 49, 58

WHO World Health Organization. 62

WHV woodchuck hepatitis virus. 63, 110, 112, 119, 120

References

Aartsma-Rus, A., Van Deutekom, J. C. T., Fokkema, I. F., Van Ommen, G. J. B., & Den Dunnen, J. T. (2006). Entries in the Leiden Duchenne muscular dystrophy mutation database: An overview of mutation types and paradoxical cases that confirm the reading-frame rule. *Muscle and Nerve*, 34(2):135–144.

Abraham, T. M. & Loeb, D. D. (2006). Base Pairing between the 5' Half of ϵ and a cis -Acting Sequence , Φ, Makes a Contribution to the Synthesis of Minus-Strand DNA for Human Hepatitis B Virus. *Journal of Virology*, 80(9):4380–4387.

Abraham, T. M. & Loeb, D. D. (2007). The topology of hepatitis B virus pregenomic RNA promotes its replication. *Journal of Virology*, 81(21):11577–84.

Abzhanov, A. (2013). Von Baer's law for the ages: Lost and found principles of developmental evolution. *Trends in Genetics*, 29(12):712–722.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002a). From DNA to RNA. In: *Molecular Biology of the Cell*, (4th ed.).

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002b). From RNA to Protein. In: *Molecular Biology of the Cell*, (4th ed.).

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002c). The Structure and Function of DNA. In: *Molecular biology of the cell*, (4th ed.), chapter The Struct.

Altschul, S. F. & Erickson, B. W. (1986). A nonlinear measure of subalignment similarity and its significance levels. *Bulletin of Mathematical Biology*, 48(5-6):617–632.

Alvarado-Mora, M. V., Romano, C. M., Gomes-Gouvêa, M. S., Gutierrez, M. F., Carrilho, F. J., & Pinho, J. R. R. (2012). Phylogenetic analysis of complete genome sequences of hepatitis B virus from an Afro-Colombian community: presence of HBV F3/A1 recombinant strain. *Virology Journal*, 9:244.

Araujo, N. M., Araujo, O. C., Silva, E. M., Villela-Nogueira, C. A., Nabuco, L. C., Parana, R., Bessone, F., Gomes, S. A., Trepo, C., & Kay, A. (2013). Identification of novel recombinants of hepatitis B virus genotypes F and G in human immunodeficiency virus-positive patients from Argentina and Brazil. *The Journal of General Virology*, 94(Pt 1):150–158.

Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, 6(OCT):319.

Baker, T. S., Krol, M. A., Johnson, J. E., Ahlquist, P., Tate, J., & Olson, N. H. (2002). RNA-controlled polymorphism in the in vivo assembly of 180-subunit and 120-subunit virions from a single capsid protein. *Proceedings of the National Academy of Sciences*, 96(24):13650–13655.

Bamford, D. H. (2003). Do viruses form lineages across different domains of life? *Research in Microbiology*, 154(4):231–236.

Bamford, D. H., Burnett, R. M., & Stuart, D. I. (2002). Evolution of viral structure. *Theoretical Population Biology*, 61(4):461–470.

Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current Opinion in Structural Biology*, 15(6):655–663.

Bartenschlager, R., Junker-Niepmann, M., & Schaller, H. (1990). The P gene product of hepatitis B virus is required as a structural component for genomic RNA encapsidation. *Journal of Virology*, 64(11):5324–32.

Bartenschlager, R. & Schaller, H. (1988). The amino-terminal domain of the hepadnaviral P-gene encodes the terminal protein (genome-linked protein) believed to prime reverse transcription. *The EMBO Journal*, 7(13):4185–92.

Bartenschlager, R. & Schaller, H. (1992). Hepadnaviral assembly is initiated by polymerase binding to the encapsidation signal in the viral RNA genome. *The EMBO Journal*, 11(9):3413–3420.

Basnak, G., Morton, V. L., Rolfsson, Ó., Stonehouse, N. J., Ashcroft, A. E., & Stockley, P. G. (2010). Viral genomic single-stranded RNA directs the pathway toward a T = 3 capsid. *Journal of Molecular Biology*, 395(5):924–936.

Beck, J., Bartos, H., & Nassal, M. (1997). Experimental confirmation of a hepatitis B virus (HBV) ϵ-like bulge-and-loop structure in avian HBV RNA encapsidation signals. *Virology*, 227(2):500–504.

Beck, J. & Nassal, M. (1998). Formation of a functional hepatitis B virus replication initiation complex involves a major structural alteration in the RNA template. *Molecular and Cellular Biology*, 18(11):6265–72.

Beck, J. & Nassal, M. (2007). Hepatitis B virus replication. *World journal of gastroenterology : WJG*, 13(1):48–64.

Beckett, D. & Uhlenbeck, O. C. (1988). Ribonucleoprotein complexes of R17 coat protein and a translational operator analog. *Journal of Molecular Biology*, 204(4):927–938.

Beckett, D., Wu, H. N., & Uhlenbeck, O. C. (1988). Roles of operator and nonoperator RNA sequences in bacteriophage R17 capsid assembly. *Journal of Molecular Biology*, 204(4):939–947.

Beekwilder, J. (1996). *Secondary structure of the RNA genome of bacteriophage Qbeta*. PhD thesis, Leiden.

Beekwilder, J., Nieuwenhuizen, R., Poot, R., & Van Duin, J. (1996). Secondary structure model for the first three domains of Qβ RNA. Control of A-protein synthesis. *Journal of Molecular Biology*, 256(1):8–19.

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

Bellman, R. (2003). *Dynamic programming*. Dover Publications.

Belyi, V. A. & Muthukumar, M. (2006). Electrostatic origin of the genome packing in viruses. *Proceedings of the National Academy of Sciences*, 103(46):17174–17178.

Benson, S. D., Bamford, J. K., Bamford, D. H., & Burnett, R. M. (1999). Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. *Cell*, 98(6):825–833.

Benson, S. D., Bamford, J. K., Bamford, D. H., & Burnett, R. M. (2004). Does common architecture reveal a viral lineage spanning all three domains of life? *Molecular Cell*, 16(5):673–685.

Bernardi, A. & Spahr, P.-F. (1972). Nucleotide sequence at the binding site for coat protein on RNA of bacteriophage R17. *Proceedings of the National Academy of Sciences*, 69(10):3033–3037.

Bertsekas, D. P. (2017). *Dynamic programming and optimal control*, (4 ed.). Athena Scientific.

Biffin, E., Harrington, M. G., Crisp, M. D., Craven, L. A., & Gadek, P. A. (2007). Structural partitioning, paired-sites models and evolution of the ITS transcript in Syzygium and Myrtaceae. *Molecular phylogenetics and evolution*, 43(1):124–39.

Billeter, M. A., Libonati, M., Viñuela, E., & Weissmann, C. (1966). Replication of viral ribonucleic acid. X. Turnover of virus-specific double-stranded ribonu-

cleic acid during replication of phage MS2 in Escherichia coli. *The Journal of Biological Chemistry*, 241(20):4750–7.

Birnbaum, F. & Nassal, M. (1990). Hepatitis B virus nucleocapsid assembly: primary structure requirements in the core protein. *Journal of Virology*, 64(7):3319–30.

Bollyky, P., Rambaut, A., Grassley, N., Carman, W., & Holmes, E. (1998). Hepatitis B virus has a recent new world evolutionary origin. *Journal of Hepatology*, 28:96.

Bollyky, P. L., Rambaut, A., Harvey, P. H., & Holmes, E. C. (1996). Recombination between sequences of hepatitis B virus from different genotypes. *Journal of Molecular Evolution*, 42(2):97–102.

Böttcher, B. & Nassal, M. (2018). Structure of mutant hepatitis B core protein capsids with premature secretion phenotype. *Journal of Molecular Biology*, 430(24):4941–4954.

Bowyer, S. M. & Sim, J. G. M. (2000). Relationships within and between genotypes of hepatitis B virus at points across the genome: Footprints of recombination in certain isolates. *Journal of General Virology*, 81(2):379–392.

Brauckmann, S. (2012). Karl Ernst von Baer (1792-1876) and evolution. *International Journal of Developmental Biology*, 56(9):653–660.

Bremer, B., Jansen, R. K., Oxelman, B., Backlund, M., Lantz, H., & Kim, K. J. (1999). More characters or more taxa for a robust phylogeny - Case study from the coffee family (Rubiaceae). *Systematic Biology*, 48(3):413–435.

Brodski, L. I., Ivanov, V. V., Kaladzidis, I. L., Leontovich, A. M., Nikolaev, V. K., Feranchuk, S. I., & Drachev, V. A. (1995). GeneBee-NET: An Internet based server for biopolymer structure analysis. *Biokhimiia*, 60(8):1221–30.

Bruno, W. J., Socci, N. D., & Halpern, A. L. (2000). Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1):189–197.

Bunka, D. H. J., Lane, S. W., Lane, C. L., Dykeman, E. C., Ford, R. J., Barker, A. M., Twarock, R., Phillips, S. E. V., & Stockley, P. G. (2011). Degenerate RNA packaging signals in the genome of Satellite Tobacco Necrosis Virus: implications for the assembly of a T=1 capsid. *Journal of Molecular Biology*, 413(1):51–65.

Cadena-Nava, R. D., Comas-Garcia, M., Garmann, R. F., Rao, A. L. N., Knobler, C. M., & Gelbart, W. M. (2012). Self-Assembly of viral capsid protein and RNA molecules of different sizes: requirement for a specific high protein/RNA mass ratio. *Journal of Virology*, 86(6):3318–3326.

Carey, J., Cameron, V., de Haseth, P. L., & Uhlenbeck, O. C. (1983a). Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. *Biochemistry*, 22(11):2601–10.

Carey, J., Lowary, P. T., & Uhlenbeck, O. C. (1983b). Interaction of R17 coat protein with synthetic variants of its ribonucleic acid binding site. *Biochemistry*, 22(20):4723–30.

Centers for Disease Control and Prevention (2008). Recommendations for identification and public health management of persons with chronic hepatitis B virus infection. *Morbidity and Mortality Weekly Report*, 57(RR-8):1–20.

Centers for Disease Control and Prevention (2018). 2018 recommended immunizations for children from birth through 6 years old. Available at: `http://www.cdc.gov/vaccines/parents/downloads/parent-ver-sch-0-6yrs.pdf`. [Accessed on 2018-06-04].

Chang, S.-f. F., Netter, H. J., Hildt, E., Schuster, R., Schaefer, S., Hsu, Y.-c. C., Rang, A., & Will, H. (2001). Duck hepatitis B virus expresses a regulatory HBx-

like protein from a hidden open reading frame. *Journal of Virology*, 75(1):161–70.

Chauhan, R., Kazim, S. N., Kumar, M., Bhattacharjee, J., Krishnamoorthy, N., & Sarin, S. K. (2008). Identification and characterization of genotype A and D recombinant hepatitis B virus from Indian chronic HBV isolates. *World Journal of Gastroenterology*, 14(40):6228–6236.

Chekaraou, M. A., Brichler, S., Mansour, W., Gal, F. L., Garba, A., Dény, P., & Gordien, E. (2010). A novel hepatitis B virus (HBV) subgenotype D (D8) strain, resulting from recombination between genotypes D and E, is circulating in Niger along with HBV/E strains. *Journal of General Virology*, 91(6):1609–1620.

Chen, B. F., Kao, J. H., Liu, C. J., Chen, D. S., & Chen, P. J. (2004). Genotypic dominance and novel recombinations in HBV genotype B and C co-infected intravenous drug users. *Journal of Medical Virology*, 73(1):13–22.

Crawford, E. M. & Gesteland, R. F. (1964). The adsorption of bacteriophage R-17. *Virology*, 22(1):165–167.

Crowther, R. A., Kiselev, N. A., Böttcher, B., Berriman, J. A., Borisova, G. P., Ose, V., & Pumpens, P. (1994). Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. *Cell*, 77(6):943–950.

Cui, C., Shi, J., Hui, L., Xi, H., Zhuoma, a., Quni, a., Tsedan, a., & Hu, G. (2002). The dominant hepatitis B virus genotype identified in Tibet is a C/D hybrid. *Journal of General Virology*, 83(11):2773–2777.

Curtis, S. A. (2003). The classification of greedy algorithms. *Science of Computer Programming*, 49(1-3):125–157.

Dai, X., Li, Z., Lai, M., Shu, S., Du, Y., Zhou, Z. H., & Sun, R. (2017). In situ

structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature*, 541(7635):112–116.

Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975.

de Levie, R. (2009). Stochastics, the Basis of Chemical Dynamics. *Journal of Chemical Education*, 77(6):771.

Deiorio-Haggar, K., Anthony, J., & Meyer, M. M. (2013). RNA structures regulating ribosomal protein biosynthesis in bacilli. *RNA Biology*, 10(7):1180–1184.

Dejean, A., Sonigo, P., Wain-Hobson, S., & Tiollais, P. (1984). Specific hepatitis B virus integration in hepatocellular carcinoma DNA through a viral 11-base-pair direct repeat. *Proceedings of the National Academy of Sciences of the United States of America*, 81(17 I):5350–5354.

Dent, K. C., Thompson, R., Barker, A. M., Hiscox, J. A., Barr, J. N., Stockley, P. G., & Ranson, N. A. (2013). The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release. *Structure*, 21(7):1225–1234.

Devkota, B., Petrov, A. S., Lemieux, S., Boz, M. B., Tang, L., Schneemann, A., Johnson, J. F., & Harvey, S. C. (2009). Structural and electrostatic characterization of pariacoto virus: Implications for viral assembly. *Biopolymers*, 91(7):530–538.

Díaz, S., Purvis, A., Cornelissen, J. H., Mace, G. M., Donoghue, M. J., Ewers, R. M., Jordano, P., & Pearse, W. D. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecology and Evolution*, 3(9):2958–2975.

Dill, J. A., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., & Ng, T. F. F. (2016). Distinct viral lineages from fish and amphibians re-

veal the complex evolutionary history of hepadnaviruses. *Journal of Virology*, 90(17):7920–7933.

Ding, Y., Chi, Y. C., & Lawrence, C. E. (2005). RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166.

Ding, Y. & Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301.

Drexler, J. F., Geipel, A., Konig, A., Corman, V. M., van Riel, D., Leijten, L. M., Bremer, C. M., Rasche, A., Cottontail, V. M., Maganga, G. D., Schlegel, M., Muller, M. A., Adam, A., Klose, S. M., Borges Carneiro, A. J., Stocker, A., Franke, C. R., Gloza-Rausch, F., Geyer, J., Annan, A., Adu-Sarkodie, Y., Oppong, S., Binger, T., Vallo, P., Tschapka, M., Ulrich, R. G., Gerlich, W. H., Leroy, E., Kuiken, T., Glebe, D., & Drosten, C. (2013). Bats carry pathogenic hepadnaviruses antigenically related to hepatitis B virus and capable of infecting human hepatocytes. *Proceedings of the National Academy of Sciences*, 110(40):16151–16156.

Dykeman, E. C., Grayson, N. E., Toropova, K., Ranson, N. A., Stockley, P. G., & Twarock, R. (2011). Simple rules for efficient assembly predict the layout of a packaged viral RNA. *Journal of Molecular Biology*, 408(3):399–407.

Dykeman, E. C., Stockley, P. G., & Twarock, R. (2010). Dynamic allostery controls coat protein conformer switching during MS2 phage assembly. *Journal of Molecular Biology*, 395(5):916–923.

Dykeman, E. C., Stockley, P. G., & Twarock, R. (2013a). Building a viral capsid in the presence of genomic RNA. *Physical Review E*, 87(2):022717–1–12.

Dykeman, E. C., Stockley, P. G., & Twarock, R. (2013b). Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and

geometry of the packaged genome. *Journal of molecular biology*, 425(17):3235–49.

Dykeman, E. C., Stockley, P. G., & Twarock, R. (2014). Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy. *Proceedings of the National Academy of Sciences*, 111(14):5361–5366.

Dykeman, E. C. & Twarock, R. (2010). All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 81(3):031908.

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, 93(23):13429–34.

Eggen, K. & Nathans, D. (1969). Regulation of protein synthesis directed by coliphage MS2 RNA. *Journal of Molecular Biology*, 39(2):293–305.

ElSawy, K. M., Caves, L. S., & Twarock, R. (2010). The impact of viral RNA on the association rates of capsid protein assembly: Bacteriophage MS2 as a case study. *Journal of Molecular Biology*, 400(4):935–947.

Endres, D. & Zlotnick, A. (2002). Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. *Biophysical Journal*, 83(2):1217–1230.

Erikson, R. L., Fenwick, M. L., & Franklin, R. M. (1964). Replication of bacteriophage RNA: Studies on the fate of parental RNA. *Journal of Molecular Biology*, 10(3):519–529.

Fallot, G., Halgand, B., Garnier, E., Branger, M., Gervais, a., Roque-Afonso, a. M., Thiers, V., Billaud, E., Matheron, S., Samuel, D., & Feray, C. (2012). Recombination of hepatitis B virus DNA in patients with HIV. *Gut*, 61(8):1197–1208.

Fallows, D. A. & Goff, S. P. (1995). Mutations in the epsilon sequences of human hepatitis B virus affect both RNA encapsidation and reverse transcription. *Journal of virology*, 69(5):3067–73.

Fares, M. A. & Holmes, E. C. (2002). A revised evolutionary history of hepatitis B virus (HBV). *Journal of Molecular Evolution*, 54(6):807–814.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791.

Felsenstein, J. & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104.

Feng, H., Beck, J., Nassal, M., & hong Hu, K. (2011). A SELEX-screened aptamer of human hepatitis B virus RNA encapsidation signal suppresses viral replication. *PLoS ONE*, 6(11).

Fenwick, M. L., Erikson, R. L., & Franklin, R. M. (1964). Replication of the RNA of Bacteriophage R17. *Science*, 146(3643):527–530.

Ferguson, N., Shepherd, D. A., Ashcroft, A. E., Freund, S. M. V., Alexander, C. G., & Jurgens, M. C. (2013). Thermodynamic origins of protein folding, allostery, and capsid formation in the human hepatitis B virus core protein. *Proceedings of the National Academy of Sciences*, 110(30):E2782–E2791.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A.,

Volckaert, G., & Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507.

Fiers, W., Contreras, R., Duerinck, F., Haegmean, G., Merregaert, J., Jou, W. M., Raeymakers, A., Volckaert, G., Ysebaert, M., Van De Kerckhove, J., Nolf, F., & Van Montagu, M. (1975). A-protein gene of bacteriophage MS2. *Nature*, 256(5515):273–278.

Fitch, W. M. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279–284.

Ford, R. J., Barker, A. M., Bakker, S. E., Coutts, R. H., Ranson, N. A., Phillips, S. E. V., Pearson, A. R., & Stockley, P. G. (2013). Sequence-specific, RNA-protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein. *Journal of Molecular Biology*, 425(6):1050–64.

Forrey, C. & Muthukumar, M. (2009). Electrostatics of capsid-induced viral RNA organization. *The Journal of Chemical Physics*, 131(10):105101.

Freeman, G. R. (1984). The emergence of stochastic theories: What are they and why are they special? *Journal of Chemical Education*, 61(11):944.

Fuerst, T. R., Niles, E. G., Studier, F. W., & Moss, B. (1986). Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T7 RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, 83(21):8122–8126.

Fukami, H. & Imahori, K. (1971). Control of translation by the conformation of messenger RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 68(3):570–3.

Gallina, A., Bonelli, F., Zentilin, L., Rindi, G., Muttini, M., & Milanesi, G. (1989). A recombinant hepatitis B core antigen polypeptide with the

protamine-like domain deleted self-assembles into capsid particles but fails to bind nucleic acids. *Journal of Virology*, 63(11):4645–4652.

Garmann, R. F., Goldfain, A. M., & Manoharan, V. N. (2019). Measurements of the self-assembly kinetics of individual viral capsids around their RNA genome. *Proceedings of the National Academy of Sciences of the United States of America*, 116(45):22485–22490.

Garmiri, P., Loua, A., Haba, N., Candotti, D., & Allain, J. P. (2009). Deletions and recombinations in the core region of hepatitis B virus genotype E strains from asymptomatic blood donors in Guinea, west Africa. *Journal of General Virology*, 90(10):2442–2451.

Garwes, D., Sillero, A., & Ochoa, S. (1969). Virus-specific proteins in Escherichia coli infected with phage Qβ. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, 186(1):166–172.

Gary, D. J., Puri, N., & Won, Y. Y. (2007). Polymer-based siRNA delivery: Perspectives on the fundamental and phenomenological distinctions from polymer-based DNA delivery. *Journal of Controlled Release*, 121(1-2):64–73.

Gascuel, O. & Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000.

Gazina, E. V., Fielding, J. E., Lin, B., & Anderson, D. A. (2000). Core Protein Phosphorylation Modulates Pregenomic RNA Encapsidation to Different Extents in Human and Duck Hepatitis B Viruses. *Journal of Virology*, 74(10):4721–4728.

Geraets, J. A., Dykeman, E. C., Stockley, P. G., Ranson, N. A., & Twarock, R. (2015). Asymmetric Genome Organization in an RNA Virus Revealed via Graph-Theoretical Analysis of Tomographic Data. *PLoS Computational Biology*, 11(3):e1004146.

Gerlich, W. H. & Robinson, W. S. (1980). Hepatitis B virus contains protein attached to the 5 terminus of its complete DNA strand. *Cell*, 21(3):801–809.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 61:2340–2361.

Gillespie, D. T. (1991). *Markov Processes: An Introduction for Physical Scientists*. Academic Press.

Gillespie, D. T. (2007). Stochastic Simulation of Chemical Kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55.

Godoy, B. A., Pinho, J. R. R., & Fagundes, N. J. (2020). Hepatitis B Virus: Alternative phylogenetic hypotheses and its impact on molecular evolution inferences. *Virus Research*, 276.

Godson, G. N. & Sinsheimer, R. L. (1967). The replication of bacteriophage MS2. *Journal of Molecular Biology*, 23:495–521.

Golmohammadi, R., Valegård, K., Fridborg, K., & Liljas, L. (1993). The refined structure of bacteriophage MS2 at 2·8 Å resolution. *Journal of Molecular Biology*, 234(3):620–639.

Goodenbour, J. M. & Pan, T. (2006). Diversity of tRNA genes in eukaryotes. *Nucleic Acids Research*, 34(21):6137–6146.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., & Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Research*, 38(Web Server):W695–W699.

Gould, S. J. (1977). *Ontogeny and phylogeny*. Belknap Press of Harvard University Press.

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2):221–4.

Grajales, A., Aguilar, C., & Sánchez, J. A. (2007). Phylogenetic reconstruction using secondary structures of Internal Transcribed Spacer 2 (ITS2, rDNA): Finding the molecular and morphological gap in Caribbean gorgonian corals. *BMC Evolutionary Biology*, 7:90.

Graybeal, A. (1998). Is It Better to Add Taxa or Characters to a Difficult Phylogenetic Problem? *Systematic Biology*, 47(1):9–17.

Groeneveld, H. (1997). *Secondary structure of bacteriophage MS2 RNA*. PhD thesis, Leiden.

Gussin, G. N. (1966). Three complementation groups in bacteriophage R17. *Journal of Molecular Biology*, 21(3):435–453.

Hagan, M. F. (2009). A theory for viral capsid assembly around electrostatic cores. *Journal of Chemical Physics*, 130(11):154709.

Hahn, C. M., Iwanowicz, L. R., Cornman, R. S., Conway, C. M., Winton, J. R., & Blazer, V. S. (2015). Characterization of a Novel Hepadnavirus in the White Sucker (Catostomus commersonii) from the Great Lakes Region of the United States. *Journal of Virology*, 89(23):11801–11811.

Haines, K. M. & Loeb, D. D. (2007). The Sequence of the RNA Primer and the DNA Template Influence the Initiation of Plus-strand DNA Synthesis in Hepatitis B Virus. *Journal of Molecular Biology*, 370(3):471–480.

Hall, B. K. (2003). Evo-Devo: evolutionary developmental mechanisms. *The International Journal of Developmental Biology*, 47(7-8):491–5.

Hamilton, W. R. (1858). Account of the icosian calculus. *Proceedings of the Royal Irish Academy*, 6:415–416.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29(2):147–160.

Hannoun, C., Norder, H., & Lindh, M. (2000). An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam. *Journal of General Virology*, 81(9):2267–2272.

Hao, R., Xiang, K., Peng, Y., Hou, J., Sun, J., Li, Y., Su, M., Yan, L., Zhuang, H., & Li, T. (2015). Naturally occurring deletion/insertion mutations within HBV whole genome sequences in HBeAg-positive chronic hepatitis B patients are correlated with baseline serum HBsAg and HBeAg levels and might predict a shorter interval to HBeAg loss and seroconversi. *Infection, Genetics and Evolution*, 33:261–268.

Haruna, I., Nishihara, T., & Watanabe, I. (1967). Template activity of various phage RNA for replicases of Qβ and VK phages. *Proceedings of the Japan Academy*, 43(5):375–377.

Haruna, I., Nozu, K., Ohtaka, Y., & Spiegelman, S. (1963). An RNA "Replicase" induced by and selective for a vbiral RMNA: Isolation and properties. *Proceedings of the National Academy of Sciences of the United States of America*, 50(5):905–11.

Haruna, I. & Spiegelman, S. (1965). Specific template requirments of RNA replicases. *Proceedings of the National Academy of Sciences of the United States of America*, 54(1962):579–587.

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

Hatton, T., Zhou, S., & Standring, D. N. (1992). RNA- and DNA-binding activities in hepatitis B virus capsid protein: a model for their roles in viral replication. *Journal of Virology*, 66(9):5232–5241.

Havert, M. B. & Loeb, D. D. (1997). cis-Acting sequences in addition to donor

and acceptor sites are required for template switching during synthesis of plus-strand DNA for duck hepatitis B virus. *Journal of Virology*, 71(7):5336–44.

Hemmi, H., Takeuchi, O., Kawai, T., Kaisho, T., Sato, S., Sanjo, H., Matsumoto, M., Hoshino, K., Wagner, H., Takeda, K., & Akira, S. (2000). A Toll-like receptor recognizes bacterial DNA. *Nature*, 408(6813):740–745.

Hoare, C. A. R. (1961). Algorithm 64: Quicksort. *Communications of the ACM*, 4(7):321.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, 125:167–188.

Hohn, T. (1969). Role of RNA in the assembly process of bacteriophage fr. *Journal of Molecular Biology*, 43(1):191–200.

Holmes, S. (2003). Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, 18(2):241–255.

Horiuchi, K. & Matsuhashi, S. (1970). Three cistrons in bacteriophage Qβ. *Virology*, 42(1):49–60.

Horiuchi, K., Webster, R. E., & Matsuhashi, S. (1971). Gene products of bacteriaphage Qβ. *Virology*, 45:429–439.

Horn, W. T., Tars, K., Grahn, E., Helgstrand, C., Baron, A. J., Lago, H., Adams, C. J., Peabody, D. S., Phillips, S. E., Stonehouse, N. J., Liljas, L., & Stockley, P. G. (2006). Structural basis of RNA binding discrimination between bacteriophages Qβ and MS2. *Structure*, 14(3):487–495.

Hu, J. & Boyer, M. (2006). Hepatitis B virus reverse transcriptase and ϵ RNA sequences required for specific interaction in vitro. *Journal of Virology*, 80(5):2141–2150.

Hung, P. P., Ling, C. M., & Overby, L. R. (1969). Self-assembly of Qβ and MS2 phage particles: Possible function of initiation complexes. *Science*, 166(3913):1638–1640.

Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.

Huy, T. T. T., Ngoc, T. T., & Abe, K. (2008). New complex recombinant genotype of hepatitis B virus identified in Vietnam. *Journal of Virology*, 82(11):5657–5663.

Huy, T. T. T., Ushijima, H., Ngoc, T. T., Ha, L. D., Hayashi, S., Sata, T., & Abe, K. (2003). Recombination of genotypes B and C in hepatitis B virus isolated from a Vietnamese patient with fulminant hepatitis. *Japanese Journal of Infectious Diseases*, 56(1):35–37.

Jacques, D. A., McEwan, W. A., Hilditch, L., Price, A. J., Towers, G. J., & James, L. C. (2016). HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. *Nature*, 536(7616):349–353.

Jamalyaria, F., Rohlfs, R., & Schwartz, R. (2005). Queue-based method for efficient simulation of biological self-assembly systems. *Journal of Computational Physics*, 204(1):100–120.

Janecek, L. L., Honeycutt, R. L., Adkins, R. M., & Davis, S. K. (1996). Mitochondrial gene sequences and the molecular systematics of the artiodactyl subfamily bovinae. *Molecular Phylogenetics and Evolution*, 6(1):107–19.

Jeong, J. K., Yoon, G. S., & Ryu, W. S. (2000). Evidence that the 5'-end cap structure is essential for encapsidation of hepatitis B virus pregenomic RNA. *Journal of Virology*, 74(12):5502–8.

Jo, E., Ryu, D. K., König, A., Park, S., Cho, Y., Park, S. H., Kim, T. H., Yoon, S. K., Ryu, W. S., Cechetto, J., & Windisch, M. P. (2020). Identification and

characterization of a novel hepatitis B virus pregenomic RNA encapsidation inhibitor. *Antiviral Research*, 175:104709.

Jou, W. M., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88.

Jow, H., Hudelot, C., Rattray, M., & Higgs, P. G. (2002). Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Molecular Biology and Evolution*, 19(9):1591–1601.

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. In: *Mammalian Protein Metabolism*, H. N. Munro, ed., (iii ed.), chapter 24, pages 21–132. Academic Press.

Junker-Niepmann, M., Bartenschlager, R., & Schaller, H. (1990). A short cis-acting sequence is required for hepatitis B virus pregenome encapsidation and sufficient for packaging of foreign RNA. *The EMBO Journal*, 9(10):3389–3396.

Kato, H., Orito, E., Gish, R. G., Bzowej, N., Newsom, M., Sugauchi, F., Suzuki, S., Ueda, R., Miyakawa, Y., & Mizokami, M. (2002). Hepatitis B e antigen in sera from individuals infected with hepatitis B virus of genotype G. *Hepatology*, 35(4):922–929.

Kawai, T. & Akira, S. (2011). Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. *Immunity*, 34(5):637–650.

Keane, S. C., Heng, X., Lu, K., Kharytonchyk, S., Ramakrishnan, V., Carter, G., Barton, S., Hosic, A., Florwick, A., Santos, J., Bolden, N. C., McCowin, S., Case, D. A., Johnson, B. A., Salemi, M., Telesnitsky, A., & Summers, M. F. (2015). Structure of the HIV-1 RNA packaging signal. *Science*, 348(6237):917–921.

Kebbekus, P., Draper, D. E., & Hagerman, P. (1995). Persistence Length of RNA. *Biochemistry*, 34(13):4354–4357.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., & Wolf, M. (2010). Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct*, 5(1):1–12.

Kenah, E., Britton, T., Halloran, M. E., & Longini, I. M. (2016). Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLoS Computational Biology*, 12(4):e1004869.

Kenney, J. M., von Bonsdorff, C. H., Nassal, M., & Fuller, S. D. (1995). Evolutionary conservation in the hepatitis B virus core structure: comparison of human and duck cores. *Structure*, 3(10):1009–19.

Kim, S., Wang, H., & Ryu, W.-S. (2010). Incorporation of eukaryotic translation initiation factor eIF4E into viral nucleocapsids via interaction with hepatitis B virus polymerase. *Journal of Virology*, 84(1):52–8.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences*, 78(1):454–458.

Kiran, D. & Kiran, D. (2017). Reliability Engineering. In: *Total Quality Management*, chapter 27, pages 391–404. Butterworth-Heinemann.

Kivenson, A. & Hagan, M. F. (2010). Mechanisms of capsid assembly around a polymer. *Biophysical Journal*, 99(2):619–628.

Kleinberg, J. & Tardos, E. (2006). *Algorithm design*. Pearson/Addison-Wesley.

Knapman, T. W., Morton, V. L., Stonehouse, N. J., Stockley, P. G., & Ashcroft, A. E. (2010). Determining the topology of virus assembly intermediates using

ion mobility spectrometry-mass spectrometry. *Rapid Communications in Mass Spectrometry*, 24(20):3033–3042.

Knaus, T. & Nassal, M. (1993). The encapsidation signal on the hepatitis B virus RNA pregenome forms a stem-loop structure that is critical for its function. *Nucleic Acids Research*, 21(17):3967–3975.

Köck, J., Nassal, M., Deres, K., Blum, H. E., & von Weizsacker, F. (2004). Hepatitis B Virus Nucleocapsids Formed by Carboxy-Terminally Mutated Core Proteins Contain Spliced Viral Genomes but Lack Full-Size DNA. *Journal of Virology*, 78(24):13812–13818.

Kolakofsky, D. & Weissmann, C. (1971a). Possible mechanism for transition of viral RNA from polysome to replication complex. *Nature: New biology*, 231(19):42–6.

Kolakofsky, D. & Weissmann, C. (1971b). Qβ replicase as repressor of Qβ RNA-directed protein synthesis. *BBA Section Nucleic Acids And Protein Synthesis*, 246(3):596–599.

Koning, R., Van Den Worm, S., Plaisier, J. R., Van Duin, J., Abrahams, J. P., & Koerten, H. (2003). Visualization by cryo-electron microscopy of genomic RNA that binds to the protein capsid inside bacteriophage MS2. *Journal of Molecular Biology*, 332(2):415–422.

Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234(2):187–208.

Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, 299(1-2):1–34.

Kramvis, A. (2014). Genotypes and genetic variability of hepatitis B virus. *Intervirology*, 57(3-4):141–150.

Kramvis, A., Kew, M., & Francois, G. (2005). Hepatitis B virus genotypes. *Vaccine*, 23(19):2409–2423.

Kramvis, A. & Kew, M. C. (1998). Structure and function of the encapsidation signal of hepadnaviridae. *Journal of Viral Hepatitis*, 5(6):357–367.

Kurbanov, F., Tanaka, Y., Fujiwara, K., Sugauchi, F., Mbanya, D., Zekeng, L., Ndembi, N., Ngansop, C., Kaptue, L., Miura, T., Ido, E., Hayami, M., Ichimura, H., & Mizokami, M. (2005). A new subtype (subgenotype) Ac (A3) of hepatitis B virus and recombination between genotypes A and E in Cameroon. *Journal of General Virology*, 86(7):2047–2056.

Lago, H., Parrott, A. M., Moss, T., Stonehouse, N. J., & Stockley, P. G. (2001). Probing the kinetics of formation of the bacteriophage MS2 translational operator complex: Identification of a protein conformer unable to bind RNA. *Journal of Molecular Biology*, 305(5):1131–1144.

Lanave, C., Preparata, G., Sacone, C., & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93.

Laoi, B. N. & Crowley, B. (2008). Molecular characterization of hepatitis B virus (HBV) isolates, including identification of a novel recombinant, in patients with acute HBV infection attending an Irish hospital. *Journal of medical virology*, 80(9):1554–64.

Le Pogam, S., Chua, P. K., Newman, M., & Shih, C. (2005). Exposure of RNA Templates and Encapsidation of Spliced Viral RNA Are Influenced by the Arginine-Rich Domain of Human Hepatitis B Virus Core Antigen (HBcAg 165-173). *Journal of Virology*, 79(3):1871–1887.

Leitner, T., Escanilla, D., Franzén, C., Uhlén, M., & Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree

analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20):10864–10869.

Levitt, N., Briggs, D., Gil, A., & Proudfoot, N. J. (1989). Definition of an efficient synthetic poly(A) site. *Genes & Development*, 3(7):1019–1025.

Lewellyn, E. B. & Loeb, D. D. (2011). The Arginine Clusters of the Carboxy-Terminal Domain of the Core Protein of Hepatitis B Virus Make Pleiotropic Contributions to Genome Replication. *Journal of Virology*, 85(3):1298–1309.

Lien, J. M., Aldrich, C. E., & Mason, W. S. (1986). Evidence that a capped oligoribonucleotide is the primer for duck hepatitis B virus plus-strand DNA synthesis. *Journal of Virology*, 57(1):229–236.

Lin, L. & Hu, J. (2008). Inhibition of Hepadnavirus Reverse Transcriptase- RNA Interaction by Porphyrin Compounds. *Journal of Virology*, 82(5):2305–2312.

Lin, T., Chen, Z., Usha, R., Stauffacher, C. V., Dai, J. B., Schmidt, T., & Johnson, J. E. (1999). The refined crystal structure of cowpea mosaic virus at 2.8 Å resolution. *Virology*, 265(1):20–34.

Lin, Y.-Y., Liu, C., Chien, W.-H., Wu, L.-L., Tao, Y., Wu, D., Lu, X., Hsieh, C.-H., Chen, P.-J., Wang, H.-Y., Kao, J.-H., & Chen, D.-S. (2015). New Insights into the Evolutionary Rate of Hepatitis B Virus at Different Biological Scales. *Journal of Virology*, 89(7):3512–3522.

Ling, C. M., Hung, P. P., & Overby, L. R. (1970). Independent assembly of Qβ and MS2 phages in doubly infected Escherichia coli. *Virology*, 40(4):920–929.

Littlejohn, M., Locarnini, S., & Yuen, L. (2016). virus and hepatitis D virus. *Cold Spring Harbor Perspectives in Medicine*, 6(1):a021360.

Liu, N., Ji, L., Maguire, M. L., & Loeb, D. D. (2004). cis-Acting sequences that contribute to the synthesis of relaxed-circular DNA of human hepatitis B virus. *Journal of Virology*, 78(2):642–9.

Liu, Y., Zhao, Q., Zhang, H., Xu, R., Li, Y., & Wei, L. (2016). A New Method to Predict RNA Secondary Structure Based on RNA Folding Simulation. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):990–995.

Lodish, H. F. (1968). Bacteriophage f2 RNA: Control of translation and gene order. *Nature*, 220(5165):345–350.

Lodish, H. F., Horiuchi, K., & Zinder, N. D. (1965). Mutants of the bacteriophage f2. V. On the production of noninfectious phage particles. *Virology*, 27(2):139–155.

Lodish, H. F. & Zinder, N. D. (1966). Mutants of the bacteriophage f2: VIII. Control mechanisms for phage-specific syntheses. *Journal of Molecular Biology*, 19(2):333–348.

Loeb, D. D., Hirsch, R. C., & Ganem, D. (1991). Sequence-independent RNA cleavages generate the primers for plus strand DNA synthesis in hepatitis B viruses: implications for other reverse transcribing elements. *The EMBO journal*, 10(11):3533–40.

Loeb, D. D., Tian, R., & Gulya, K. J. (1996). Mutations within DR2 independently reduce the amount of both minus- and plus-strand DNA synthesized during duck hepatitis B virus replication. *Journal of virology*, 70(12):8684–8690.

Loeb, T. & Zinder, N. D. (1961). A bacteriophage containing RNA. *Proceedings of the National Academy of Sciences*, 47(3):282–289.

Lopez, L., Flichman, D., Mojsiejczuk, L., Gonzalez, M. V., Uriarte, R., Campos, R., Cristina, J., & Garcia-Aguirre, L. (2015). Genetic variability of hepatitis B virus in Uruguay: D/F, A/F genotype recombinants. *Archives of Virology*, 160(9):2209–2217.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(26):1–14.

Lorenz, R., Hofacker, I. L., & Stadler, P. F. (2016). RNA folding with hard and soft constraints. *Algorithms for Molecular Biology*, 11(1):8.

Luo, K., Liu, Z., He, H., Peng, J., Liang, W., Dai, W., & Hou, J. (2004). The putative recombination of hepatitis B virus genotype B with PreC/C region of genotype C. *Virus Genes*, 29(1):31–41.

MacDonald, D. M., Holmes, E. C., Lewis, J. C. M., & Simmonds, P. (2000). Detection of hepatitis B virus infection in wild-born chimpanzees (Pan troglodytes verus): Phylogenetic relationships with human and other primate genotypes. *Journal of Virology*, 74(9):4253–4257.

Maguire, M. L. & Loeb, D. D. (2010). cis-acting sequences that contribute to synthesis of minus-strand DNA are not conserved between hepadnaviruses. *Journal of Virology*, 84(24):12824–12831.

Malys, N. & Nivinskas, R. (2009). Non-canonical RNA arrangement in T4-even phages: Accommodated ribosome binding site at the gene 26-25 intercistronic junction. *Molecular Microbiology*, 73(6):1115–1127.

Martinez-Urreaga, J., Mira, J., & Gonzáles-Fernández, C. (2003). Introducing stochastic simulation of chemical reactions using the Gillespie algorithm and MATLAB. *Chemical Engineering Science*, 37(1):14–19.

Mathews, D. H. (2006). Revolutions in RNA Secondary Structure Prediction. *Journal of Molecular Biology*, 359(3):526–532.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.

Proceedings of the National Academy of Sciences of the United States of America, 101(19):7287–7292.

Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940.

Matthews, K. S. & Cole, R. D. (1972). Shell formation by capsid protein of f2 bacteriophage. *Journal of Molecular Biology*, 65(1):1–15.

McAdams, H. H. & Arkin, A. (1999). It's a noisy business! Genetic regulation at the nanomolar scale. *Trends in genetics : TIG*, 15(2):65–9.

Michitaka, K., Tanaka, Y., Horiike, N., Duong, T. N., Chen, Y., Matsuura, K., Hiasa, Y., Mizokami, M., & Onji, M. (2006). Tracing the history of hepatitis B virus genotype D in Western Japan. *Journal of Medical Virology*, 78(1):44–52.

Milich, D. R., Chen, M. K., Hughes, J. L., & Jones, J. E. (1998). The secreted hepatitis B precore antigen can modulate the immune response to the nucleocapsid: a mechanism for persistence. *Journal of Immunology (Baltimore, Md.: 1950)*, 160(4):2013–2021.

Mira, J., Fernández, C. G., & Urreaga, J. M. (2003). Two examples of deterministic versus stochastic modeling of chemical reactions. *Journal of Chemical Education*, 80(12):1488–1493.

Miyakawa, Y. & Mizokami, M. (2003). Classifying hepatitis B virus genotypes. *Intervirology*, 46(6):329–338.

Miyake, T., Haruna, I., Shiba, T., Ito, Y. H., & Yamane, K. (1971). Grouping of RNA phages based on the template specificity of their RNA replicases. *Proceedings of the National Academy of Sciences of the United States of America*, 68(9):2022–2024.

Mizokami, M., Orito, E., ichi Ohba, K., Ikeo, K., Lau, J. Y. N., & Gojobori, T. (1997). Constrained evolution with respect to gene overlap of hepatitis B virus. *Journal of Molecular Evolution*, 44(1 Supplement):83–90.

Molnar-Kimber, K. L., Summers, J. W., & Mason, W. S. (1984). Mapping of the cohesive overlap of duck hepatitis B virus DNA and of the site of initiation of reverse transcription. *Journal of Virology*, 51(1):181–191.

Moore, C. H., Farron, F., Bohnert, D., & Weissmann, C. (1971). Possible origin of a minor Virus specific protein (A1) in Qβ particles. *Nature New Biology*, 234(50):204–206.

Morgan, S. R. & Higgs, P. G. (1996). Evidence for kinetic effects in the folding of large RNA molecules. *Journal of Chemical Physics*, 105(16):7152–7157.

Morozov, V., Pisareva, M., & Groudinin, M. (2000). Homologous recombination between different genotypes of hepatitis B virus. *Gene*, 260:55–65.

Mühlemann, B., Jones, T. C., De Barros Damgaard, P., Allentoft, M. E., Shevnina, I., Logvin, A., Usmanova, E., Panyushkina, I. P., Boldgiv, B., Bazartseren, T., Tashbaeva, K., Merz, V., Lau, N., Smrčka, V., Voyakin, D., Kitov, E., Epimakhov, A., Pokutta, D., Vicze, M., Price, T. D., Moiseyev, V., Hansen, A. J., Orlando, L., Rasmussen, S., Sikora, M., Vinner, L., Osterhaus, A. D., Smith, D. J., Glebe, D., Fouchier, R. A., Drosten, C., Sjögren, K. G., Kristiansen, K., & Willerslev, E. (2018). Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature*, 557(7705):418–423.

Mulyanto, Depamede, S. N., Surayah, K., Tjahyono, A. A. H., Jirintai, Nagashima, S., Takahashi, M., & Okamoto, H. (2010). Identification and characterization of novel hepatitis B virus subgenotype C10 in Nusa Tenggara, Indonesia. *Archives of Virology*, 155(5):705–715.

Mulyanto, Depamede, S. N., Wahyono, A., Jirintai, Nagashima, S., Takahashi, M., & Okamoto, H. (2011). Analysis of the full-length genomes of novel hepati-

tis B virus subgenotypes C11 and C12 in Papua, Indonesia. *Journal of Medical Virology*, 83(1):54–64.

Mulyanto, Pancawardani, P., Depamede, S. N., Wahyono, A., Jirintai, S., Nagashima, S., Takahashi, M., Nishizawa, T., & Okamoto, H. (2012). Identification of four novel subgenotypes (C13-C16) and two inter-genotypic recombinants (C12/G and C13/B3) of hepatitis B virus in Papua Province, Indonesia. *Virus Research*, 163(1):129–140.

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., Mayo, M. A., & Summers, M. D. (1995). *Virus taxonomy: the classification and nomenclature of viruses. Sixth report of the International Committee on Taxonomy of Viruses*. Springer Vienna.

Nassal, M. (1992). The arginine-rich domain of the hepatitis B virus core protein is required for pregenome encapsidation and productive viral positive-strand DNA synthesis but not for virus assembly. *Journal of Virology*, 66(7):4107–16.

Nassal, M., Junker-Niepmann, M., & Schaller, H. (1990). Translational inactivation of RNA function: Discrimination against a subset of genomic transcripts during HBV nucleocapsid assembly. *Cell*, 63(6):1357–1363.

Nassal, M. & Rieger, A. (1996). A bulged region of the hepatitis B virus RNA encapsidation signal contains the replication origin for discontinuous first-strand DNA synthesis. *Journal of Virology*, 70(5):2764–2773.

Nathans, D., Oeschger, M. P., Polmar, S. K., & Eggen, K. (1969). Regulation of protein synthesis directed by coliphage MS2 RNA. *Journal of Molecular Biology*, 39(2):279–292.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Neuman, B. W. & Buchmeier, M. J. (2016). Supramolecular Architecture of the Coronavirus Particle. In: *Advances in Virus Research*, volume 96, pages 1–27. Academic Press Inc.

NHS (2015). Hepatitis B vaccine. Available at: `https://www.nhs.uk/conditions/vaccinations/hepatitis-b-vaccine/`. [Accessed on 2018-06-04].

Norder, H., Couroucé, A., & Magnius, L. O. (1994). Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology*, 198(2):489–503.

Norder, H., Couroucé, A. M., Coursaget, P., Echevarria, J. M., Lee, S. D., Mushahwar, I. K., Robertson, B. H., Locarnini, S., & Magnius, L. O. (2004). Genetic diversity of hepatitis B virus strains derived worldwide: Genotypes, subgenotypes, and HBsAg subtypes. *Intervirology*, 47(6):289–309.

Nowak, M. A., Bonhoeffer, S., Hill, A. M., Boehme, R., Thomas, H. C., & McDade, H. (1996). Viral dynamics in hepatitis B virus infection. *Proceedings of the National Academy of Sciences*, 93(9):4398–4402.

Okamoto, H., Tsuda, F., Sakugawa, H., Sastrosoewignjo, R. I., Imai, M., Miyakawa, Y., & Mayumi, M. (1988). Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *Journal of General Virology*, 69:2575–83.

Olinger, C. M., Venard, V., Njayou, M., Bola Oyefolu, A. O., Maiga, I., Kemp, A. J., Omilabu, S. a., le Faou, A., & Muller, C. P. (2006). Phylogenetic analysis of the precore/core gene of hepatitis B virus genotypes E and A in West Africa: New subtypes, mixed infections and recombinations. *Journal of General Virology*, 87(5):1163–1173.

Olsthoorn, R. C. L. (1996). *Structure and evolution of RNA phages*. PhD thesis, Leiden.

Oropeza, C. E. & McLachlan, A. (2007). Complementarity between epsilon and phi sequences in pregenomic RNA influences hepatitis B virus replication efficiency. *Virology*, 359(2):371–381.

Osiowy, C., Giles, E., Tanaka, Y., Mizokami, M., & Minuk, G. Y. (2006). Molecular Evolution of Hepatitis B Virus over 25 Years. *Journal of Virology*, 80(21):10307–10314.

Osiowy, C., Gordon, D., Borlang, J., Giles, E., & Villeneuve, J. P. (2008). Hepatitis B virus genotype G epidemiology and co-infection with genotype A in Canada. *Journal of General Virology*, 89(12):3009–3015.

Osiowy, C., Kaita, K., Solar, K., & Mendoza, K. (2010). Molecular characterization of hepatitis B virus and a 9-year clinical profile in a patient infected with genotype I. *Journal of Medical Virology*, 82(6):942–948.

Ostrow, K. M. & Loeb, D. D. (2002). Characterization of the cis-acting contributions to avian hepadnavirus RNA encapsidation. *Journal of virology*, 76(18):9087–95.

Overby, L. R., Barlow, G. H., Doi, R. H., Jacob, M., & Spiegelman, S. (1966). Comparison of two serologically distinct ribonucleic acid bacteriophages. I. Properties of the viral particles. *Journal of Bacteriology*, 91(1):442–8.

Owiredu, W. K., Kramvis, A., & Kew, M. C. (2001a). Molecular analysis of hepatitis B virus genomes isolated from black African patients with fulminant hepatitis B. *Journal of Medical Virology*, 65(3):485–92.

Owiredu, W. K. B. A., Kramvis, A., & Kew, M. C. (2001b). Hepatitis B virus DNA in serum of healthy black African adults positive for hepatitis B surface antibody alone: Possible association with recombination between genotypes A and D. *Journal of Medical Virology*, 64(4):441–454.

Pal, S. K., Ray, S. S., & Ganivada, A. (2017). RNA secondary structure prediction: Soft computing perspective. *Studies in Computational Intelligence*, 712(1):195–222.

Patel, N., White, S. J., Thompson, R. F., Bingham, R., Weiß, E. U., Maskell, D. P., Zlotnick, A., Dykeman, E. C., Tuma, R., Twarock, R., Ranson, N. A., & Stockley, P. G. (2017). HBV RNA pre-genome encodes specific motifs that mediate interactions with the viral core protein that promote nucleocapsid assembly. *Nature Microbiology*, 2(8):17098.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., & Ho, D. D. (1996). HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586.

Pestova, T. V. & Kolupaeva, V. G. (2002). The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes and Development*, 16(22):2906–2922.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., & Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5:50.

Piel, F. B., Steinberg, M. H., & Rees, D. C. (2017). Sickle Cell Disease. *New England Journal of Medicine*, 376(16):1561–1573.

Pollack, J. R. & Ganem, D. (1993). An RNA stem-loop structure directs hepatitis B virus genomic RNA encapsidation. *Journal of Virology*, 67(6):3254–3263.

Poot, R. A., Tsareva, N. V., Boni, I. V., & van Duin, J. (1997). RNA folding kinetics regulates translation of phage MS2 maturation gene. *Proceedings of the National Academy of Sciences*, 94(19):10110–10115.

Porterfield, J. Z., Dhason, M. S., Loeb, D. D., Nassal, M., Stray, S. J., & Zlotnick, A. (2010). Full-Length Hepatitis B Virus Core Protein Packages Viral and

Heterologous RNA with Similarly High Levels of Cooperativity. *Journal of Virology*, 84(14):7174–7184.

Proctor, J. R. & Meyer, I. M. (2013). CoFold: An RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Research*, 41(9).

Propst Ricciuti, C. (1976). The effect of host cell starvation on virus induced lysis by MS2 bacteriophage. *Journal of General Virology*, 31(3):323–330.

Quarleri, J. (2014). Core promoter: A critical region where the hepatitis B virus makes decisions. *World Journal of Gastroenterology*, 20(2):425–435.

Rall, L. B., Standring, D. N., Laub, O., & Rutter, W. J. (1983). Transcription of hepatitis B virus by RNA polymerase II. *Molecular and Cellular Biology*, 3(10):1766–73.

Reuter, J. S. & Mathews, D. H. (2010). RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129.

Rieger, A. & Nassal, M. (1996). Specific hepatitis B virus minus-strand DNA synthesis requires only the 5' encapsidation signal and the 3'-proximal direct repeat DR1. *Journal of Virology*, 70(1):585–9.

Robertson, H., Webster, R. E., & Zinder, N. D. (1968). Bacteriophage coat protein as repressor. *Nature*, 218(5141):533–6.

Robertson, H. D. & Lodish, H. F. (1970). Messenger characteristics of nascent bacteriophage RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 67(2):710–716.

Rodríguez, F., Oliver, J. L., Marín, A., & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142(4):485–501.

Rogers, G. W., Richter, N. J., Lima, W. F., & Merrick, W. C. (2001). Modulation of the Helicase Activity of eIF4A by eIF4B, eIF4H, and eIF4F. *Journal of Biological Chemistry*, 276(33):30914–30922.

Rogers, G. W., Richter, N. J., & Merrick, W. C. (1999). Biochemical and kinetic characterization of the RNA helicase activity of eukaryotic initiation factor 4A. *Journal of Biological Chemistry*, 274(18):12236–12244.

Rolfsson, Ó., Middleton, S., Manfield, I. W., White, S. J., Fan, B., Vaughan, R., Ranson, N. A., Dykeman, E., Twarock, R., Ford, J., Cheng Kao, C., & Stockley, P. G. (2016). Direct Evidence for Packaging Signal-Mediated Assembly of Bacteriophage MS2. *Journal of Molecular Biology*, 428(2):431–448.

Rolfsson, O., Toropova, K., Morton, V., Francese, S., Basnak, G., Thompson, G. S., Homans, S. W., Ashcroft, A. E., Stonehouse, N. J., Ranson, N. A., & Stockley, P. G. (2008). RNA packing specificity and folding during assembly of the bacteriophage MS2. *Computational and Mathematical Methods in Medicine*, 9(3-4):339–349.

Romaniuk, P. J., Lowary, P., Wu, H. N., Stormo, G., & Uhlenbeck, O. C. (1987). RNA binding site of R17 coat protein. *Biochemistry*, 26(6):1563–8.

Rose, A. S., Bradley, A. R., Valasatava, Y., Duarte, J. M., Prlic, A., & Rose, P. W. (2018). NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, 34(21):3755–3758.

Roseman, A. M., Crowther, R. A., Berriman, J. A., Wynne, S. A., & Butler, P. J. G. (2005). A structural model for maturation of the hepatitis B virus core. *Proceedings of the National Academy of Sciences*, 102(44):15821–15826.

Routh, A., Domitrovic, T., & Johnson, J. E. (2012). Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proceedings of the National Academy of Sciences*, 109(6):1907–1912.

Rumnieks, J. & Tars, K. (2017). Crystal structure of the maturation protein from bacteriophage Qβ. *Journal of Molecular Biology*, 429(5):688–696.

Ryu, D.-K., Kim, S., & Ryu, W.-S. (2008). Hepatitis B virus polymerase suppresses translation of pregenomic RNA via a mechanism involving its interaction with 5' stem-loop structure. *Virology*, 373(1):112–23.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Sakamoto, T., Tanaka, Y., Orito, E., Co, J., Clavio, J., Sugauchi, F., Ito, K., Ozasa, A., Quino, A., Ueda, R., Sollano, J., & Mizokami, M. (2006). Novel subtypes (subgenotypes) of hepatitis B virus genotypes B and C among chronic liver disease patients in the Philippines. *Journal of General Virology*, 87(7):1873–1882.

Sakurai, T., Miyake, T., Shiba, T., & Watanabe, I. (1968). Isolation of a possible fourth group of RNA phage. *Japanese Journal of Microbiology*, 12(4):544–546.

Samoilov, M. S. & Arkin, A. P. (2006). Deviant effects in molecular reaction pathways. *Nature Biotechnology*, 24(10):1235–1240.

Schöniger, M. & Von Haeseler, A. (1994). A Stochastic Model for the Evolution of Autocorrelated DNA Sequences.

Scotland, R. W., Olmstead, R. G., & Bennett, J. R. (2003). Phylogeny Reconstruction: The Role of Morphology. *Systematic Biology*, 52(4):539–548.

Sede, M., Lopez-Ledesma, M., Frider, B., Pozzati, M., Campos, R. H., Flichman, D., & Quarleri, J. (2014). Hepatitis B virus depicts a high degree of conservation during the immune-tolerant phase in familiarly transmitted chronic hepatitis B infection: Deep-sequencing and phylogenetic analysis. *Journal of Viral Hepatitis*, 21(9):650–661.

Seibel, P., Muller, T., Dandekar, T., Schultz, J., & Wolf, M. (2006). 4SALE - A tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, 7(1):498.

Selzer, L., Katen, S. P., & Zlotnick, A. (2014). The hepatitis B virus core protein intradimer interface modulates capsid assembly and stability. *Biochemistry*, 53(34):5496–504.

Selzer, L. & Zlotnick, A. (2015). Assembly and release of hepatitis B virus. *Cold Spring Harbor Perspectives in Medicine*, 5(12):a021394.

Shi, S., Zhang, X. L., Zhao, X. L., Yang, L., Du, W., & Wang, Y. J. (2019). Prediction of the RNA Secondary Structure Using a Multi-Population Assisted Quantum Genetic Algorithm. *Human Heredity*, 84(1):1–8.

Shi, W., Carr, M. J., Dunford, L., Zhu, C., Hall, W. W., & Higgins, D. G. (2012a). Identification of novel inter-genotypic recombinants of human hepatitis B viruses by large-scale phylogenetic analysis. *Virology*, 427(1):51–59.

Shi, W., Zhu, C., Zheng, W., Carr, M. J., Higgins, D. G., & Zhang, Z. (2012b). Subgenotype reclassification of genotype B hepatitis B virus. *BMC Gastroenterology*, 12(1):116.

Shiba, T. & Suzuki, Y. (1981). Localization of A protein in the RNA-A protein complex of RNA phage MS2. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, 654(2):249–255.

Shin, M.-K., Lee, J., & Ryu, W.-S. (2004). A Novel cis-Acting Element Facilitates Minus-Strand DNA Synthesis during Reverse Transcription of the Hepatitis B Virus Genome. *Journal of Virology*, 78(12):6252–6262.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., & Higgins, D. G. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1):539–539.

Simmonds, P. & Midgley, S. (2005). Recombination in the Genesis and Evolution of Hepatitis B Virus Genotypes. *Journal of Virology*, 79(24):15467–15476.

Simonsen, C. C. & Levinson, A. D. (1983). Analysis of processing and polyadenylation signals of the hepatitis B virus surface antigen gene by using simian virus 40-hepatitis B virus chimeric plasmids. *Molecular and Cellular Biology*, 3(12):2250–2258.

Sokal, R. R. & Michener, C. D. (1958). *A statiscal method for evaluating systematic relationships*. University of Kansas.

Starkman, S. E., MacDonald, D. M., Lewis, J. C., Holmes, E. C., & Simmonds, P. (2003). Geographic and species association of hepatitis B virus genotypes in non-human primates. *Virology*, 314(1):381–393.

Stockley, P. G., Ranson, N. A., & Twarock, R. (2013a). A new paradigm for the roles of the genome in ssRNA viruses. *Future Virology*, 8(6):531–543.

Stockley, P. G., Rolfsson, O., Thompson, G. S., Basnak, G., Francese, S., Stonehouse, N. J., Homans, S. W., & Ashcroft, A. E. (2007). A Simple, RNA-Mediated Allosteric Switch Controls the Pathway to Formation of a T = 3 Viral Capsid. *Journal of Molecular Biology*, 369(2):541–552.

Stockley, P. G., Twarock, R., Bakker, S. E., Barker, A. M., Borodavka, A., Dykeman, E., Ford, R. J., Pearson, A. R., Phillips, S. E. V., Ranson, N. a., & Tuma, R. (2013b). Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. *Journal of Biological Physics*, 39(2):277–87.

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques*, 28(6):1102–1104.

Strimmer, K. & von Haeseler, A. (1996). Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution*, 13(7):964–969.

Suárez-Díaz, E. & Anaya-Muñoz, V. H. (2008). History, objectivity, and the construction of molecular phylogenies. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, 39(4):451–468.

Sugauchi, F., Orito, E., Ichida, T., Kato, H., Sakugawa, H., Kakumu, S., Ishida, T., Chutaputti, A., Lai, C. L., Gish, R. G., Ueda, R., Miyakawa, Y., & Mizokami, M. (2003). Epidemiologic and virologic characteristics of hepatitis B virus genotype B having the recombination with genotype C. *Gastroenterology*, 124(4):925–932.

Sugauchi, F., Orito, E., Ichida, T., Kato, H., Sakugawa, H., Kakumu, S., Ishida, T., Chutaputti, A., Lai, C.-L., Ueda, R., Miyakawa, Y., & Mizokami, M. (2002). Hepatitis B virus of genotype B with or without recombination with genotype C over the precore region plus the core gene. *Journal of Virology*, 76(12):5985–92.

Sugiyama, T. & Nakada, D. (1967). Control of translation of MS2 RNA cistrons by MS2 coat protein. *Proceedings of the National Academy of Sciences of the United States of America*, 57(6):1744–1750.

Sugiyama, T. & Nakada, D. (1968). Translational control of bacteriophage MS2 RNA cistrons by MS2 coat protein: Polyacrylamide gel electrophoretic analysis of proteins synthesized in vitro. *Journal of Molecular Biology*, 31(3):431–440.

Suh, A., Brosius, J., Schmitz, J., & Kriegs, J. O. (2013). The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses. *Nature Communications*, 4(1):1791.

Summers, J. & Mason, W. S. (1982). Replication of the genome of a hepatitis

B-like virus by reverse transcription of an RNA intermediate. *Cell*, 29(2):403–415.

Sun, S., Rao, V. B., & Rossmann, M. G. (2010). Genome packaging in viruses. *Current Opinion in Structural Biology*, 20(1):114–120.

Sundram, A., Jumanla, N., & Ehlers, M. M. (2006). Genotyping of F-RNA coliphages isolated from wastewater and river water samples. *Water SA*, 32(1):65–70.

Suwannakarn, K. (2005). A novel recombinant of Hepatitis B virus genotypes G and C isolated from a Thai patient with hepatocellular carcinoma. *Journal of General Virology*, 86(11):3027–3030.

Suzuki, M. (1989). SPXX, a frequent sequence motif in gene regulatory proteins. *Journal of Molecular Biology*, 207(1):61–84.

Sweeney, B., Zhang, T., & Schwartz, R. (2008). Exploring the parameter space of complex self-assembly through virus capsid models. *Biophysical Journal*, 94(3):772–783.

Takyar, S., Hickerson, R. P., & Noller, H. F. (2005). mRNA helicase activity of the ribosome. *Cell*, 120(1):49–58.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution*, 9(4):678–87.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–26.

Tang, H. & McLachlan, A. (2002). A pregenomic RNA sequence adjacent to DR1 and complementary to epsilon influences hepatitis B virus replication efficiency. *Virology*, 303(1):199–210.

Tange, O. (2011). Gnu parallel-the command-line power tool. *The USENIX Magazine*, 36(1):42–47.

Tarjan, R. (1971). Depth- first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):114–121.

Tatematsu, K., Tanaka, Y., Kurbanov, F., Sugauchi, F., Mano, S., Maeshiro, T., Nakayoshi, T., Wakuta, M., Miyakawa, Y., & Mizokami, M. (2009). A genetic variant of hepatitis B virus divergent from known human and ape genotypes isolated from a Japanese patient and provisionally assigned to new genotype J. *Journal of Virology*, 83(20):10538–10547.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17:57–86.

Tavares, P., Harris, J. R., & Bhella, D. (2018). Protein-RNA Interactions in the Single-Stranded RNA Bacteriophages. In: *Virus Protein and Nucleoprotein Complexes (Subcellular Biochemistry)*, J. R. Harris & D. Bhella, ed., volume 88, chapter 13, pages 305–328. Springer.

Telford, M. J., Wise, M. J., & Gowri-Shankar, V. (2005). Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the bilateria. *Molecular Biology and Evolution*, 22(4):1129–1136.

Temin, H. M. (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Molecular biology and evolution*, 2(6):455–68.

Tillier, E. R. & Collins, R. A. (1998). High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, 148(4):1993–2002.

Tiollais, P., Pourcel, C., & Dejean, A. (1985). The hepatitis B virus. *Nature*, 317(6037):489–95.

Toropova, K., Basnak, G., Twarock, R., Stockley, P. G., & Ranson, N. A. (2008). The Three-dimensional Structure of Genomic RNA in Bacteriophage MS2: Implications for Assembly. *Journal of Molecular Biology*, 375(3):824–836.

Tuerk, C., Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510.

Tuthill, T. J., Harlos, K., Walter, T. S., Knowles, N. J., Groppelli, E., Rowlands, D. J., Stuart, D. I., & Fry, E. E. (2009). Equine rhinitis A virus and its low pH empty particle: clues towards an aphthovirus entry mechanism? *PLoS pathogens*, 5(10):e1000620.

Twarock, R., Leonov, G., & Stockley, P. G. (2018). Hamiltonian path analysis of viral genomes. *Nature Communications*, 9(1):2021.

Twarock, R. & Stockley, P. G. (2019). RNA-Mediated Virus Assembly: Mechanisms and Consequences for Viral Evolution and Therapy. *Annual Review of Biophysics*, 48(1):495–514.

Valegård, K., Liljas, L., Fridborg, K., & Unge, T. (1990). The three-dimensional structure of the bacterial virus MS2. *Nature*, 345(6270):36–41.

Valegård, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J., & Liljas, L. (1994). Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature*, 371(6498):623–626.

Valegård, K., Murray, J. B., Stonehouse, N. J., Van Den Worm, S., Stockley, P. G., & Liljas, L. (1997). The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions. *Journal of Molecular Biology*, 270(5):724–738.

Valentine, R. C. & Strand, M. (1965). Complexes of F-pili and RNA bacteriophage. *Science*, 148(3669):511–513.

Vandenberghe, A., Min Jou, W., & Fiers, W. (1975). 3'-Terminal nucleotide sequence (n = 361) of bacteriophage MS2 RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 72(7):2559–2562.

Vasquez, C., Granboulan, N., & Franklin, R. M. (1966). Structure of the ribonucleic acid bacteriophage R17. *Journal of bacteriology*, 92(6):1779–86.

Verrier, E. R., Yim, S. A., Heydmann, L., El Saghire, H., Bach, C., Turon-Lagot, V., Mailly, L., Durand, S. C., Lucifora, J., Durantel, D., Pessaux, P., Manel, N., Hirsch, I., Zeisel, M. B., Pochet, N., Schuster, C., & Baumert, T. F. (2018). Hepatitis B virus evasion from cyclic guanosine monophosphateadenosine monophosphate synthase sensing in human hepatocytes. *Hepatology*, 68(5):1695–1709.

Vincent, I. E., Zannetti, C., Lucifora, J., Norder, H., Protzer, U., Hainaut, P., Zoulim, F., Tommasino, M., Trépo, C., Hasan, U., & Chemin, I. (2011). Hepatitis b virus impairs tlr9 expression and function in plasmacytoid dendritic cells. *PLoS ONE*, 6(10):e26315.

Viñuela, E., Algranati, I. D., & Ochoa, S. (1967). Synthesis of virus-specific proteins in Escherichia coli infected with the RNA bacteriophage MS2. *European journal of biochemistry*, 1(1):3–11.

Waddell, P. J. & Steel, M. A. (1997). General Time-Reversible Distances with Unequal Rates across Sites: Mixing Γ and Inverse Gaussian Distributions with Invariant Sites. *Molecular Phylogenetics and Evolution*, 8(3):398–414.

Wang, G. H. & Seeger, C. (1992). The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. *Cell*, 71(4):663–670.

Wang, J. C.-Y., Dhason, M. S., & Zlotnick, A. (2012). Structural Organization of Pregenomic RNA and the Carboxy-Terminal Domain of the Capsid Protein of Hepatitis B Virus. *PLoS Pathogens*, 8(9):e1002919.

Wang, J. C.-Y., Nickens, D. G., Lentz, T. B., Loeb, D. D., & Zlotnick, A. (2014). Encapsidated hepatitis B virus reverse transcriptase is poised on an ordered RNA lattice. *Proceedings of the National Academy of Sciences of the United States of America*, 111(31):11329–34.

Wang, Z., Hou, J., Zeng, G., Wen, S., Tanaka, Y., Cheng, J., Kurbanov, F., Wang, L., Jiang, J., Naoumov, N. V., Mizokami, M., & Qi, Y. (2007). Distribution and characteristics of hepatitis B virus genotype C subgenotypes in China. *Journal of Viral Hepatitis*, 14(6):426–434.

Wang, Z., Liu, Z., Zeng, G., Wen, S., Qi, Y., Ma, S., Naoumov, N. V., & Hou, J. (2005). A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. *Journal of General Virology*, 86(4):985–990.

Ward, R., Shive, K., & Valentine, R. (1967). Capsid protein of f2 as translational repressor. *Biochemical and Biophysical Research Communications*, 29(1):8–13.

Ward, R., Strand, M., & Valentine, R. C. (1968). Translational repression of f2 protein synthesis. *Biochemical and Biophysical Research Communications*, 30(3):310–317.

Warnow, T. (2017). *Computational Phylogenetics*. Cambridge University Press.

Watanabe, I., Miyake, T., Sakurai, T., Shiba, T., & Ohno, T. (1967a). Isolation and grouping of RNA phages. *Proceedings of the Japan Academy*, 43(3):204–209.

Watanabe, I., Nishihara, T., Kaneko, H., Toshizo, S., & Osawa, S. (1967b). Group Characteristics of RNA Phages. *Proceedings of the Japan Academy*, 43(3):210–213.

Wayne, K. (2001). Princeton COS 423 Analysis of Algorithms Lectures. Available at: `http://www.cs.princeton.edu/{~}wayne/cs423/`. [Accessed on 2019-02-14].

Weiner, A. M. & Weber, K. (1971). Natural read-through at the UGA termination signal of Qβ coat protein cistron. *Nature New Biology*, 234(50):206–209.

Weissmann, C. & Borst, P. (1963). Double-stranded ribonucleic acid formation in vitro by MS2 phage-induced RNA synthetase. *Science*, 142(3596):1188–1191.

Wheeler, Q., Assis, L., & Rieppel, O. (2013). Phylogenetics: Heed the father of cladistics. *Nature*, 496(7445):295–296.

Wiese, K. C., Deschenes, A. A., & Hendriks, A. G. (2008). RnaPredict - An evolutionary algorithm for RNA secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(1):25–41.

Wiley, E. O. (1981). *Phylogenetics : theory and practice of phylogenetic systematics*. Wiley-Blackwell.

Will, H., Reiser, W., Weimer, T., Pfaff, E., Büscher, M., Sprengel, R., Cattaneo, R., & Schaller, H. (1987). Replication strategy of human hepatitis B virus. *Journal of virology*, 61(3):904–11.

Winter, R. B. & Gold, L. (1983). Overproduction of bacteriophage Qβ maturation (A2) protein leads to cell lysis. *Cell*, 33(3):877–885.

Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J., & Müller, T. (2008). ProfDistS: (profile-) distance based phylogeny on sequence - Structure alignments. *Bioinformatics*, 24(20):2401–2402.

World Health Organization (2017). Hepatitis B. Available at: `http://www.who.int/news-room/fact-sheets/detail/hepatitis-b`. [Accessed on 2018-05-31].

Wright, A. M., Lloyd, G. T., & Hillis, D. M. (2016). Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Systematic Biology*, 65(4):602–611.

Wu, J., Meng, Z., Jiang, M., Pei, R., Trippler, M., Broering, R., Bucchi, A., Sowa, J. P., Dittmer, U., Yang, D., Roggendorf, M., Gerken, G., Lu, M., & Schlaak, J. F. (2009). Hepatitis B virus suppresses toll-like receptor-mediated innate immune responses in murine parenchymal and nonparenchymal liver cells. *Hepatology*, 49(4):1132–1140.

Wu, Y. L., Shen, C. L., & Chen, X. Y. (2019). Antiviral treatment for chronic hepatitis B: Safety, effectiveness, and prognosis.

Wuchty, S., Fontana, W., Hofacker, I. L., & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65.

Wynne, S. A., Crowther, R. A., & Leslie, A. G. (1999). The crystal structure of the human hepatitis B virus capsid. *Molecular Cell*, 3(6):771–780.

Yang, J., Xing, K., Deng, R., Wang, J., & Wang, X. (2006). Identification of Hepatitis B virus putative intergenotype recombinants by using fragment typing. *Journal of General Virology*, 87(8):2203–2215.

Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, 13(5):303–314.

Ye, L., Zhang, Y., Mei, Y., Nan, P., & Zhong, Y. (2010). Detecting putative recombination events of hepatitis B virus: An updated comparative genome analysis. *Chinese Science Bulletin*, 55(22):2373–2379.

Young, I. & Coleman, A. W. (2004). The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects: A Drosophila example. *Molecular Phylogenetics and Evolution*, 30(1):236–242.

Zehender, G., Ebranati, E., Gabanelli, E., Sorrentino, C., Presti, A. L., Tanzi, E., Ciccozzi, M., & Galli, M. (2014). Enigmatic origin of hepatitis B virus: An ancient travelling companion or a recent encounter? *World Journal of Gastroenterology*, 20(24):7622–7634.

Zeng, G., Wang, Z., Wen, S., Jiang, J., Wang, L., Cheng, J., Tan, D., Xiao, F., Ma, S., Li, W., Luo, K., Naoumov, N. V., & Hou, J. (2005). Geographic distribution, virologic and clinical characteristics of hepatitis B virus genotypes in China. *Journal of Viral Hepatitis*, 12(6):609–617.

Zhou, B., Xiao, L., Wang, Z., Chang, E. T., Chen, J., & Hou, J. (2011). Geographical and ethnic distribution of the HBV C/D recombinant on the Qinghai-Tibet Plateau. *PLoS ONE*, 6(4).

Zlotnick, A. (1994). To build a virus capsid: An equilibrium model of the self assembly of polyhedral protein complexes. *Journal of Molecular Biology*, 241(1):59–67.

Zlotnick, A., Cheng, N., Conway, J. F., Booy, F. P., Steven, A. C., Stahl, S. J., & Wingfield, P. T. (1996). Dimorphism of Hepatitis B Virus Capsids Is Strongly Influenced by the C-Terminus of the Capsid Protein. *Biochemistry*, 35(23):7412–7421.

Zuber, J., Sun, H., Zhang, X., McFadyen, I., & Mathews, D. H. (2017). A sensitivity analysis of RNA folding nearest neighbor parameters identifies a subset of free energy parameters with the greatest impact on RNA secondary structure prediction. *Nucleic Acids Research*, 45(10):6168–6176.

Zuker, M. (1989). On Finding All Suboptimal Foldings of an RNA Molecule. *Science*, 244(4900):48–52.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415.

Zvelebil, M. J. & Baum, J. O. (2008). *Understanding bioinformatics*. Garland Science.