Universiteit Leiden

Master Computer Science

Intersegmental RNA-RNA interactions in influenza A virus genomes: a bioinformatics perspective

Name: Ruben A. Walen
Student ID: s1911724
Date: 30/06/2021

Specialisation: Bioinformatics

1st supervisor: Alexandre P. Goultiaev
2nd supervisor: Katherine J. Wolstencroft

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank my parents, Cees and Fatima, for their endless support in my academic endeavours.

My girlfriend Myrthe, who has advised me on many things, and has helped to keep me sane.

Last but not least, I would like to thank my supervisors Sacha and Katy for providing me with and supporting me in this extremely interesting and extraordinarily complex topic, in a time where it could not have felt more relevant.

PRELIMINARIES

Conventions

- This thesis is intended to be understandable for people with some background in bioinformatics and/or life sciences
- Page numbers in in-line citations are numbered from the first page of the article/book
- The inter-RNA interaction notation is as follows: *28 NP 670-712 MP 355-395* stands for: the interaction ranked 28+1=29th in the source dataset, between the NP segment at nucleotides 670-712 and MP segment at 355-395, measured from the 5' end. This can be abbreviated as: *2 NP MP*. The ranking is usually by reads-per-million (RPM) in (Dadonaite et al. 2019)

Abbreviations

- cf.: *confer*, meaning compare, or used to refer to relevant literature
- w.r.t.: with respect to
- vRNA: viral RNA
- vRNP: viral ribonucleoprotein: an RNA-protein complex containing a particular vRNA segment, viral nucleoproteins and viral polymerase subunits
- nt: nucleotide
- inter-RNA interaction: intersegmental RNA-RNA interaction
- strain names are sometimes abbreviated, e.g. A/California/07/2009 → California/2009
- vRNA segment abbreviations:

Segment common name/protein encoded	**Segment abbreviation**
Haemagglutinin	HA
Neuraminidase	NA
Polymerase A subunit	PA
Polymerase B1 subunit	PB1
Polymerase B2 subunit	PB2
Nucleoprotein	NP
Non-structural protein	NS
Matrix protein	MP

ABSTRACT

Genome packaging is the process by which a virus loads its genome into newly produced viral particles. The unique nature of the influenza genome complicates genome packaging: its genome is divided into eight unique segments, each carrying genes crucial for the proper functioning of the virus. There is experimental evidence that most influenza virions carry all eight segments with low multiplicity, a fact that cannot be explained using a random model of incorporation of genome segments. It is not known how influenza viruses manage to *selectively package* the unique genome segments into new virions in this manner, but experimental evidence points to the involvement of direct inter-RNA base-pairing interactions between genome segments. This thesis aims to build upon existing datasets on these inter-RNA interactions in strains of influenza A in order to extract novel information. By extrapolating the interaction set from a reference strain to other strains of influenza A, insights were gained into the potential conservation and similarity of these interactions on the sequence- and inter-RNA structure level. Limited evidence was found for conservation of inter-RNA interactions in general. In-depth analysis of a few specific interactions indicated the potential existence of some partly conserved sequence- and structural elements. The degree of similarity was lower in general in more genomically distinct strains. The findings underline the plasticity of inter-RNA interaction networks. This may have implications for selective packaging and reassortment processes in influenza.

KEYWORDS

Influenza, genome packaging, selective packaging, RNA structure, computational virology

INTRODUCTION

Overview of influenza

Influenza, known colloquially as ‘the flu’, is a contagious disease caused by influenza viruses. Influenza is responsible for a high annual burden of disease in humans, including many deaths (CDC 2020) (Iuliano et al. 2018). Besides humans, influenza viruses can cause disease in other animals, such as wild- and domesticated fowl and pigs. Influenza outbreaks in this context can have an impact on agriculture (Alders et al. 2014).

The symptoms of influenza mainly affect the respiratory system (El Ramahi and Freifeld 2019). Severe complications are possible, including viral pneumonia resulting in acute respiratory distress syndrome (Short et al. 2014). In severe influenza, bacterial secondary pneumonia also poses a significant risk (Beigel 2008, Sect. Complications) (Morris, Cleary, and Clarke 2017). Influenza can also lead to non-respiratory symptoms, and exclusively non-respiratory presentations are possible in a minority of cases (Chow et al. 2020). Although antiviral agents exist that are at least somewhat effective against influenza (Dobson et al. 2015) (Hsu et al. 2012), there are no definitive treatments for influenza (Beigel 2008, Sect. Antiviral Treatments). Besides antivirals, treatments are supportive and symptomatic in nature, including mechanical ventilation for severe influenza-associated pneumonia (Ríos et al. 2011). Influenza vaccines are effective in preventing or controlling influenza, but challenges remain in the development and maintenance of influenza vaccines and vaccination strategies (Houser and Subbarao 2015). The evolution of resistance to antivirals is a growing problem in influenza treatment (Lampejo 2020).

Owing to its respiratory-centric manifestations, influenza spreads mainly through respiratory droplets (Weinstein et al. 2003), and possibly through aerosols (Tellier 2006) (Cowling et al. 2013), although controversy exists about the latter (Lemieux et al. 2007). Transfer through fomites (objects or surfaces contaminated by virus) may also be possible (Weinstein et al. 2003) (Boone and Gerba 2007). In the Western hemisphere, influenza in humans is most prevalent in winter months, but nearer to the tropics, outbreaks happen year-round (Viboud, Alonso, and Simonsen 2006) (Moura 2010). Influenza has also been responsible for several pandemics, including the 1918 pandemic (also known as ‘Spanish flu’, although the term is now considered improper), which killed about 50 million people globally (Taubenberger and Morens 2006). More recently, an influenza virus caused the 2009 swine flu pandemic. The propensity of influenza to cause recurrent outbreaks, both seasonally and in the context of pandemics, is due to several factors. Its relatively high mutation rate, multiplicity of variants and global spread in combination with maintenance in animal reservoirs in close proximity to humans and tendency towards ‘spillover’ events establishes this virus as one of the most persistent infectious disease threats to human health (Shao et al. 2017) (Lyons and Lauring 2018).

CLASSIFICATION AND NOMENCLATURE

Influenza viruses are single-stranded, negative-sense RNA viruses (i.e. -ssRNA, or Class V in the Baltimore classification) (Bouvier and Palese 2008). This means that the influenza genomic viral RNA must first be transcribed to opposite sense (positive sense) mRNA in other to produce viral proteins. Influenza viruses are enveloped, meaning that the outer layer of the viral particles consists of host cell membrane lipids, with antigenic spike proteins protruding. Inside the virion, matrix proteins coat the viral envelope. Influenza viruses package a multi-subunit viral RNA polymerase which is capable of both transcribing mRNA from the viral genomic RNA, and replicating the genomic RNA via a 'cRNA' intermediate (te Velthuis and Fodor 2016) (Park et al. 2003).

Influenza viruses belong to the family *Orthomyxoviridae*, which contains the influenza genera or species *A, B, C* and *D* (Su et al. 2017), along with a few other non-influenza viruses. Of these, influenza A and B are most common in humans (Belshe 2010). Influenza A is the main cause of influenza pandemics and is also common in seasonal/annual outbreaks (Layne, Monto, and Taubenberger 2009) and it will be the focus of this thesis. Influenza A is divided into *serovars* or *subtypes* based on the identity of its immunogenic haemagglutinin (HA) and neuraminidase (NA) spike proteins (Bouvier and Palese 2008), yielding the common notation of *HxNy* (e.g. H1N1 and H5N1), where *x* and *y* note the haemagglutinin and neuraminidase subtypes respectively. Each subtype is then further subdivided into various strains based on more precise genetic differences, yielding notations such as *A/California/7/2009(H1N1)pdm* for a particular strain of 2009 swine flu pandemic influenza A virus found in California.

INFLUENZA GENETICS

One particular aspect of influenza genomics drives the generation of new strains of influenza virus in a manner that is uncommon among viruses: its segmented genome. The influenza A genome, which encodes up to 14 genes depending on the specific strain (Eisfeld, Neumann, and Kawaoka 2015), is 'spread out' across eight different single-stranded, negative-sense RNA segments (Bouvier and Palese 2008). Each of these segments is unique and does not share genes with other segments. This means that each of the segments is necessary for the correct functioning and replication of influenza virus (Bouvier and Palese 2008). A missing segment(s) means the virus is unable to produce one or more of its viral proteins, resulting in partially- or non-functional virus. The segmented nature of influenza genomes (which it shares with other members of the *Orthomyxoviridae* family) is not unique among *-ssRNA* viruses (Lowen 2018), but the multiplicity of segmentation (eight segments) in influenza viruses is high relative to most other segmented viral genomes (Chaitanya 2019) (McDonald et al. 2016, Tab. 1), including *Bunyavirales-* and *Arenaviridae* segmented genomes which usually have two or three unique segments. The segmentation of influenza virus genomes complicates viral replication (Hutchinson et al. 2010), as each of the unique segments replicates independently, but all eight need to be packaged together into a new viral particle.

Table 1: Overview of the eight influenza A viral RNA (vRNA) segments and known encoded proteins. The lengths in nucleotides of each vRNA segment is given based on the A/WSN/1933[H1N1] strain, available for example in the Influenza Research Database (fludb.org). Encoded proteins are based mostly on (Bouvier and Palese 2008), with additional citations to support in virio functions. An overview of the viral polymerase (with PA, PB1, PB2 subunits) is available in (te Velthuis and Fodor 2016), although the individual subunits may have other functions as well. The list of proteins encoded may not be exhaustive, as new viral protein functions are discovered occasionally.

Segment abbreviation	**Length (nt)** *A/WSN/1933[H1N1]*	**Proteins encoded**	**Functions**
PB1	2341	PB1 (polymerase B1 subunit)	Viral polymerase basic subunit 1
		PB1-F2	Pro-apoptotic function and interferon inhibition (Varga et al. 2012)
PB2	2341	PB2 (polymerase B2 subunit)	Viral polymerase basic subunit 2
PA	2233	PA (polymerase A subunit)	Viral polymerase acidic subunit (te Velthuis and Fodor 2016)
		PA-X	Endonucleolytic activity, immune suppression? (Bavagnoli et al. 2015) (Hayashi, MacDonald, and Takimoto 2015)
HA	1775	HA (haemagglutinin)	Surface antigen with receptor binding and cell entry functions (Kosik and Yewdell 2019)
NP	1565	NP (nucleoprotein)	Binds viral RNA (Hu et al. 2017)
NA	1409	NA (neuraminidase)	Surface antigen with glycoprotein cleavage capabilities, viral release functions (Kosik and Yewdell 2019)
MP *or* M	1027	M1 (matrix protein 1)	Coats viral envelope. Involved in viral budding, viral ribonucleoprotein export, and other functions, cf. e.g. (Ruigrok et al. 2000) (Bui et al. 2000)
		M2 (matrix protein 2)	Proton channel. Regulates pH during cell entry/viral maturation (Pielak and Chou 2011)
NS	890	NS1 (non-structural protein 1)	Blocking host immune response, interferes with host mRNA processing, promote viral mRNA, various other functions (Hale et al. 2008)
		NS2/NEP (nuclear export protein)	Facilitates viral ribonucleoprotein export from nucleus (Neumann, Hughes, and Kawaoka 2000)

The segmented nature of the influenza genome enables a particular mechanism of recombinational genetic transfer among different viral strains, known as *reassortment* (Lowen 2018). Reassortment in influenza occurs when two or more strains of influenza co-infect one cell, resulting in the co-localisation of viral RNA (vRNA) segments from different strains. In this case, segments from these differing strains can blend to form new combinations of segments. These new recombinant strains may have properties different from the underlying strains, such as greater adaptability to the host and the ability to dodge host immune responses to existing strains due to the expression of a new combination of the two surface antigens HA and NA, a phenomenon known as *antigenic shift* (Lowen 2018) (Webster and Govorkova 2014). *Antigenic shift* refers to a recombination event between different strains of a virus leading to the generation of a new combination of its antigens (in this case, the HA and NA antigens) (Webster and Govorkova 2014). Antigenic shift necessitates on one hand

the existence of multiple viral antigens, and on the other hand, a mechanism for such recombination events to occur. This contrasts with *antigenic drift*, which is the change such viral antigens may undergo as a result of normal mutational processes. Influenza A undergoes both antigenic shift and antigenic drift processes, the former through reassortment of the separate vRNA segments coding for the HA and NA antigens. Note however that reassortment of the HA- and NA-encoding vRNA segments are not the only relevant reassortment events, although such antigenic shift events are commonly focused on in literature.

Such reassortment events may have been responsible for past outbreaks, including major pandemics such as the 2009 swine flu pandemic (Tao, Steel, and Lowen 2014), which likely first arose from reassortment events in farm pigs from what were originally avian-, swine- and human influenza strains of different serotypes (Smith et al. 2009). Another area where the potential for reassortment is a major concern are outbreaks of H5Nx avian influenza in fowl (Nuñez and Ross 2019). H5Nx-serotype strains (including strains of the well-known H5N1 and H5N8 serotypes) cause outbreaks mainly in wild- and domesticated fowl and have spread globally. This includes so-called highly pathogenic avian influenza (HPAI) strains, which are thought to be much more deadly to humans than the widely circulating human influenza A strains. However, these strains are not currently known to be able to spread sustainably from human to human due to their generally low infectivity in humans. Fears exist that a highly infectious (among humans) novel strain of HPAI H5Nx influenza may arise from a reassortment of low-infectivity HPAI H5Nx with a strain that is more adaptable to human infection (Nuñez and Ross 2019). A more human-like intermediate host such as swine may serve as a ‘stepping stone’ in such reassortment processes (Ma, Kahn, and Richt 2008).

Due in part to the recognised threat of reassortment in generating novel strains of influenza, monitoring systems for detecting such new strains have been put in place (WHO 2021). Predicting the likelihood of specific reassortment events happening between specific strains would be highly useful in contributing to such efforts, but our understanding of the mechanisms behind influenza reassortment and the consequences for reassortment probabilities is still quite basic (Gerber et al. 2014) (Vijaykrishna, Mukerji, and Smith 2015). Many studies are now pointing to the influence of influenza *genome packaging* on the reassortment process, see e.g. (Essere et al. 2013) (Gerber et al. 2014) (Vijaykrishna, Mukerji, and Smith 2015) (Hutchinson et al. 2010).

The process of ensuring that each unique influenza vRNA segment is represented in newly produced viral particles during replication is known as *genome packaging* (Hutchinson et al. 2010), although the term also refers more generally to the process of packaging viral genetic material into new particles (Sun, Rao, and Rossmann 2010). There has been much discussion about the nature of genome packaging in influenza (Shafiuddin and Boon 2019). The two main models proposed are *random packaging* and *selective packaging*. It is known that all influenza vRNA segments associate with viral nucleoprotein (NP) and polymerase units to form *viral ribonucleoprotein* (vRNP) *complexes* (te Velthuis and Fodor 2016) (Noda et al. 2012). In the *random packaging model*, vRNA segments are incorporated at random into new viral particles, i.e. the vRNA sequences themselves do not affect the packaging process or the probability of being included in a given *constellation* of vRNPs. In this model, genome packaging rests only on the ability to discern viral RNA to be packaged from

non-viral RNA (i.e. native mRNA), and from opposite sense viral RNA transcripts which are created during RNA polymerisation. A simple calculation shows that random selection with replacement of eight such vRNPs would result in a probability of approximately 0.24% of correctly including all eight unique segments in the complex (Hutchinson et al. 2010):

Equation 1: Naïve calculation of the probability of including all eight unique vRNA segments in a random selection with replacement of eight segments in e.g. cytoplasm of an infected cell.

$$P_{random} = \frac{8!}{8^8} \approx 0.0024$$

Although the applicability of this calculation is limited, this number seems implausibly low, and indeed the efficiency of generation of complete influenza particles has been found to be higher (Shafiuddin and Boon 2019) (Hutchinson et al. 2010). The probability of packaging at least one of each unique vRNA could in principle be increased by packaging a larger number of (random) segments into each virion (Hutchinson et al. 2010 p. 3). However, further evidence suggests that generally, just one of each unique vRNA is included in influenza particles (Shafiuddin and Boon 2019) (Gerber et al. 2014). This evidence includes direct observations of vRNPs in influenza particles using electron microscopy, see e.g. (Noda et al. 2012). Additionally, influenza is probably able to initiate infection at low multiplicity (few initial viruses), thus indicating that a model in which multiple partially complete virions infect a cell to complement missing segments (a *multipartite* virus model) is unlikely (Hutchinson et al. 2010, p. 3). This points to a model of *selective packaging*, in which the vRNAs/vRNPs themselves are involved in the packaging process. In this model, the vRNAs/vRNPs interact in order to ensure that each unique vRNA segment is incorporated (at least or only once) into budding virions. This selective packaging model is also thought to be critical in driving reassortment probabilities in the scenario of cell co-infection by different influenza strains (Gerber et al. 2014). Several mechanisms for selective packaging have been proposed.

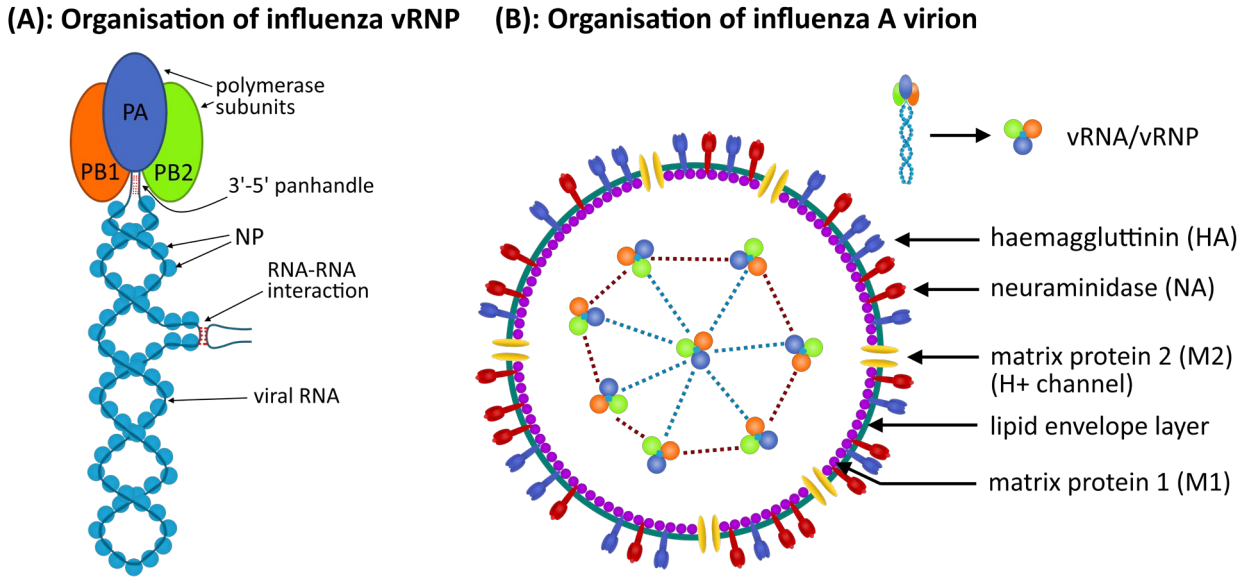


Figure 1: Organisation of influenza vRNP and virions. ***(A):*** *The organisation of a viral RNA segment (vRNA) into viral ribonucleoprotein (vRNP). Nucleoprotein wraps the vRNA strand, which folds into a helical shape. The vRNA helix is held together at the 3' and 5' ends through base pairing, forming a 'panhandle structure'. Viral polymerase subunits bind to this panhandle structure. A potential inter-RNA interaction with an external vRNA element is also shown.* ***(B):*** *The organisation of an influenza A virion. The viral envelope consists of host cell lipids and other membrane components, with viral M1 matrix protein coating the inside of the envelope. Not shown is that the M1 protein may also bind vRNA/vRNP during the budding process. Also shown are the M2 viral proton channel and neuraminidase (NA) and haemagglutinin (HA) spike proteins in the viral envelope. Inside the virion, the eight different vRNPs form an interacting network. Interactions between vRNPs are likely responsible for selective packaging in influenza A. The nature of these interactions is suspected to involve direct inter-RNA interaction. This network of vRNP interactions is thought to occur in a '7+1'-pattern, with one central vRNP. It is possible that each vRNP interacts with this central vRNP, shown as blue dotted lines. Another possibility is that the peripheral vRNPs interact with their neighbours, shown as crimson dotted lines. The precise identity of the vRNPs in this pattern is not known, and it is also not known whether this organisation is static.*

GENOME PACKAGING IN INFLUENZA

In the *selective packaging* model of influenza genetics, the virus needs to be able to discern vRNA on two levels: firstly, to distinguish viral (negative sense) genomic vRNA from opposite sense viral RNA and other cellular RNAs, and secondly, to distinguish individual vRNAs from one another (Gerber et al. 2014 p. 6). Elements of the viral genomic sequences that enable these two levels of discrimination are called *packaging signals*. A general packaging signal required for recognition of the negative (genomic) sense vRNA during replication has been known for several decades, cf. (Hutchinson et al. 2010, p. 7). Partially complementary sequences are present in the 3' and 5' ends of all influenza A vRNA segments, which causes the ends to bond and the segment to adopt a 'panhandle structure', i.e. the interior of the segment forms a large loop with the two ends binding through partial complementarity of the general packaging signals on both ends. This panhandle is known to associate with viral proteins even before the budding stage, mainly viral polymerase subunits (te Velthuis and Fodor 2016). The panhandle and the rest of the viral RNA also bind viral nucleoprotein (NP), which has many functions in replication (Hu et al. 2017), but is not known to bind specifically to certain vRNA elements (Williams et al. 2018). The resulting vRNP complex consisting of vRNA and viral proteins undergoes transport and changes in a complicated series of steps from

initial replication to budding of new virions, and from cell entry to replication upon infecting a new cell, cf. e.g. (Eisfeld, Neumann, and Kawaoka 2015).

The *partial* complementarity of the general packaging/panhandle formation signal means that the opposite (positive) sense complementary RNA which is synthesised from the viral genomic (negative sense) RNA by viral RNA polymerase during replication does not form the exact same panhandle binding interaction, which allows for discrimination of the correct sense viral RNA from both non-viral RNA and opposite sense viral RNA. The mechanism of exclusion for the wrong sense viral RNA is thought to be nuclear export, i.e. such wrong sense viral RNA material is not exported from the nucleus whilst the correct sense vRNA segments are. Note that this first level of discrimination is required in both the random packaging model and the specific packaging model, and such packaging signals are common among viruses in general (Coffin, Hughes, and Varmus 1997) (Masters 2019).

Although evidence is mounting that selection on the basis of unique vRNA segments is crucial to influenza genome packaging, it is not known how exactly this happens. Panhandle recognition and prevention of nuclear export is likely the mechanism of discrimination for general packaging, but these mechanisms are not enough to explain the discrimination between individual segments, since the panhandle packaging region is highly conserved for all segments (even among different strains) (Gerber et al. 2014, p. 1), and all vRNPs are exported from the nucleus (Hutchinson et al. 2010, Fig. 2). Non-sequence features that could affect selective packaging include discrimination on the basis of segment length (Shafiuddin and Boon 2019, p. 3), and self-repulsion (Venev and Zeldovich 2013). However, several of the influenza vRNA segments have similar lengths and corresponding vRNP sizes, so this alone would not be enough to explain selective packaging. Self-repulsion of identical segments is reported to have an effect on selective packaging (Venev and Zeldovich 2013), but it is not known how such self-repulsion would arise between these identical segments, but not between different segments (Gerber et al. 2014, p. 4).

Protein-RNA interactions could drive selective packaging through selective binding of unique vRNAs/vRNPs, perhaps in different binding sites depending on vRNA sequence elements (Shafiuddin and Boon 2019, p. 3-4), which could form the basis of a supramolecular complex packaging such selective proteins with the different vRNAs. However, no such vRNA-discrimination selective viral protein is known in influenza (Gerber et al. 2014, p. 3). The most obvious choice, viral nucleoprotein (NP), is thought to bind to the vRNA in a non-specific and semi-regular manner (Williams et al. 2018). Recently, some evidence was found that nucleoprotein binding is probably not random and not uniform, leading to speculations that it might also play a role in (specific) packaging (Le Sage et al. 2018). Influenza NP also seems to possess only one RNA binding groove (Ye, Krug, and Tao 2006) (Tarus et al. 2012), although this is not completely certain (Labaronne et al. 2016). Other viral influenza proteins are either not known to bind to vRNA at all (Hutchinson et al. 2010, p. 10), or bind non-specifically, such as the panhandle-structure-binding polymerase subunits.

OPEN QUESTIONS

It is becoming increasingly apparent, both through the elimination of other hypotheses and the appearance of new studies, that RNA-RNA interactions are involved in influenza selective packaging. Several reviews are available that deal with the lines of evidence and the RNA-genomic locations of selective packaging sites that may undergo such interactions, cf. e.g. (Hutchinson et al. 2010; Gerber et al. 2014; Shafiuddin and Boon 2019; Li et al. 2021). However, little is known about the nature of such interactions: does it involve direct binding between RNA nucleotides in different vRNAs, or does the interaction involve some kind of intermediate target, such as a protein or RNA structure? Some regions involved in the intersegmental RNA interactions are known, but would all of these be involved in such direct RNA-RNA interactions? At what stage of replication do interactions form, and how stable are they? Do the interactions constitute a network involving all eight unique vRNA segments, and if so, what kind of organisation underlies this network? Are these interactions, and the underlying RNA elements, conserved in different strains and different serotypes, or even in other influenza species? This is just a grasp from some of the open questions on the topic of selective packaging in influenza, the answers to which will provide key insights into the mechanisms of packaging and reassortment in influenza, yielding information on reassortment risks and potentially opening avenues for future therapies.

With this thesis, I aim to provide some insight into recent data on RNA-RNA interactions in influenza packaging from a bioinformatics perspective. The novelty and rapidly developing nature of this field necessitate new approaches, and key insights may be hidden in data that is already available. Chiefly, I aim to use bioinformatics-based techniques to analyse in-depth the data on possible direct RNA-RNA interactions in influenza A that have recently been described. I will mainly be looking at possible conservation of these inter-RNA interactions among different strains and serotypes of influenza A, both in terms of sequence conservation of the RNA elements, and structural conservation of the interactions. Relevant research questions include: are the recently discovered direct inter-RNA interactions in influenza A conserved among different strains and serotypes? If so, is this conservation mainly sequence-based or structure-based? Are some interactions more or less conserved than other interactions, and does this imply that certain vRNAs are more important than others for maintaining selective packaging-involved RNA-RNA interactions?

RNA-RNA interactions in influenza A

RNA STRUCTURE

The central dogma of molecular biology states that DNA is used as a template to make RNA, which is then used as a template to make proteins. Proteins then carry out most of the functions an organism needs to thrive. However, DNA and RNA are more than just templates for transcription and translation. Both molecules can adopt rich structures via base pairing and folding, by which they can be involved in various mechanisms useful to organisms and viruses. In RNA viruses, RNA structure is an indispensable factor in various viral functions, so much so that the evolution of such viruses may be constrained by conservation of RNA structures (Smyth et al. 2018). For influenza, the 3'-5' UTR panhandle structure which forms in each of the vRNAs is an example of a functional RNA structure, which in this case is important for viral packaging and complexing with the viral polymerase subunits. However, there are other (potentially conserved) RNA structures in the influenza genome, although not much is known about their functions, cf. (Gultyaev, Fouchier, and Olsthoorn 2010) (Ferhadian et al. 2018).

Importantly, these are all examples of RNA structures formed by *intra-RNA interactions* (or *cis*-(RNA-RNA) interactions), meaning that these structures are formed when nucleotides on the same RNA molecule form base pairs. These *cis*-interactions can occur on a short range to form localised RNA structures, but long-range *cis*-interactions are also known. Such long-range *cis*-interactions are found not just in influenza, but are widespread across RNA viruses, where they are involved in critical viral functions, cf. (Nicholson and White 2014). These interacting RNA elements are on the same molecule, but they can lie thousands of nucleotides apart, meaning that complex regulatory mechanisms, perhaps involving large-scale RNA structure and competing RNA elements, may be involved in getting such interactions to occur at the right stages of viral life cycles (Nicholson and White 2014).

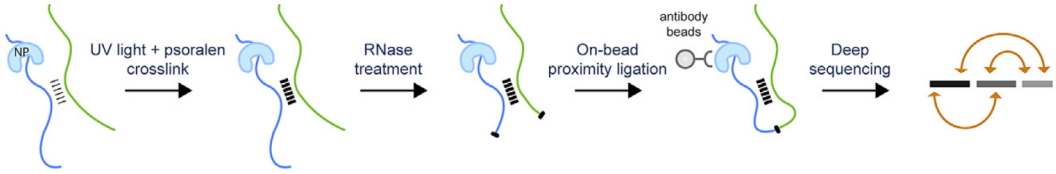
RNA-RNA INTERACTIONS IN INFLUENZA A

To explain selective packaging phenomena however, we need to study the potential for intersegmental *inter-RNA interactions* (or *trans*-(RNA-RNA) interactions). Direct *inter*-RNA interactions in the influenza genome would mean base pairing between nucleotides in different vRNA segments. Finding such interactions could mean that the vRNA sequences themselves are at least partly responsible for selective packaging in influenza, and that the conservation or lack thereof would probably be critical in reassortment compatibility between influenza strains. As discussed before, evidence has been mounting that some form of direct intersegmental RNA-RNA interaction is indeed involved in selective packaging in influenza, but up until recently, no techniques existed to study these potential interactions. In 2012 and in 2013, two studies showed using mutagenesis and electron tomography that certain regions of vRNAs in particular strains are involved in interactions with other vRNAs (Fournier et al. 2012; Gavazzi, Isel, et al. 2013), and pinpointed the specific nucleotides that could be involved in direct base pairing between vRNAs. Later in 2013, another study showed evidence of a functional inter-RNA interaction between two vRNA elements in the PB1 and NS segments of a H5N1 influenza A strain (Gavazzi, Yver, et al. 2013). Non-complementary

mutation of nucleotides in the vRNA elements responsible for this interaction greatly reduced packaging efficiency. The *in vitro* nature of this study limited its virological relevance, but this was a major step towards proving that direct inter-RNA interaction occurs in influenza, and that such interactions are important for selective packaging.

To conclusively demonstrate the existence of such intersegmental RNA-RNA interactions, *in virio* studies were conducted in recent years. Two seminal studies by (Dadonaite et al. 2019) and (Le Sage et al. 2020) were able to prove the existence *in virio* of hundreds such inter-RNA interactions between all vRNAs in various strains of influenza A. The data from these studies will be the focal point of analysis for this thesis. These studies used two novel molecular methods to capture these inter-RNA interactions in the *in virio* context: SPLASH for (Dadonaite et al. 2019) and 2CIMPL for (Le Sage et al. 2020). The 2CIMPL method was developed for the Le Sage study, whilst the SPLASH technique was described in (Aw et al. 2016). Although there are important differences between these methods, they are similar in that they both employ a crosslinking technique to stabilise the otherwise 'fragile' RNA-RNA base pairing, and then use sequencing followed by a mapping algorithm to determine the genomic locations of the interacting RNA elements. See Figure 2 for a schematic overview adapted from the respective studies. Importantly, 2CIMPL also incorporates an immunoprecipitation step targeting viral nucleoprotein (in this specific study), but this did not lead to bias in favour of high-nucleoprotein RNA regions according to the authors. A comparison of results between the two studies in the (Le Sage et al. 2020) paper revealed some major differences in terms of interactions found, even in highly similar strains. The reasons for these discrepancies are not clear yet. A review of inter-RNA interactions (*trans*-RNA-RNA interactions) in influenza A and other segmented RNA viruses is available, including evidence from the Dadonaite et al. study, cf. (Newburn and White 2019).

(A) *Le Sage et al., 2020: 2CIMPL*



(B) *Dadonaite et al., 2019: SPLASH*

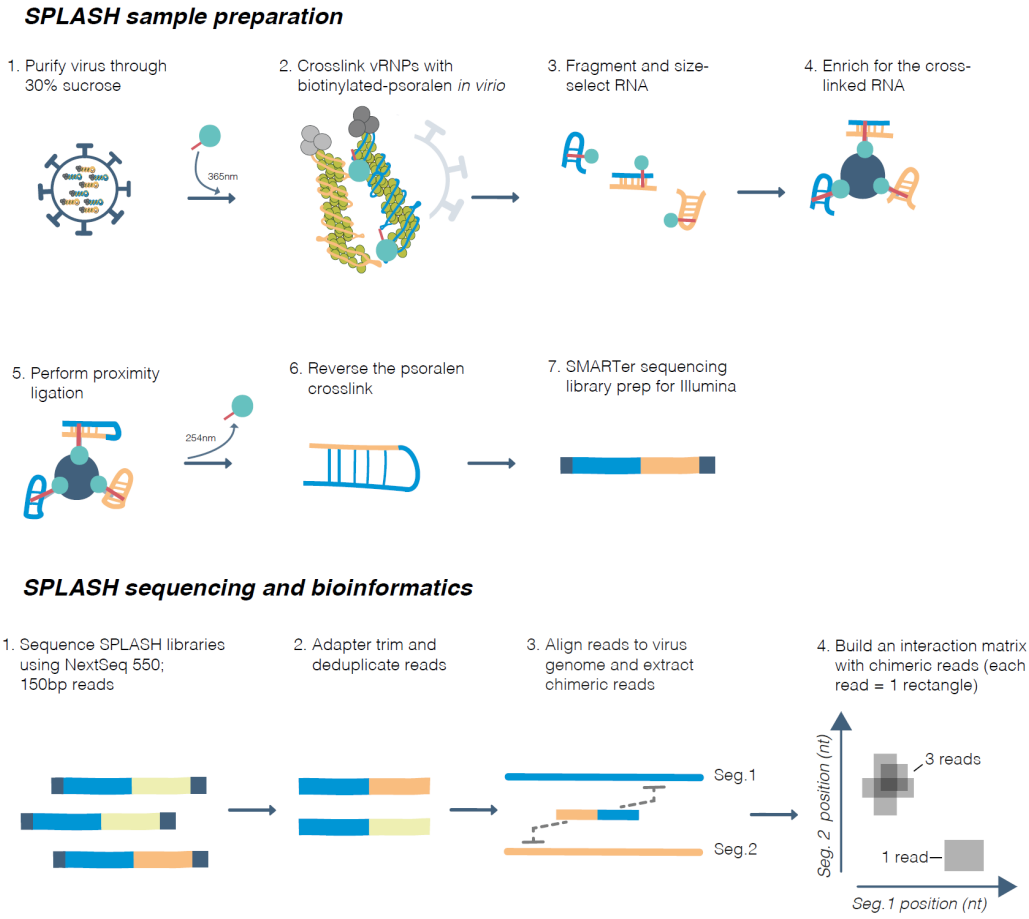


*Figure 2: Schematic overviews of Le Sage et al. 2CIMPL method and Dadonaite et al. SPLASH method for detecting inter-RNA interactions in influenza A. **(A)**: 2CIMPL method, from: (Le Sage et al. 2020, Figure 1). **(B)**: SPLASH method, from: (Dadonaite et al. 2019, Supplement Figure 7). Both use a crosslinking technique on vRNA bound in vRNP in virio. The crosslinked interactions are isolated using RNase digestion and proximity ligation. The 2CIMPL method then includes an immunoprecipitation step targeting viral nucleoprotein (NP) to isolate interacting vRNA elements. Both methods employ a bioinformatics-based mapping algorithm on the 'chimeric' reads to determine the genomic locations of the two interacting RNA elements.*

These studies showed that a complex network of inter-RNA interactions exists in various strains of influenza A. Mutagenesis experiments in these studies revealed a significant role of these interactions in selective packaging, and potential consequences for reassortment compatibility. A built-in degree of redundancy in these networks was also noted, which could be useful in case some interactions are broken by natural processes such as mutation. Remarkably, some of these redundant interactions were not reproduced in virtually identical replicates between (Dadonaite et al. 2019) and (Le Sage et al. 2020), or even in separate replicates of the same strains within these studies. Of course, this is just the first step in what is needed to determine the extent, timing, and functions of these inter-RNA interactions. Importantly, these methods are able to pinpoint to high accuracy the regions involved in such an inter-RNA interaction, but they cannot determine exactly which nucleotides within the interacting regions are involved in binding and what opposite-strand nucleotides they bind to. In other words, these methods are not able to determine the structure of an inter-RNA interaction. In the next section, I will explain how inter-RNA interaction structures can be predicted, and how interactions can be extended to other strains based only on sequence information.

METHODS

In this section, I will describe the experiments and general methods that were developed as part of this thesis. All code written for this thesis project is available along with documentation in a Git repository at: https://gitlab.com/ruben.walen/infla-rna-interactions. The repository also contains results of experiments performed.

Pipeline and experiments

Various separate experiments were performed to analyse the inter-RNA interaction data from several different perspectives. The data from (Dadonaite et al. 2019) and (Le Sage et al. 2020) serves as the basis for this thesis: they provide information on the genomic locations (involved segments and specific nucleotide ranges) of inter-RNA interactions occurring in multiple strains of influenza A. The interaction extrapolation step described below is a key prerequisite to most of these experiments, because it allows for the extension of inter-RNA interaction data from one reference strain to a set of new (unanalysed) strains. This enables analysis of the interacting RNA regions on both the sequence level and the structure level, the latter requiring a further inter-RNA structure prediction step using the *intaRNA* algorithm. A general graphical overview of the data processing- and analysis pipeline is shown in Figure 3.

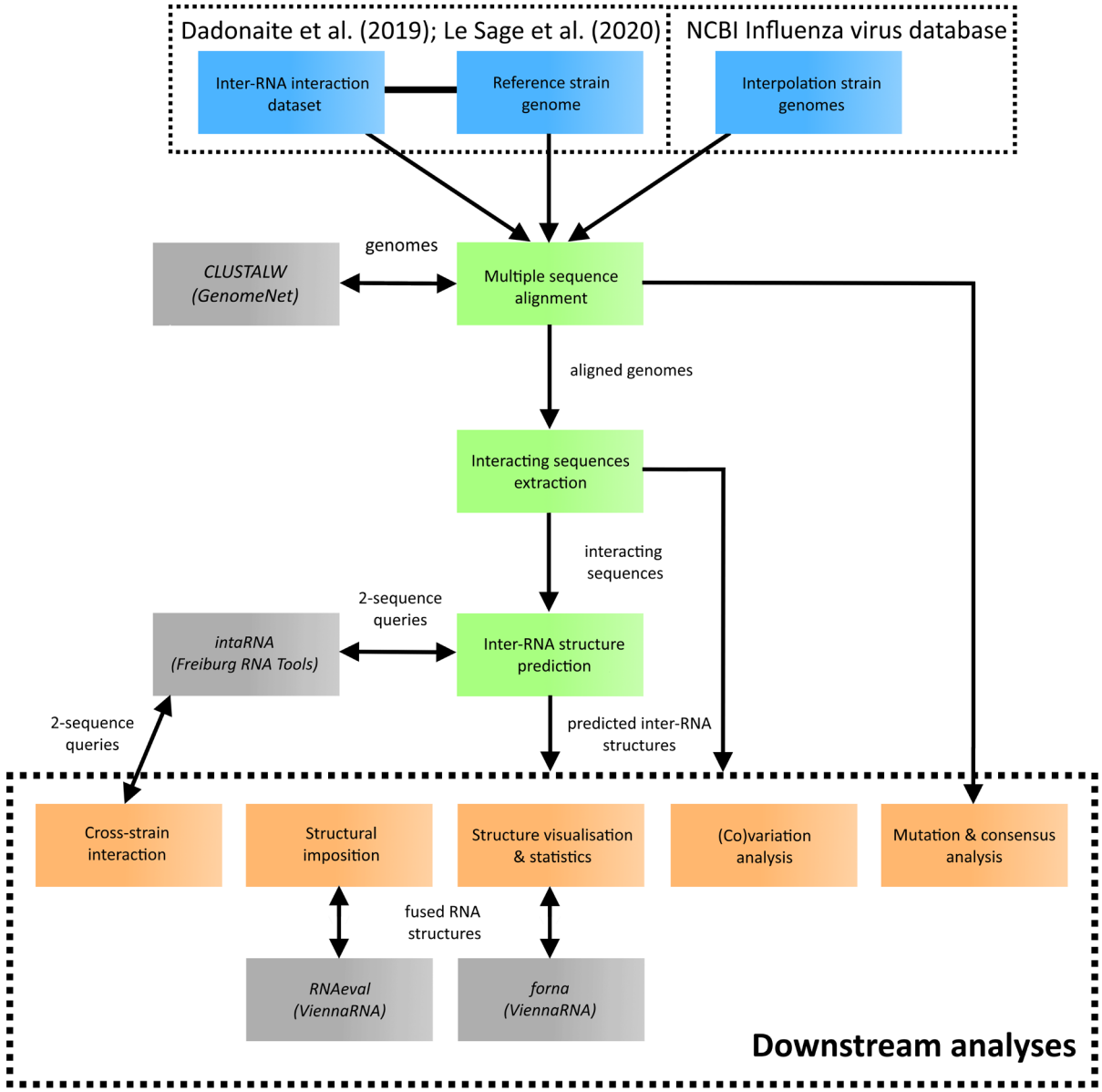


Figure 3: General overview of the data processing- and analysis pipeline for this thesis. ***Blue:*** *data sources;* ***green:*** *processing steps;* ***orange:*** *analysis steps;* ***grey:*** *external tools used.*

INTER-RNA INTERACTION EXTRAPOLATION

I developed a method to re-analyse the *in virio* inter-RNA interaction data from one strain (the *reference strain*) using the genomic sequence of another strain. The method is as follows:

- multiple sequence alignment (using *GenomeNet CLUSTALW, genome.jp*) of the new strains to the reference strain, for each vRNA segment
- use the genomic location data for inter-RNA interactions, which was derived experimentally for the reference strain in either Dadonaite et al or Le Sage et al., to generate pairs of sequences of interacting RNA elements for each strain (including the reference strain) for each interaction
- execute the *intaRNA* algorithm on every pair of interacting RNA elements for each strain and gather the results: the minimum free energy inter-RNA interaction structure and the corresponding free energy value, if a significant interaction was found

This method relies on extrapolating the experimentally derived genomic locations of discrete inter-RNA interactions in one strain to other strains based on multiple sequence alignment of vRNA segments. This enables comparative analysis of these interacting RNA elements on a sequence level, and on a predicted structure level. The major limitation of this approach is that there is no guarantee that the inter-RNA interactions in the reference strain can be extended to the same (homologous) genomic location in other strains. (Dadonaite et al. 2019) and (Le Sage et al. 2020), and the additional analysis of their data in the Results section of this thesis, showed that inter-RNA interactivity can differ greatly between strains and even between replicates of the same strains. Thus, it is highly unlikely that extrapolating interactions from one reference strain to a new strain will yield a fully accurate profile of inter-RNA interactions in the new strain, especially if the strains have greatly differing genomes or even different serotypes. What it can be used for is to determine to which degree the inter-RNA interacting elements are *conserved* or *similar* between strains on a sequence level and on a predicted structure level, whether some interactions are more conserved than others, and what implications this could have on genome packaging and reassortment in influenza A.

The method uses a *reference strain*, which supplies the reference viral genome and the genomic locations of inter-RNA interactions in that strain. The genomic location of an inter-RNA interaction consists of the start- and end positions of the interacting RNA regions on both interacting vRNA segments. No additional information, such as indications of RNA structure, is supplied or required, although the number of reads per million (RPM) in Dadonaite et al. data and the number of occurrences in Le Sage et al. data can be useful as an indicator of the frequency of occurrence of each discrete interaction. Using multiple sequence alignment, analogous RNA regions are then derived in the new strains, where the same inter-RNA interaction would occur as in the reference strain if it were conserved. Then, an inter-RNA structure prediction algorithm is applied to derive the minimum free energy (MFE) structure (and corresponding free energy value) of those interacting RNA regions. Here, Freiburg RNA Tools *intaRNA* will be used, as it quite modern and is useful in this context, and it was also employed in the Dadonaite et al. study, albeit with different (non-traceable) parameters and perhaps on an older version. It is possible that the algorithm will not yield a

significant inter-RNA structure for a given interaction if no structure has a predicted free energy value below 0.00 kcal/mol, especially for the non-reference strains. In this case one can assume that the likelihood of the extrapolated interaction existing in that strain is low, noting that all inter-RNA interactions found in the Dadonaite et al. study did have an associated significant inter-RNA structure. In the Dadonaite et al. study, nucleotide-scale accessibility constraints were used to guide the inter-RNA structure prediction algorithm, which they derived using a powerful *in virio* approach for the strains they analysed, cf. (Dadonaite et al. 2019, Fig. 1). Because these experimentally-derived accessibility constraints are not available for the strains used here for extrapolation, such information cannot not be included here.

The Dadonaite et al. data includes several reference strains for the purpose of this method of analysis: A/WSN/1933[H1N1] (WSN), A/Puerto Rico/8/1934[H1N1] (PR8) and A/Udorn/1972[H3N2] (Udorn, also known as A/Udorn/307/1972). WSN will be used as a basis reference strain for most of the analysis here, owing to the fact that Dadonaite et al. include in the data an 'average of replicates' for the two WSN replicates they analysed, which I regard as more reliable than the data for individual replicates. Also of interest is the use of the Udorn strain as a reference, as it is of a different serotype. An important assumption is that the applicability of interaction extrapolation to new strains is reduced the more genomically distant they are from the reference strain. The Le Sage et al. data is also suitable for use as a reference strain, but since they also and exclusively analyse the WSN strain for wild-type, I will mainly stick to Dadonaite et al. WSN.

INTER-RNA STRUCTURE PREDICTION

Based on advances in the prediction of nucleic acid folding (DNA/RNA folding), algorithms have been developed for the determination of *minimum free energy* (MFE) structures of RNA-RNA interactions. Here, *RNA structure* refers to a configuration of the nucleotide bonds within an RNA molecule (or between two or more interacting RNA molecules). The MFE structure is the RNA structure that occupies the lowest possible free energy state within the *folding landscape* of an RNA molecule, or in this case two interacting RNA molecules. It is well known in the field of RNA structural biology (and in structural biology in general) that MFE structures are not the end-all-be-all of RNA structure biology (Zuker 1989), since RNA molecules can and do adopt non-MFE structures, and the RNA folding energy landscape may depend on the local environment. However, a determination of the MFE structure of an RNA molecule, or in this case the binding structure of two interacting RNA molecules, can provide insight into the properties of that structure, notably including the 'strength' or stability of the structure. The MFE value of an inter-RNA interaction gives an indication of the expected stability of such an interaction: low MFE values correspond to more stable structures, whilst higher MFE values may be associated with weaker interactions that may even represent algorithmic artifacts instead of biologically relevant interactions, absent other evidence. Algorithms that are able to determine the MFE structure of an inter-RNA interaction are therefore useful as a first bioinformatics step for studying the expected properties of the influenza inter-RNA interactions that may be so important for selective packaging.

One such algorithm, *intaRNA* (Busch, Richter, and Backofen 2008), which was developed by researchers at the University of Freiburg, was used in the Dadonaite et al. study to compute the expected MFE structures of inter-RNA interactions. This algorithm, which was originally published in 2008 and further enhanced in 2017 (Mann, Wright, and Backofen 2017), is part of a long line of similar inter-RNA structure prediction algorithms relying on dynamic programming methods, which are in turn based on standard RNA folding algorithms and energy parameters. Several of these algorithms can still be considered 'contemporary', i.e. incorporating some of the new insights in recent years, and each of the corresponding publications seems to claim dominance in some particular field of application. An independent benchmark placed *intaRNA* as one of the top 'competitors' in various domains of application (Umu and Gardner 2017). None of these algorithms seem to have been developed specifically for predicting viral inter-RNA interactions. Therefore, *intaRNA* was chosen as the algorithm to use in the methodology of this thesis in order to remain close to the Dadonaite et al. results whilst still working with a high-quality algorithm.

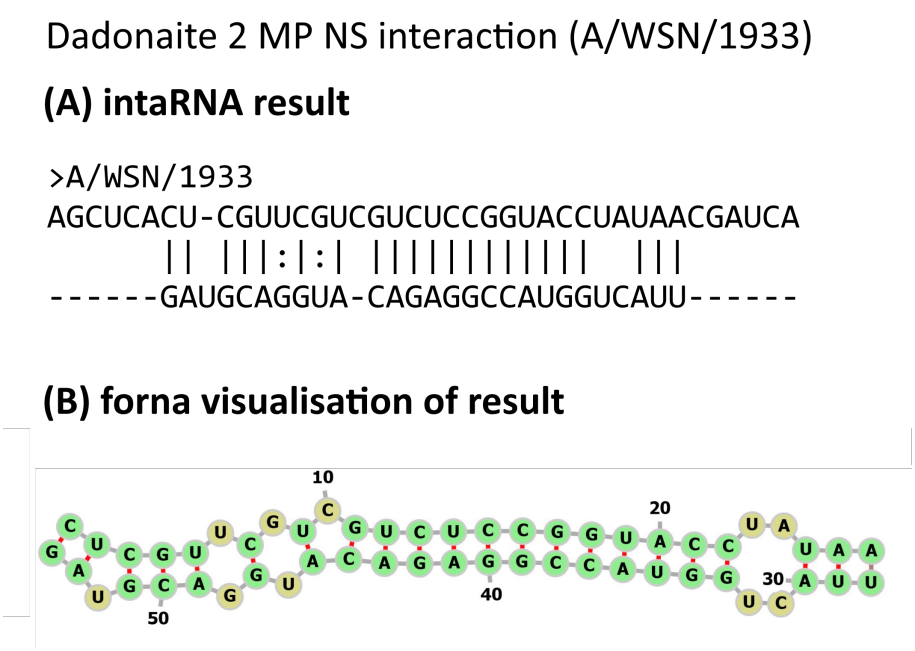


Figure 4: An example of an inter-RNA interaction structure predicted by intaRNA. The interaction is between the MP and NS segments of A/WSN/1933[H1N1] based on data from (Dadonaite et al. 2019). ***(A):*** *The raw output of the intaRNA software.* ***(B):*** *a visualization of the above interaction using the ViennaRNA forna RNA structure visualisation tool. G-U alternative base pairs not included in this visualisation.*

Like the other inter-RNA interaction prediction algorithms, *intaRNA* is based on a dynamic programming method, in this case meaning a method that recursively reduces the original problem to a series of smaller sub-problems, which are divided again until they are solvable. The solved subproblems are then compiled into a solution to the original problem. More concretely, the original problem comprises finding the MFE interaction structure of two RNA sequences, here these will be the RNA regions that were found to be involved in an interaction. Each recursion step divides this problem into smaller sequences, until they become solvable by means of assessment by some

scheme of RNA (sub)structure energy estimation, such as the well-known Nearest Neighbour Database (Turner and Mathews 2010), which is based on real molecular free energy measurements.

Then, a traceback procedure is used to compile the 'global' minimum free energy structure (and calculate the corresponding MFE value) based on these local estimations. The *intaRNA* algorithm presents optimisations on these procedures, including a more efficient recursion procedure and the ability to find and initiate from a *seed* region, a small region of (relatively) uninterrupted inter-RNA base pairing that can be used as a basis for computing the rest of the RNA structure. Also important is the ability to include *accessibility* constraints, meaning that restrictions to base pairing accessibility due to e.g. *intra*-RNA structure and RNA-protein binding can be included as a constraint on the individual nucleotide scale to provide a more accurate inter-RNA structure and MFE energy estimate. Although accessibility data is not available for most of the influenza strains that will be included in this thesis, this constraint was used by Dadonaite et al. for their inter-RNA interaction structure computations, using RNA accessibility constraints measured *in virio* at the nucleotide scale.

One important limitation of most if not all current inter-RNA structure prediction approaches is that they cannot predict the existence of *pseudoknots* and other 'higher-order' RNA (sub-)structures. *Pseudoknots* are a family of various types of RNA structures (Peselis and Serganov 2014) characterised by 'non-linear' base pairing that make them unrepresentable using the standard bracket notation, and difficult to predict using standard dynamic programming methods, although specialised methods for pseudoknot prediction exist (Jabbari, Wark, and Montemagno 2018). They exist natively in structures of various RNA molecules and even have functional implications in viruses (Brierley, Pennell, and Gilbert 2007). Although the concept of pseudoknots does not directly translate to context of inter-RNA interaction structure, the 'non-linear' form of binding seen in pseudoknotting could be extended to inter-RNA interaction, in which case the correct structure of such an interaction would not be elucidated by existing algorithms including *intaRNA*. Resolving this issue is outside of the scope of this thesis, and the small size of the interacting RNA regions in the influenza A inter-RNA interactions should conformationally limit non-linear binding. The existence of more complicated and large-scale inter-RNA interaction structures, possibly involving multiple interacting sites, is an interesting prospect however.

Utilising inter-RNA interaction structure prediction allows for extrapolation and analysis of the influenza RNA-RNA interaction data not just on the sequence level, but also on the (predicted) structure level. This two-level comparative analysis of inter-RNA interactions across different strains will be the central theme of this thesis to tackle some of the questions posed. If someday a technique becomes available that can determine the actual RNA structure of inter-RNA interactions, preferably *in virio*, this analysis can perhaps be extended using those experimentally determined inter-RNA structures.

For the interaction structure prediction, *intaRNA 2.0* version 3.2.1 precompiled (Windows 64 bit) binary, dated November 2020, is used. Unless stated otherwise, the parameter set used is: default settings, with 'exact' (most precise) prediction mode, no accessibility constraint, and no seed constraint.

EXTRAPOLATION STRAINS

Thousands of sequences of influenza A genomes are available online in scattered databases, including in the *Influenza Research Database* (located at fludb.org) and the NCBI *Influenza Virus Database* (NCBI IVD) (Bao et al. 2008). For the purpose of the analysis performed here, it is important to have complete genomes, containing all eight genome segments and all material in each of the genome segments (no significant sequencing gaps). However, the pipeline built here is able to deal with incomplete genomes if necessary. As a first step for analysis and exploration of results, H1N1 vaccine strains in the NCBI IVD were picked with (Dadonaite) WSN as the reference strain, because these vaccine strains could be considered to constitute a form of consensus concerning the human-infecting H1N1 strains occurring in the time period in which the vaccine derived from that strain was made. The following (all H1N1) strains fit this description within the NCBI IVD:

*Box 1: **H1N1** interpolation strain set; A/WSN/1933[H1N1] (WSN) as the reference strain. WSN has a small, presumably sequencing-related gap in the NA segment.*

- A/Bayern/7/1995 (absent: PA, PB1, PB2; incomplete: HA, NP, MP)
- A/Beijing/262/1995
- A/Brazil/11/1978
- A/Brisbane/59/2007
- A/California/07/2009
- A/Chile/1/1983
- A/Michigan/45/2015
- A/New Caledonia/20/1999
- A/Singapore/6/1986
- A/Solomon Islands/3/2006 (absent: NS, PB1, PB2; incomplete: PA)
- A/South Dakota/06/2007
- A/USSR/90/1977

The WSN, PR8 and Udorn sequences were taken from the Dadonaite et al. supplemental data (first replicates). The low quality of the ‘Bayern’ and ‘Solomon Islands’ sequences pose challenges for the analysis of these strains, but all other strains are relatively complete, bar potential sequencing problems in the near-end regions on both sides of vRNA segments in several strains. The chronological and geographical variety of these strains offers a solid basis of analysis for H1N1 strains of influenza A.

Owing to the role of H3N2 strains in causing human outbreaks and pandemics, a selection of H3N2 vaccine strains with a large chronological and geographical spread was compiled:

*Box 2: **H3N2** interpolation strain set; A/Udorn/72[H3N2] as the reference strain. WSN is also possible as a reference strain but is less relevant due to greater sequence dissimilarity.*

- A/Moscow/10/1999
- A/Bangkok/1/1979
- A/Beijing/353/1989
- A/Leningrad/360/1986
- A/Wellington/01/2004
- A/Beijing/32/1992
- A/Wisconsin/67/2005
- A/Brisbane/10/2007
- A/Shandong/9/1993
- A/Philippines/2/1982
- A/Victoria/210/2009 (large gaps in PB1, PB2 segments, smaller gaps in HA, PA)
- A/Texas/50/2012
- A/Victoria/361/2011
- A/Ohio/02/2012
- A/Colorado/06/2017
- A/Brisbane/1/2012
- A/Brisbane/6/2012
- A/South Australia/55/2014
- A/Kansas/14/2017

For this strain set, A/Udorn/72[H3N2] (Udorn) is most fitting as the reference strain due to its serotypic and genomic similarity. However, WSN (or even PR8) could also be used as the reference strain, especially when working with shared WSN-Udorn interactions in the Dadonaite et al. dataset.

To increase the variety of serotypes and genetic material, some analysis was also performed with a set of mixed serotype (not necessarily vaccine-related) strains, with WSN as the reference strain. The selection contains several non-human strains, as well as the 2009 pandemic A/Belgium/145-MA/2009[H1N1] strain and the 1968 pandemic A/Hong Kong/01/1968[H3N2] strain. A/Anhui/1-BALF_RG1/2013[H7N9] was responsible for an novel-serotype outbreak with an unusually high human case fatality rate in 2013 (Watanabe et al. 2013).

*Box 3: **Mixed serotype** interpolation strain set; A/WSN/1933[H1N1] as the reference strain.*

- A/Puerto Rico/8/1934-Korea/426/1968[H2N2]
- A/Anhui/1-BALF_RG1/2013[H7N9]
- A/Belgium/145-MA/2009[H1N1]
- A/Brisbane/59/2007[H1N1]
- A/Hong Kong/01/1968[H3N2]
- A/duck/Hokkaido/Vac-3/2007[H5N1]
- A/duck/Zhejiang/6DK19-MA/2013[H5N2]
- A/mallard/Alberta/70/2017[H7N3]

Lastly, a selection of H5Nx strains was made to study those serotypes in greater depth, due to the increasing potential of these serotypes in causing outbreaks of influenza in humans. The selection forms a chronologically and geographically varied mixture of strains from different animal sources (including humans, birds, and swine) and including several different neuraminidase (NA/Nx) serovars:

*Box 4: **H5Nx** serotypes interpolation strain set; A/WSN/1933[H1N1] as the reference strain.*

- A/Anhui/1/2005[H5N1]
- A/Vietnam/UT36282/2010[H5N1]
- A/Egypt/MOH-NRC-8434/2014[H5N1]
- A/Changsha/1/2014[H5N6]
- A/Yunnan/0127/2015[H5N6]
- A/duck/Mongolia/54+47/01[H5N1]
- A/duck/Hokkaido/Vac-3/2007[H5N1]
- A/duck/Moscow/4182-C/2017[H5N3]
- A/duck/Zhejiang/6DK19-MA/2013[H5N2]
- A/Anas Platyrhynchos/Belgium/10811-6/2019[H5N6]
- A/swine/Banten/UT2071/2005[H5N1]
- A/swine/Zhejiang/SW57/2015[H5N1]
- A/Crow/Aghakhan/2017[H5N8]
- A/swan/Krasnodar/44/2017[H5N8]

INTER-RNA INTERACTION DISTRIBUTIONS

In this experiment, profiles are created of the number of unique inter-RNA interactions sites overlapping each nucleotide position along the genome segments. This is done based only on the raw (Dadonaite et al. 2019) and (Le Sage et al. 2020) supplemental data, structure prediction was not necessary here. Finding the distribution of hotspots along the eight vRNA segments for each strain is straightforward based on the data. Each nucleotide in each segment is simply annotated with the number of unique interactions overlapping that nucleotide in the dataset. The three strains available in the Dadonaite et al. data are: A/WSN/1933[H1N1] (WSN), A/Puerto Rico/8/1934[H1N1] (PR8) and A/Udorn/72[H3N2] (Udorn). Note that the Udorn strain is of a different serotype than the first two, although this does not necessarily mean that it is also dissimilar to WSN and PR8 in non-antigenic segments. For WSN, the data from the ‘average’ of the two reproduced replicates is used, whilst for PR8 and Udorn, the data on the first replicate is used. For Le Sage et al., the data on the

first replicate of their WSN strain is used, as well as the *HSmut* strain created by the researchers as a variant of WSN in which the NP hotspot was disrupted using synonymous mutations. The distributions were normalised by setting the maximal value across all segments to one, i.e. setting the highest peak to one and normalising the distributions for all segments to that peak value. This allows for cross-comparison of hotspots across segments. A second normalisation procedure including reads-per-million (RPM) information was omitted due to lack of observed effect on the resulting distributions.

SEQUENCE- AND STRUCTURE CONSENSUS

Based on the data at hand, several different ideas are possible for further analysis of the inter-RNA interactions without performing additional web-lab experiments. Firstly, sequence- and (predicted) structure consensus can be computed using the strains gathered (reference strain and extrapolation strains) at the nucleotide level to get a view of conserved sequence- and structural elements inside interactions. For *sequence consensus*, multiple sequence alignment followed by consensus scoring is sufficient. For structure consensus, this is a little more complicated because there is no set way to align (inter-)RNA structures. One method is to align the underlying sequences, and then assign inter-RNA bond information to each nucleotide, i.e. whether that nucleotide engages in inter-RNA base pairing in that specific interaction. In this way, sequence alignment is used to align the inter-RNA structures, and consensus scoring can be performed on the structures. This must be performed separately for the two interacting RNA elements, resulting in two separate *per-nucleotide structure consensus* profiles aligned to the interacting sequences.

A method to obtain a single structure consensus profile for an interaction is also possible. It involves placing the predicted inter-RNA structure and corresponding two interacting sequences back into the original two sequences used for the structure prediction query. This is necessary because most predicted interaction structures involve sub-sequences of the original query sequences, which can be used for structure-oriented re-alignment of those query sequences and the inter-RNA binding pattern between them. These structure-oriented aligned binding patterns can then be used for consensus scoring, yielding a single *structure-oriented consensus* profile. A graphical overview is shown in Figure 5.

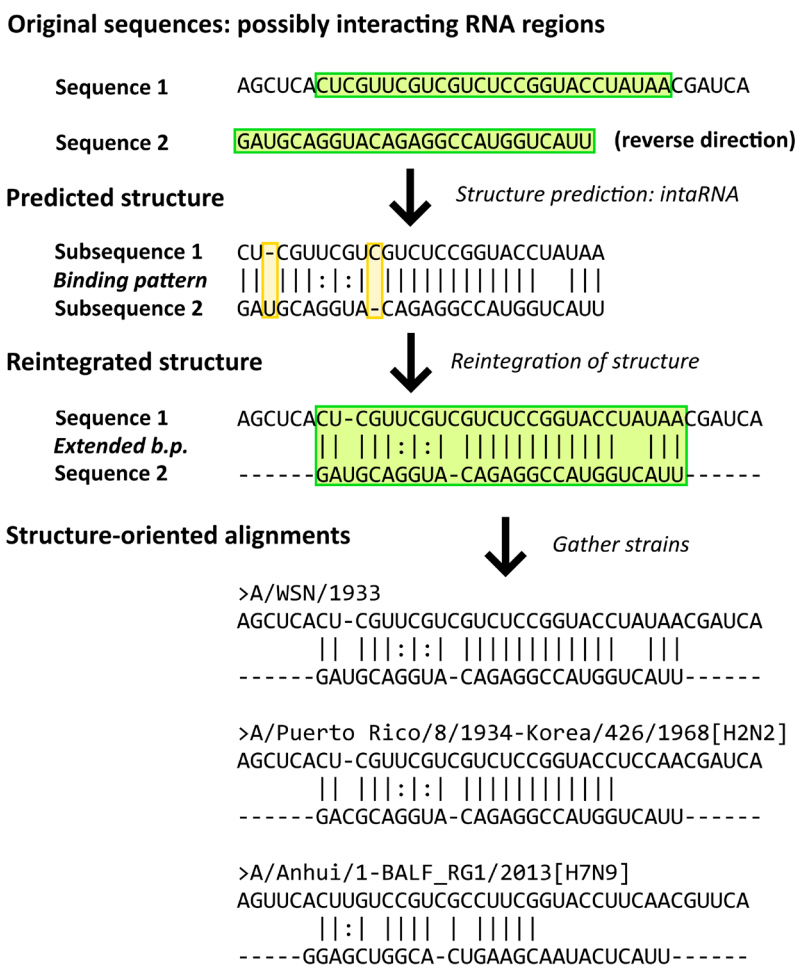


*Figure 5: Overview of structure-oriented realignment of inter-RNA interaction sequences. The two 'original' sequences or query sequences are derived e.g. by the mapping algorithm used to deduce which RNA regions are involved in a particular inter-RNA interaction. The predicted interaction structure between sub-sequences of those query sequences is reintegrated into the query sequences, after which comparative analysis between strains of binding patterns is possible for that interaction. The **green** bars show the predicted interacting sub-sequences, and interaction binding pattern. The **yellow** bars show gaps introduced in the sub-sequence structural alignment due to loops in the predicted interaction structure.*

In the absence of a precise and complete phylogenetic tree describing the evolution of each strain, the resulting alignments for determining sequence- and structure consensus can be interpreted simply as a histogram at each position (nucleotide or structure position). For example, for sequence consensus, each nucleotide is either one of the four RNA nucleotide bases ('A', 'U', 'C', 'G'), or an alignment gap (symbol '-'), forming a discrete distribution over the various strains (reference strain and extrapolation strains). There are various measures of 'consensus' applicable to such a histogram, including variance from the histogram mean, histogram entropy and mean pairwise Hamming distance.

Mean pairwise Hamming distance (*mpd*, technically its inverse), as used by (Gog et al. 2007) in a similar experiment, is simply the proportion of equal pairs (of sequences, nucleotides, etc.) among the total number of possible pairs of values:

Equation 2: Inverse mean pairwise Hamming distance defined over a histogram with N bins and M total values (all bins).

$$MPD = \frac{1}{M^2}\sum_{i=1}^{N}(p(x_i)M)^2$$

Where M is the (unnormalised) sum of the histogram across all bins, e.g. the number of sequences, N indicates the number of possible histogram values or *bins*, e.g. the number of possible nucleotides plus alignment gap symbol for sequence consensus, and $p(x_i)$ is the histogram probability (normalised to M) of symbol x_i, e.g. $p(A)$ is the number of occurrences of nucleotide 'A'. Without additional normalisation, the upper limit (full consensus) of this score is always one, but the lower limit (minimal consensus) is close to $\frac{M}{N^2}$. I will mainly use the histogram entropy measure as a consensus score, which measures the entropy of the histogram configuration as follows:

Equation 3: Histogram entropy for N bins.

$$S = -\sum_{i=1}^{N} p(x_i)\log p(x_i)$$

This can then be mapped to an entropy consensus score between zero and one, where zero indicates minimal consensus and one indicates maximum consensus:

Equation 4: Histogram entropy consensus score from histogram entropy S, with N bins.

$$C = 1 - \frac{S}{\log N}$$

The sequence consensus, per-nucleotide structure consensus, and reintegrated structure consensus profiles for inter-RNA interactions will be computed as a measure of nucleotide-scale conservation, which can provide evidence for the existence of conserved sequence- or structural elements in such interactions.

MUTATION

It is difficult to track mutagenesis across strains for widely divergent influenza strains, as a full phylogenetic tree for influenza A is not well-defined, and this is complicated even further by the occurrence of reassortment processes between strains. Nevertheless, it may be useful in some cases to use the inter-RNA interaction reference strain as a basis strain for mutation analysis, with the extrapolation strains as variants. In this case, sequence mutation rates, consisting of the substitution rate, deletion rate and insertion rate, can be computed per nucleotide or over a section of the genome based on comparison of this reference strain to the extrapolation strains. In this context, the overall 'mutation rate' refers to the sum of these three individual mutation rates. 'Conservation

rate' then refers to the inverse of this sum, i.e. one minus the mutation rate. The conservation rate in this context is equal to the notion of 'sequence identity' often seen in genomics.

Strain	Sequence	Mutation rate
Ref.	ACUUGC-C	(=0.000)
Var. 1	AGUUGC-C	(1+0+0)/8=0.125
Var. 2	A-AUCG-C	(3+1+0)/8=0.500
Var. 3	A-UUGCCA	(1+1+1)/8=0.375

*Figure 6: Mutations in a sequence alignment w.r.t. a reference sequence (top row). Substitutions are in **yellow**, deletions in **red**, insertions in **cyan**. The mutation rate for a strain is calculated as an average of SNPs at each position in the alignment. Mutation rates per position can be calculated by summing SNPs along a column and dividing by the number of variant strains.*

COVARIATION

In the context of genomics, *covariation* occurs when pairs (or greater multiples) of sequence elements (most commonly individual nucleotides) share an evolutionary constraint, resulting in co-dependent/cross-correlated mutational patterns (Dutheil 2012). Such an evolutionary constraint can be imposed by RNA structure, in which case pairs of nucleotides involved in a (intra- or inter-RNA) nucleotide bond may not mutate independently due to conservation of that bond being important for maintaining the RNA structure. A *covariation event* occurs when mutation does occur in those particular nucleotides in such a way that the bond (and the RNA structure) is conserved, e.g. if both nucleotides mutate to a pair of complementing nucleotides. Evidence of the occurrence of covariation events within some sequence element may be seen as evidence of the presence of such an evolutionary constraint, in this case, the presence of an important, conserved inter-RNA interaction, or elements of that interaction.

A well-informed study of covariation must take into account the phylogeny of the genomes being studied to elucidate which specific covariation events could have taken place (Dutheil 2012). Covariation is often detected and quantified on the basis of measures of *mutual information* between genome elements (Dutheil 2012). In the context of this thesis, proper statistical analysis of covariation processes is not available due to the complexities and unknowns of influenza phylogeny, and uncertain quality of inter-RNA interaction structures. Instead, I will look at the possibility of individual covariation events within inter-RNA interactions, supported by phylogenetic information when possible.

Due to the shifting nature of (predicted) inter-RNA interaction structures, a definition for covariation in this context is not immediately obvious. The key idea is that a covariation event is a double mutation in nucleotides on different segments that acts to conserve a base pair in an inter-RNA interaction. However, the (predicted) inter-RNA structure depends on the configuration of all nucleotides in the interaction regions, so base pair conservation is not solely determined by mutation events in the two nucleotides that engage in the bond. Nevertheless, the most sensible definition of a covariation event

in the context of inter-RNA interactions is: a double mutation in two nucleotides that base-pair in an inter-RNA interaction in a reference strain, resulting in conservation of that specific base pair in the inter-RNA interaction in the variant strain. This is illustrated in Figure 7.

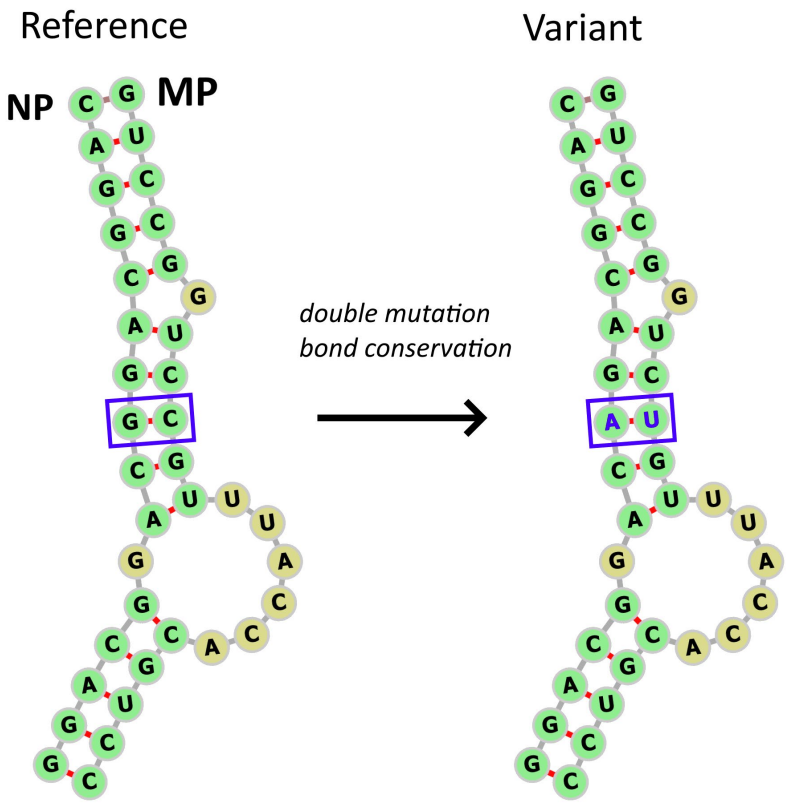


*Figure 7: Graphical illustration of covariation principle in an example inter-RNA interaction between the NP and MP segments (Dadonaite et al. WSN 28 NP 670-712 MP 355-395). Depicted is the predicted interaction structure in the reference strain, and the same interaction in a fictional variant strain with a double mutation (**blue** box) in two base pairing nucleotides, which results in conservation of that base pair instead of loss of complementarity. forna is used to visualise the RNA structures.*

STRUCTURAL IMPOSITION

To provide another perspective into similarities and differences in predicted inter-RNA interaction structures across strains, the *structural imposition* experiment is introduced. In this method, a predicted inter-RNA interaction structure for one strain is imposed unto homologous sequences in another. This imposed structure is then scanned for incompatible base pairings, i.e. when complementarity is lost due to variations in the imposed strain. Additionally, if two non-base-paired complementary nucleotides are directly opposite one another in the imposed structure, a new base pairing is introduced between them. This *rescanned* imposed structure is then evaluated using the ViennaRNA *RNAeval* tool, which estimates the free energy of *intra*-RNA structures, because *intaRNA* is not available for direct free energy estimation on existing structures. In order to transform the *inter*-RNA structures (consisting of two interacting (sub-)sequences) into *intra*-RNA structures (consisting of just one RNA sequence with internal base pairing), five unpaired ‘G’ nucleotides are inserted between the two interacting sequences, after which the inter-RNA interaction can be

converted to an intra-RNA interaction in the fused sequence. The quintuple ‘G’ insert is introduced in order to avoid impossible stereochemical configurations (i.e. base pairing between two nucleotides that are directly connected on the backbone), which are incompatible with *RNAeval*. This quintuple ‘G’ insert has a moderate (single digit) effect on the predicted free energy, but this is compensated by also evaluating the original (non-imposed) inter-RNA structure in this way. *RNAeval* can yield results different from *intaRNA* predicted MFE values mainly due to algorithmic- and parametric differences, but in most cases, they were found to at least correlate. A graphical overview of the method is given in Figure 8.

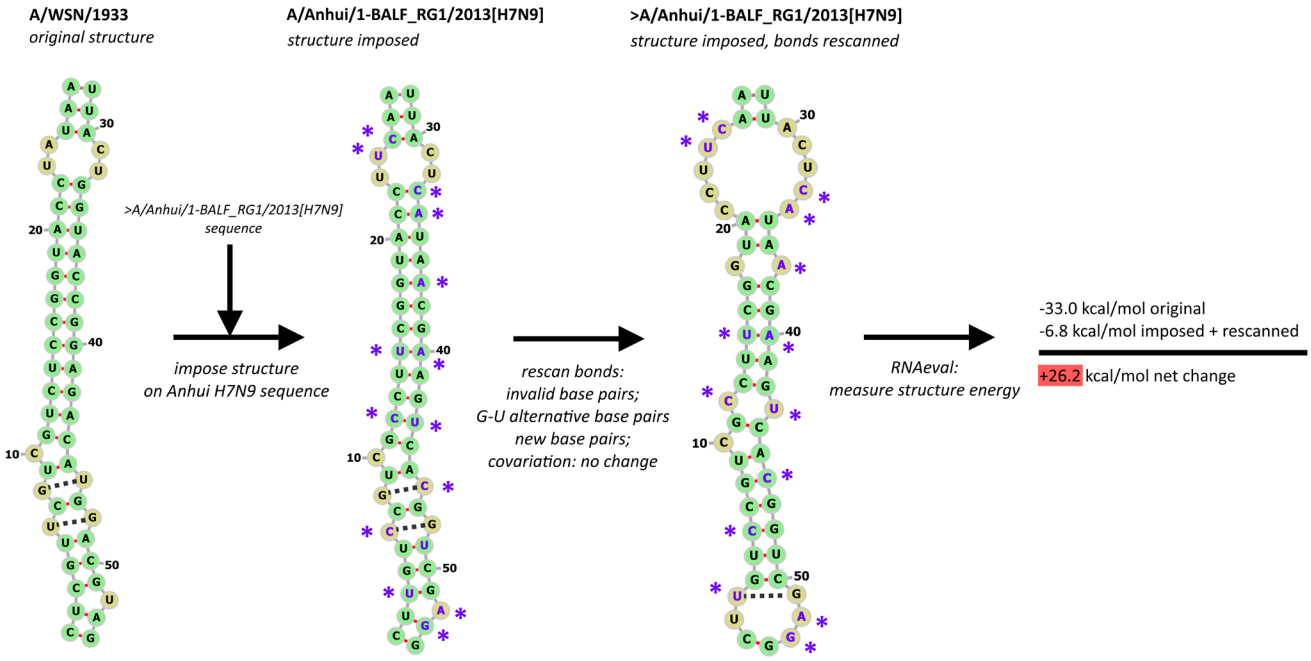


Figure 8: An overview of the structural imposition method, with the Dadonaite et al. 2 MP 382-420 NS 605-631 interaction as an example. The predicted structure of A/WSN/1933 for this interaction is imposed on the sequence of A/Anhui/1-BALF_RG1/2013[H7N9], with mutations marked in purple. Then, each inter-RNA base pair is re-examined for validity: non-complementary bonds are removed, complementary nucleotides directly opposite one another in the imposed structure form a new bond, opposing G-U base pairs form alternative base pairs, and all other bonds are left unchanged (including in cases of covariation). The ViennaRNA RNAeval tool is then used to evaluate the free energy of the original structure and the rescanned imposed structure for comparison. A loop of 5 ‘G’ nucleotides is inserted in the middle of the bracket-notation form of these interactions (here, between the ‘A’ and ‘U’ nucleotides at the top of the structures) before RNAeval evaluation to avoid impossible stereochemical configurations which cause the algorithm to glitch. forna is used to visualise the RNA structures.

The purpose of the structural imposition experiment is to measure the *compatibility* of a predicted inter-RNA structure for one set of interacting nucleotides with another set of interacting nucleotides. If there is a large positive net change in free energy for the rescanned imposed structure, then the imposed strain likely has a different optimal structure for that interaction, if it has a good optimal structure at all. This may indicate a lack of interaction conservation, and it may have consequences for e.g. reassortment compatibility between strains.

CROSS-STRAIN INTERACTION

The cross-strain interaction experiment focuses on elucidating the compatibility of two strains for specific inter-RNA interactions, which is potentially relevant in understanding reassortment compatibility between strains. In the cross-strain interaction experiment, the interaction site sequence for one segment for one particular strain is combined in an *intaRNA* query with the sequence for the opposing segment from a different strain. In this way, genetic material from two different strains is combined to check whether the particular interaction between those two strains is feasible. There are two cross-strain combinations which can be made per interaction and per strain pair, depending on which strain 'donates' which segment. The resulting predicted structures and corresponding MFE values can then be compared to the in-strain predictions for both involved strains. Unlike the structural imposition experiment, *intaRNA* is used directly here for structure- and free energy computation.

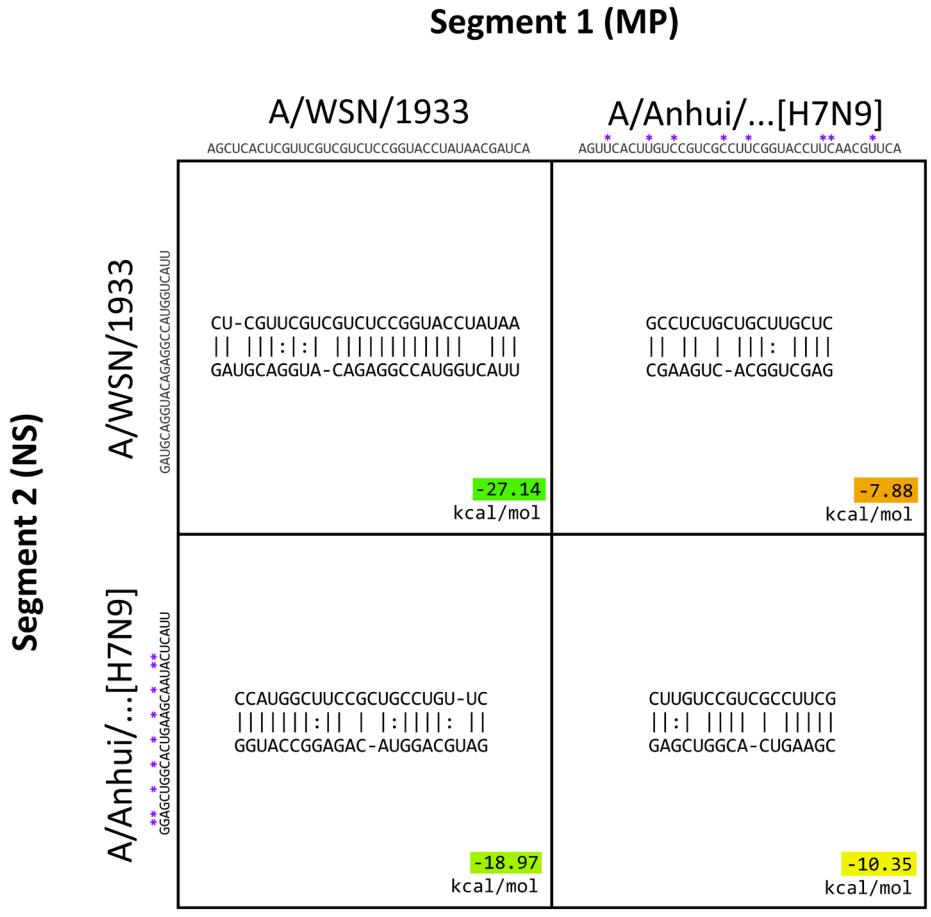


Figure 9: An overview of the cross-strain interaction method. Shown here are the cross-strain computations for strains A/WSN/1933[H1N1] and A/Anhui/1-BALF_RG1/2013[H7N9] for the Dadonaite et al. 2 MP 382-420 NS 605-631 interaction, between the MP and NS segments. Using intaRNA, the predicted interaction structure was computed for each combination of interaction site sequences from each strain. The (non-reintegrated) predicted structure and MFE value (bottom rights) for each combination are shown in the boxes. The purple stars mark mutations in the H7N9 strain interaction site sequences w.r.t the H1N1 strain. The cross-strain interactions are in the ***bottom-left*** *and* ***top-right*** *boxes.*

VISUALISATION

There are multiple ways to visualise inter-RNA interactions, but no specific tools are available for this purpose yet. In this thesis, the main ways in which inter-RNA interaction structures are visualised are: *intaRNA-like* and *forna*, cf. Figure 4. The *intaRNA*-like visualisation method aligns the two interacting sequences, with potential insertion of gap symbols (i.e. '-') for nucleotides opposite loops, and shows the *binding pattern* of nucleotides between these sequences. Here, the colon symbol is used to visualise G-U alternative base pairs. Sometimes, the interaction structural alignment is first 'reintegrated' into the full query sequences (cf. Figure 5), because usually only parts of the full interaction site sequences form the predicted optimal inter-RNA interaction structure.

The *forna* visualisation technique relies on fusing the two interacting sequences (*without* a middle insert, cf. Structural imposition section), and linearising the binding pattern between the sequences to bracket notation as if it were a single RNA molecule. This fused RNA structure is then visualised using the ViennaRNA *forna* tool, available at http://rna.tbi.univie.ac.at/forna/, which can only handle single-molecule RNA structures. A limitation of this approach is that it cannot handle the full interaction site sequences when only parts of those sequences constitute the predicted interaction structure (i.e. the aforementioned reintegration method is not possible) because an improper loop would appear around the sequence fusion point in the visualisation.

RESULTS

Known inter-RNA interactions in influenza A

As discussed in the Introduction, so-called *packaging signals* or *packaging sites* exist in the influenza genome: regions of the viral RNA that are involved in genome packaging and/or selective packaging. Both general- and selective packaging seem to be driven by RNA elements near the 3' and 5' ends of each vRNA segment, mostly in the untranslated regions (UTRs), but also sometimes overlapping the coding regions, cf. (Hutchinson et al. 2010, Fig. 4). However, the recent (Dadonaite et al. 2019) and (Le Sage et al. 2020) studies, amongst others, seem to suggest that the RNA elements responsible for inter-RNA interactions, which are thought to be a key factor in selective packaging, are located all over the vRNA segments, with many occurrences well inside coding regions. Evidence from the earlier (Gavazzi, Yver, et al. 2013) and (Fournier et al. 2012) studies seems to suggest that such inter-RNA interactions well inside coding regions are indeed important in maintaining selective packaging. These hundreds of unique inter-RNA interactions seem to form a network involving all eight vRNA segments. These findings seem to contradict the idea that it is mostly the untranslated regions around segment ends that are involved in genome packaging in influenza A.

The data in the Dadonaite et al. and Le Sage et al. studies seems to suggest that the distribution of inter-RNA interactions is far from uniform across the vRNA segments. Certain regions, deemed *hotspots*, seem to engage in far more interactions than other regions of the vRNA. In particular, one hotspot in the NP (nucleoprotein-encoding) vRNA segment was analysed in depth by the Le Sage et al. team, who concluded that this hotspot is central to the RNA-RNA interaction network, but that a built-in degree of redundancy exists within this network to compensate for potential mutations in the interacting RNA elements including this hotspot. Such redundancy could imply a degree of robustness of the inter-RNA interactions to mutation. By extension, this could indicate that the selective packaging process in influenza A could be somewhat robust to random genomic mutations through this redundant network of RNA-RNA interactions. Also interesting is the observation that some vRNA regions seem to not engage in any inter-RNA interactions at all, I will dub these *silent regions*. In this section, I will take a look at the distribution of inter-RNA interactions in the various strains analysed in the Dadonaite and Le Sage studies, with the aim of discovering potential patterns in inter-segment interactivity and in hotspots and silent regions and ascertaining whether these patterns are conserved across strains and across the two studies.

GENERAL OVERVIEW OF INTER-RNA INTERACTIONS IN INFLUENZA A

Before analysing the distributions of the inter-RNA interactions over the segments, it is prudent to take a look at a more global overview of the number of interactions per segment, and at the number of interactions between each pair of segments.

Table 2 shows the absolute number of interactions involving each segment in the Dadonaite et al. data for the WSN strain, and the number of interactions averaged over the length in nucleotides of

each segment. The same is showed for the Udorn strain of Dadonaite et al. in Table 3. Since PR8 is quite similar in this regard to WSN, this strain has been omitted. First off, fewer interactions were found for the Udorn strain overall in this experiment. For both replicates of WSN, around 850 interactions were found in Dadonaite et al. (611 combined interactions in the 'average' of replicates), compared to only 112 and 312 respectively for the two replicates of Udorn (of which the first was selected for further analysis here). Why this might be the case is unclear, but it might be useful to check in a future study whether this is a real reflection of differing numbers of total interactions in the two strains, or a consequence of the method and/or statistical evaluation. For PR8, 1275 and 600 interactions were found for the two respective replicates, again widely differing numbers for replicates which are supposed to be almost equal genetically. For the Le Sage et al. study, 675 and 1240 interactions were found for the two replicates of WSN respectively.

From looking at Table 2 for Dadonaite et al. WSN, it is apparent that the longer segments (PB2, PB1, PA) engage in more interactions overall, but that the shorter segments (MP, NS) have more interactions relative to segment length. An exception to this is the NA segment, which seems to engage in relatively few interactions in both regards. For Udorn in Table 3, the reverse is true: NA engages in the most interactions of any segment, both absolutely and relatively. Here, the three longest segments engage in the fewest interactions relative to segment length. It should be noted that the low number of overall interactions may reduce the statistical significance of these findings in the case of Udorn. To this end, replicate 2 of Udorn in Dadonaite et al. has also been included. Here, NA is still relatively strong in terms of the number of relative interactions, but the long PB2 and mid-length HA segments are now also relatively strongly represented. Interestingly, the short segments MP and NS here engage in few interactions both absolutely and relatively speaking. Overall, the picture of which segments are central to the interaction network in terms of the number of interactions they engage in is inconclusive. The only real conclusion is that all eight segments seem to engage in inter-RNA interactions in the strains analysed.

Table 2: Interaction coverage table for A/WSN/1933[H1N1] (WSN), average of replicates based on data from (Dadonaite et al. 2019). ***Interpretation:*** *although the polymerase segments have the most interactions overall, the two smallest segments, MP and NS, surprisingly have the most interactions per nucleotide. The NP segment, which is middling in length, has a relatively high number of interactions per nt potentially due to the NP hotspot.*

Segment	Length (nt)	Number of interactions	Avg. interactions per nt
PB2	2341	194	0.0829
PB1	2341	204	0.0871
PA	2233	193	0.0864
HA	1775	136	0.0766
NP	1565	166	0.1061
NA	1409	91	0.0646
MP	1027	128	0.1246
NS	889	110	0.1237

Table 3: Interaction coverage table for A/Udorn/1972[H3N2] (Udorn), both replicates shown based on data from (Dadonaite et al. 2019). ***Interpretation:*** *the overall number of interactions found in both replicates was unfortunately lower, but the relative number of interactions with the polymerase segments has decreased, whereas the NA (neuraminidase) segment now has relatively many interactions, possibly due to the presence of several new hotspots in the second half of this segment.*

Segment	Length (nt)	Number of interactions	Avg. interactions per nt	*Replicate 2* Length (nt)	No. inter.	Avg. inter.
PB2	2341	29	0.0124	2341	123	0.0525
PB1	2341	32	0.0137	2341	92	0.0393
PA	2233	27	0.0121	2233	88	0.0394
HA	1774	30	0.0169	1765	98	0.0555
NP	1565	33	0.0211	1565	74	0.0473
NA	1466	38	0.0259	1466	72	0.0491
MP	1027	18	0.0175	1027	46	0.0448
NS	890	17	0.0191	890	31	0.0348

More specifically, it is compelling to look at the number of unique interactions between each *pair* of segments in the various strains, to gain an understanding of how segments interact with one another. Multiple discrete interactions were found between most pairs of segments in both datasets for the various strains analysed, with each discrete interaction consisting of different interacting RNA regions, although (part of) the same RNA region on one segment can interact separately with multiple other RNA regions on other segments. Figure 10 shows a heatmap of the number of interactions between each pair of segments for the WSN (average of replicates) strain in Dadonaite et al., Figure 11 shows the same for the two replicates of Udorn in Dadonaite et al. In both cases, the three long polymerase segments PB2, PB1 and PA form a block dense with interactions. In WSN, the PB2-PB1 and PB2-PA interactions are particularly numerous, the same goes for Udorn. The polymerase segments also seem involved in a large number of interactions with the HA, NP and perhaps MP segments in WSN. In Udorn replicate 2, HA seems to engage in many interactions with the polymerase segments as well. HA also interacts more commonly in Udorn with the NP and NA segments. In WSN, there exists a relatively large number of interactions between the MP and NP segments. Also notable are the absence of interactions between MP-HA and NA-NS in WSN, and near-absence of interactions of PB2-NS in Udorn replicates 1 and 2.

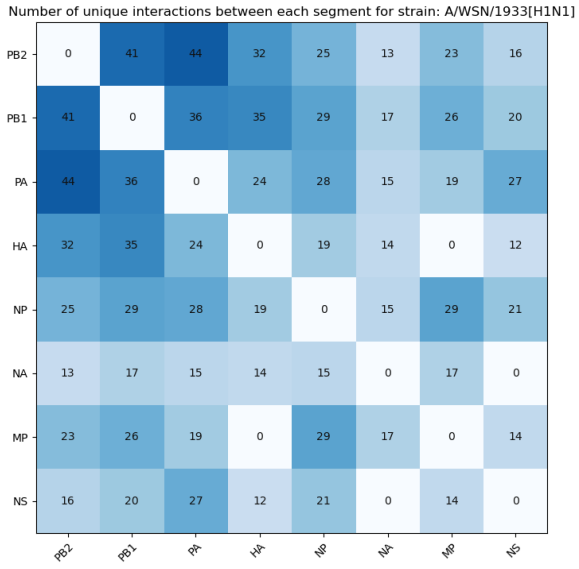


Figure 10: Unique interactions found between segments in A/WSN/1933, average of two replicates, from the (Dadonaite et al. 2019) data. ***Note****: This figure is symmetric along the diagonal axis.*

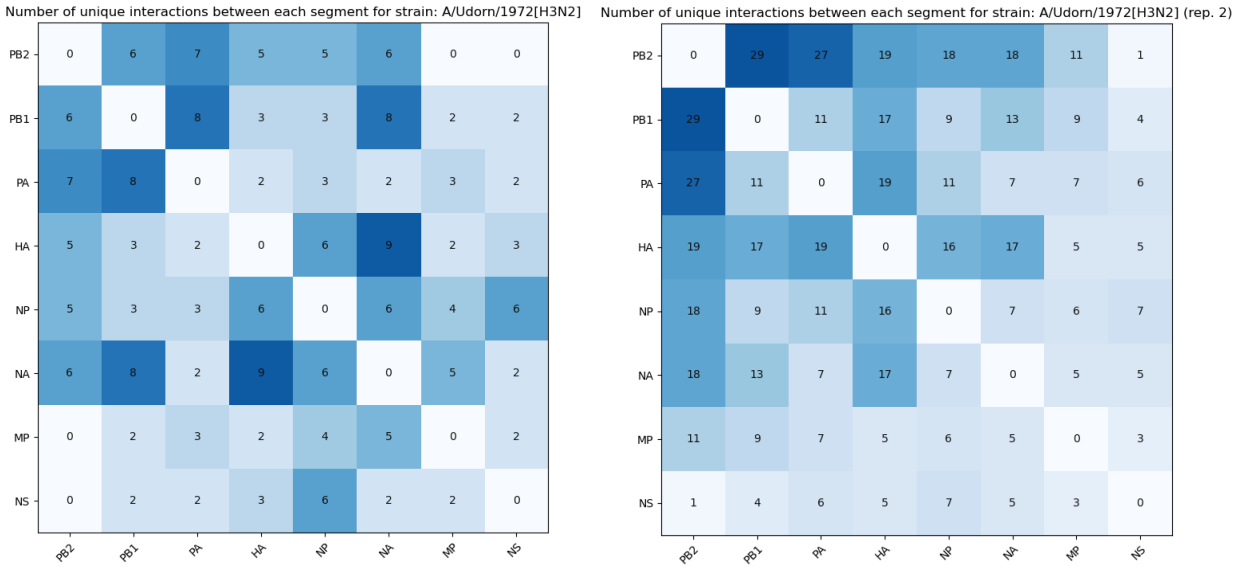


Figure 11: Unique interactions found between segments in A/Udorn/1972, both replicates displayed ***left/right****, from the (Dadonaite et al. 2019) data.* ***Note****: These figures are symmetric along the diagonal axes.*

A more precise overview of the ~50 highest RPM interactions for WSN and Udorn is available in Figure 12. It is apparent that inter-RNA interaction networks are complex and non-uniform across segments, and that they differ between the WSN and Udorn strains. Based on the results so far, it is also evident that most segments engage in at least one interaction with each other segment, indicating that a model in which one segment is responsible for mediating interactions between other segments is probably unlikely. This may also mean that the '7+1'-vRNP architecture is not static and could simply be a consequence of spatial constraints instead of a consequence of a central mediating vRNP.

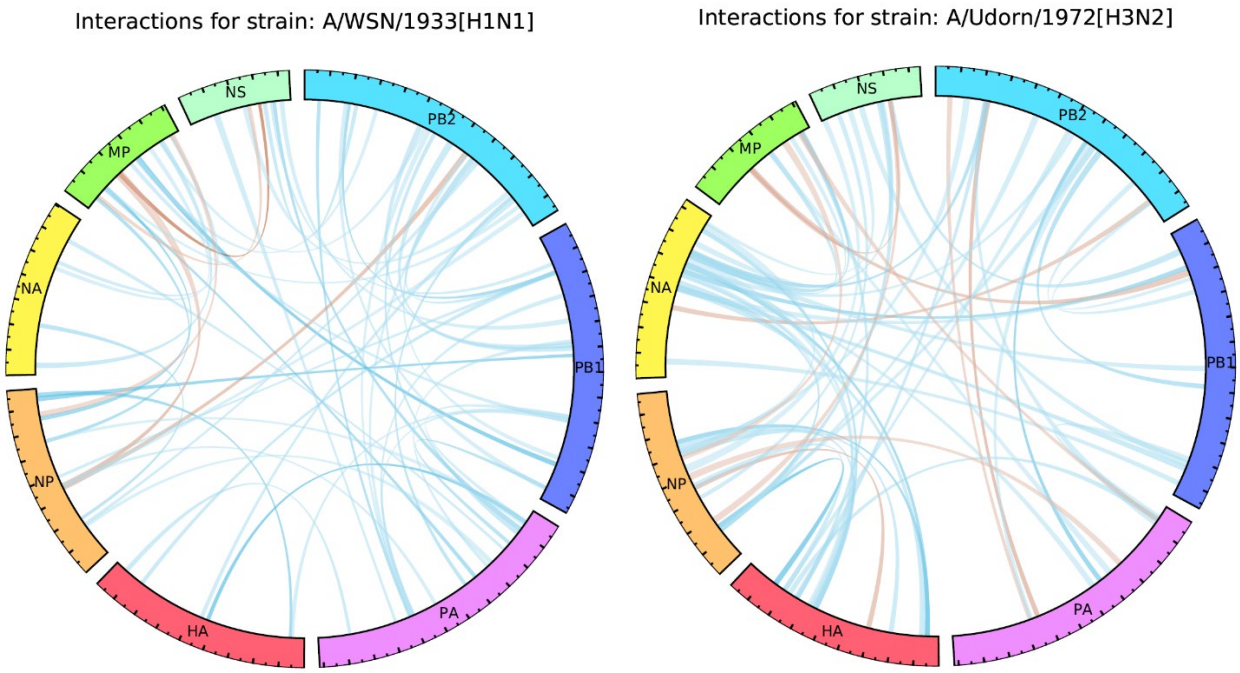


Figure 12: ***Left:*** *Circos (pyCircos) plot of 54 (out of 611) interactions with highest RPM in the Dadonaite et al. WSN average of replicates dataset.* ***Right:*** *the same plot for 56 (out of 112) interactions in Udorn, first replicate. This is a recreation of the Circos plots in Dadonaite et al., e.g. (Dadonaite et al. 2019, Fig. 2). Triple conserved (WSN, PR8, Udorn) interactions in the Dadonaite et al. dataset are marked in* ***red****. All segments are ordered clockwise in the 5' to 3' direction.* ***Interpretation:*** *although some interactions, including those marked in red, are conserved across the two strains, the overall interaction networks are markedly different. Both interaction networks are also highly non-uniform across the segment lengths. Even in this smaller selection, it seems that most segments have at least one interaction with all other segments, no 'central' segment in the network is apparent.*

Overall, no clear and/or conserved pattern is visible in terms of which segments interact with which other segments, but some segment pairs do engage in more unique interactions than others. However, this does not directly mean that these segments interact relatively less with one another, since the strength of interactions and other factors such as co-localisation and site accessibility must also be taken into account. In the Dadonaite et al. study, the number of reads per million (RPM) may give insight into how common each unique interaction is, i.e. how likely that interaction is to be found at a given time in the given strain. For interaction sites which can have multiple interaction partners and for chains of interactions in which multiple configurations of interactions are possible, this could provide a view into the ensembles of interactions that are likely to exist at the same time. The wide distribution of RPM values for the interactions found and low correlation to predicted interaction MFE values (computed by Dadonaite et al.) as shown in Figure 13 (for WSN) probably indicates that not every possible interaction is present all the time. There may be multiple possible ensembles of co-occurring interactions due to (partly) overlapping interaction sites and/or differing segment co-localisation patterns.

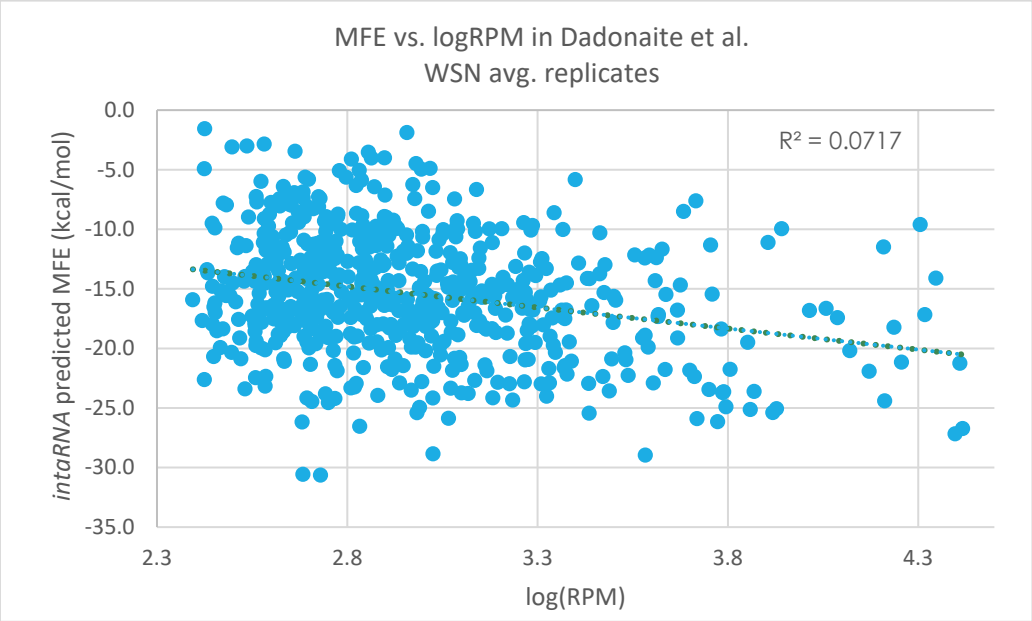


Figure 13: Scatter plot of intaRNA-predicted free energy values of the predicted RNA structure of each interaction in the (Dadonaite et al. 2019) WSN average of replicates dataset, plotted against log10 of reads per million (RPM) values of interactions. ***Interpretation:*** *the RPM values of interactions have a broad distribution, and the correlation with predicted free energy values, i.e. stability, of the corresponding RNA interaction structures is low.* ***Note:*** *the predicted MFE values here were taken from the Dadonaite et al. data, they were not computed using the methodology of this thesis.*

DISTRIBUTIONS OF INTER-RNA INTERACTIONS IN INFLUENZA A

The results on the inter-RNA interaction distributions across the various strains for which data is available are shown in Figure 14 to Figure 18. Broadly taken, it is clear that the distribution of inter-RNA interactions is highly non-uniform for all vRNA segments in all strains. All eight vRNA segments seem to be involved in inter-RNA interactions in the strains analysed, although not all of them seem to be involved equally. Multiple hotspots and silent regions seem to exist for each vRNA segment, and not all of these seem to be conserved across all strains.

Firstly, there are notable differences between the distributions derived from Dadonaite et al. data versus Le Sage et al. data, even when comparing the same strain (WSN). In particular, the peak corresponding to the NP hotspot is far and away the strongest in the Le Sage et al. WSN non-HSmut (intact hotspot) strain (Figure 17), whilst the peak corresponding to the NP hotspot in the Dadonaite et al. WSN strain (Figure 14) is of similar strength to that of several other hotspots, such as those in the PA and PB1 segments. This indicates, as Le Sage et al. already investigated, that the 2CIMPL and SPLASH methods employed in the two studies do not yield equal or even similar results due to as of yet unknown factors. This highlights the importance of data validation using a variety of approaches. Further research might elucidate the causes of the divergences of the two methods.

The Le Sage et al. HSmut strain (Figure 18), which features synonymous mutations in the NP hotspot nucleotides, shows a general increase in the peak size of other hotspots, a consequence of the normalisation procedure. It is also clear that the synonymous mutations in this NP hotspot induce other changes, such as the relative shrinking of the PA ~1400 nt hotspot peak. A possible

explanation is that this hotspot interacts frequently with the NP hotspot in the WSN strain. Other hotspots, such as the PA ~100 nt hotspot, could be unaffected by the NP hotspot disruption if they do not interact with that hotspot. This is the simplest explanation, but it may not be the only one, as higher-order network effects may come into play here as well. For example, the disruption of the NP hotspot may ‘free’ other interacting RNA elements, which could then interact with yet other RNA elements that normally bind the PA ~1400 nt hotspot. In a cascading effect, this could then reduce the (normalised) interactivity of that hotspot through disruption of another site. Indeed, only one interaction between this PA hotspot and the NP hotspot is found in the Le Sage et al. data on WSN non-HSmut (replicate 1), between PA 1384-1414 nt and NP 619-706 nt. This would indicate that there must be higher-order network effects at play, because the falling out of one interaction is not enough to explain the reduction in normalised peak size of the PA hotspot. The possibility of changes in (intra- and inter-)RNA structure due to the disruption of the NP hotspot can also not be discounted as a factor in the complex changes observed in the interaction network.

Looking at the figures for the three strains WSN, PR8 and Udorn (Figure 14, Figure 15, Figure 16 respectively) in the data derived from Dadonaite et al., it is clear that there are several differences in the distributions of inter-RNA interactions between these strains. The figures have been annotated using markers to show several points of interest, with cyan markers indicating peaks that are deemed to be conserved in all three strains, and black markers indicating potential silent regions (conserved or non-conserved). The blue marker indicates the NP hotspot, and the red marker indicates the PA hotspot described earlier in this section. Even though WSN and PR8 are quite closely related (Dadonaite et al. 2019), several differences are apparent between these two strains as well. An obvious discrepancy is the strong and wide peak in the NA ~900 nt region of PR8, constituting a hotspot on par with the NP hotspot in terms of peak size. In WSN, this peak is significantly lower, indicating (relatively) fewer interactions at this site. In Udorn, it is unclear whether this peak exists at all, but many other strong peaks arise in the NA segment. Between WSN and PR8, several such differences in peak strengths can be observed, but it also seems that there is a large degree of similarity in the overall distribution of peaks and silent regions. This may indicate that the interaction networks of these two strains are quite similar. From this observation, it is tempting to speculate that these two strains may be ‘compatible’ in the sense that co-packaging and therefore reassortment of segments from these two strains may be possible. The reason for the differing peak strengths in many hotspots in these two strains is not immediately obvious: this may be a consequence of the method used to find the inter-RNA interactions, or it may be a genuine reflection of differing numbers of interactions at these hotspot sites.

The differences between the H1N1 WSN/PR8 strains and the H3N2 Udorn strain are much more obvious. Many peaks and silent regions, including the NP hotspot, are still conserved to some degree. Others however have disappeared or seem to have shifted in position. It is unclear whether the PA hotspot that was mentioned before still exists, and in general the PA segment seems ‘quieter’. The pattern for the HA and NA segments seems to be completely different, including a very strong and broad new peak around ~1100 nt in NA. This seems logical, as the Udorn strain has different serotypes for both the HA and NA proteins compared to the H1N1 WSN and PR8 strains, and

consequently the underlying genotypes must be different as well. A sensible conjecture is that the co-packaging/reassortment probabilities between Udorn segments and WSN/PR8 segments are probably lower due to inter-RNA interaction incompatibilities. Dadonaite et al. investigated this to some extent, cf. (Dadonaite et al. 2019, Fig. 4), and found incompatibilities that could be recovered using silent nucleotide substitutions. This is an interesting avenue for future research, and a new study that is currently in preprint may shed some further light on the topic of reassortment probabilities and gene constellations in the context of genome packaging in influenza A (Trifkovic et al. 2021). Using a combination of the Dadonaite et al. data, influenza A sequencing data, and *intaRNA* computations, I will try to study the potential for cross-strain inter-RNA interaction later in this thesis.

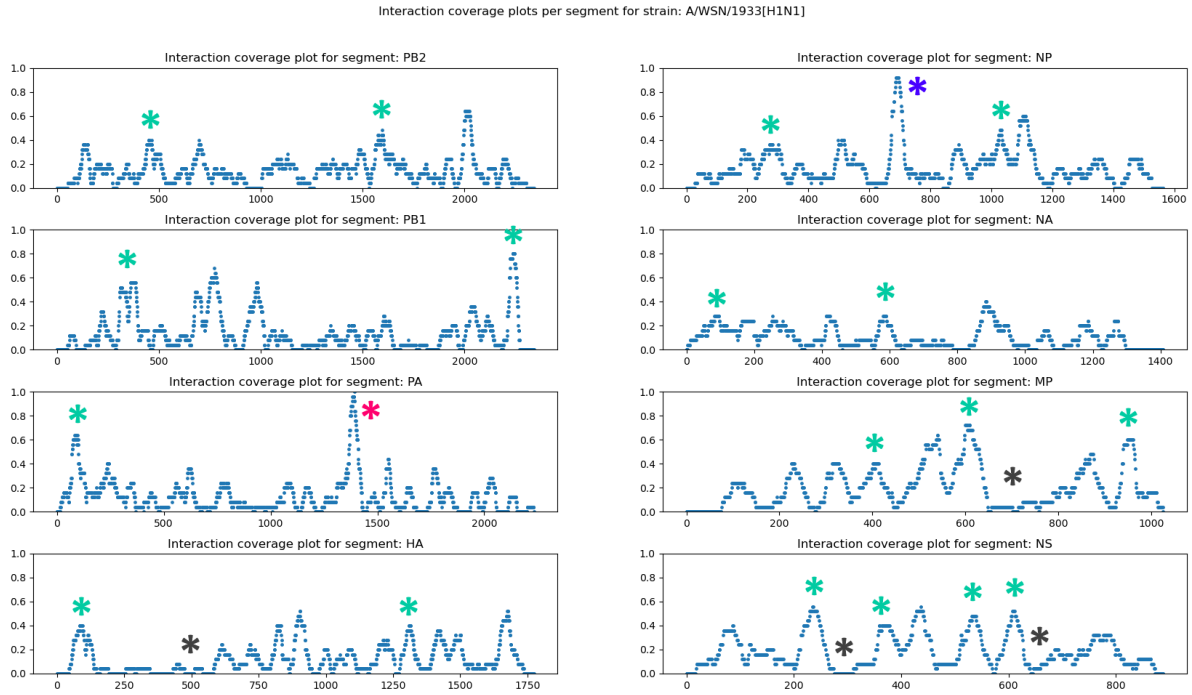


Figure 14: Interaction coverage plots per segment for strain A/WSN/1933[H1N1] (WSN), average of replicates based on data from (Dadonaite et al. 2019). The number of interactions covering each nucleotide position is shown, normalised to the highest coverage value across all segments. Nucleotides on the ***x-axis*** *are numbered from 5' to 3'.* ***Colours:*** *cyan: conserved hotspots or peaks found in all three Dadonaite strains; black: silent regions; blue: the NP hotspot; red: another strong WSN hotspot in the PA segment.* ***Markers:*** *star: feature is present in this strain; question mark: feature may be present in this strain, but ambiguity exists; X: feature is present in other strains but not in this strain.* ***Interpretation:*** *the NP hotspot is clearly visible as a high peak at around 700 nt in the bottom right (NP) plot. However, an even stronger, narrow hotspot exists at the red marker in the PA segment. Various other hotspots, lesser peaks and silent regions are seen as well, including the cyan-marked hotspots, which are conserved across the three strains Dadonaite et al. analysed. Interestingly, both the 3' and 5' ends do not seem to play a major role in terms of strong hotspots for most segments in this strain, except for the 3' end of PB1 and perhaps the 5' ends of PA and HA. In other segments, some interactions may occur in the 3' and 5' regions, but not on the scale of the stronger hotspots.*

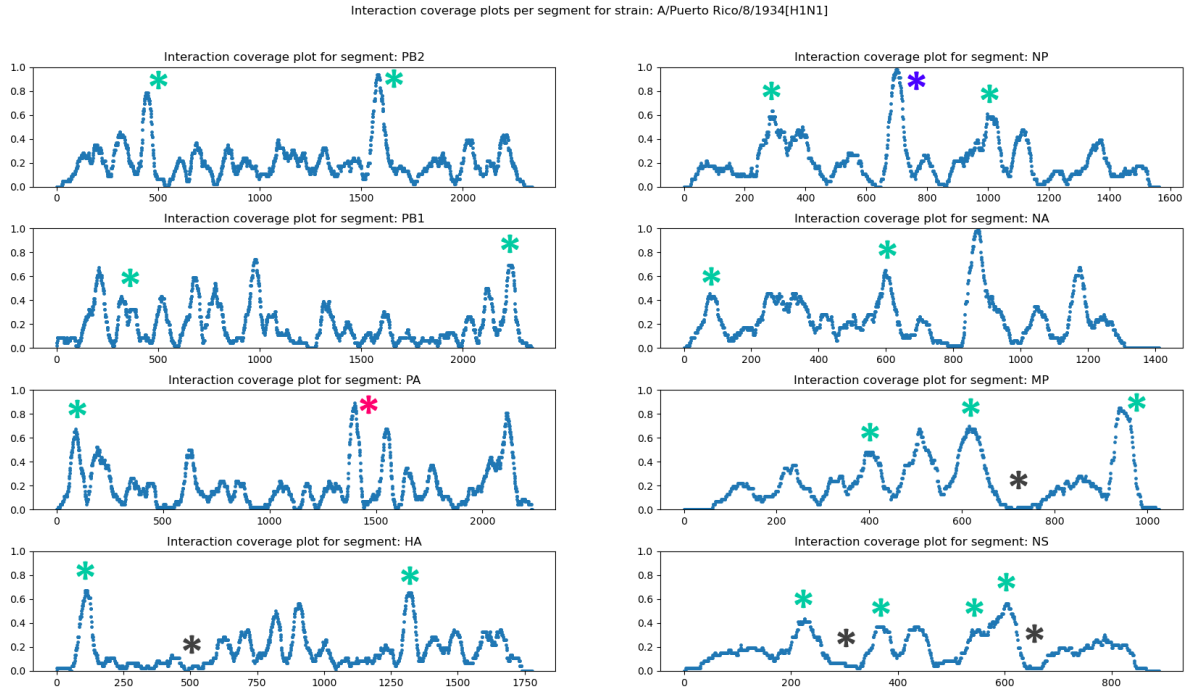


Figure 15: Interaction coverage plots per segment for strain A/Puerto Rico/1934[H1N1] (PR8) replicate 1, based on data from (Dadonaite et al. 2019). For the description of the colours and markers, see Figure 14. ***Interpretation****: the NP hotspot is clearly visible as a high peak at around 700 nt in the bottom right (NP) plot. Many features seen in the WSN distribution are conserved here.*

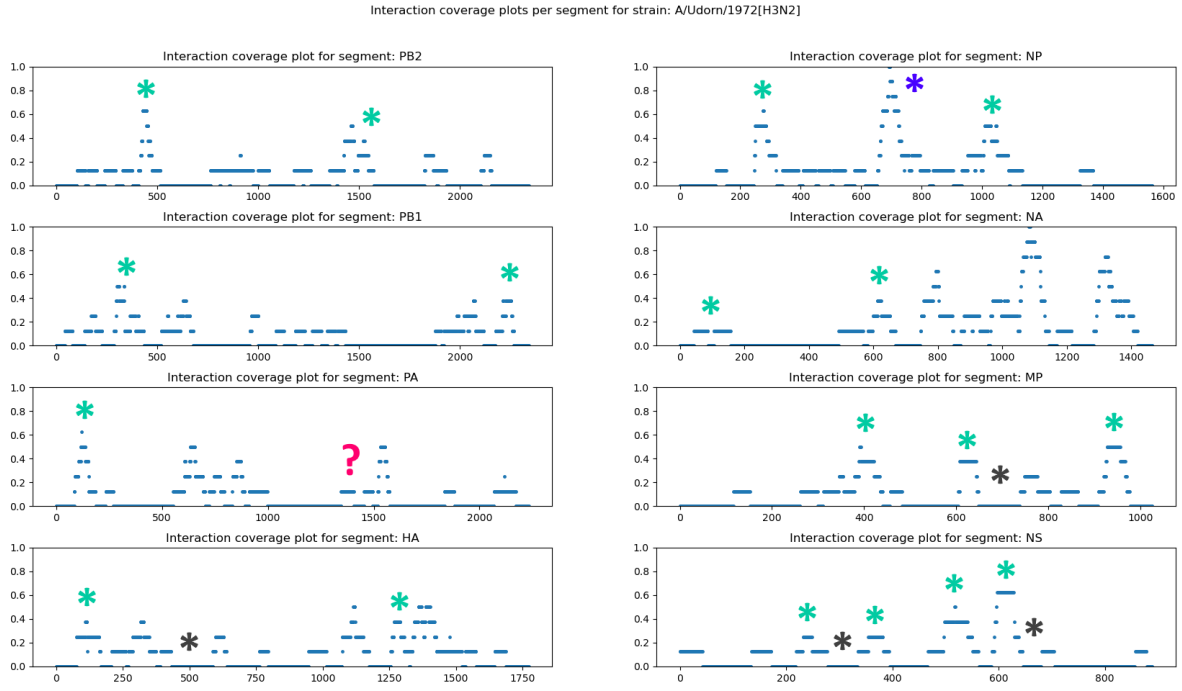


Figure 16: Interaction coverage plots per segment for strain A/Udorn/1972[H3N2] (Udorn), average of replicates based on data from (Dadonaite et al. 2019). For the description of the colours and markers, see Figure 14. ***Interpretation****: the NP hotspot is visible around 700 nt in the bottom right (NP) plot. Other hotspots are also visible, some of which are not seen in H1N1 strains, including a moderate NA ~1100 nt hotspot, and other new hotspots in the second half of the NA segment.*

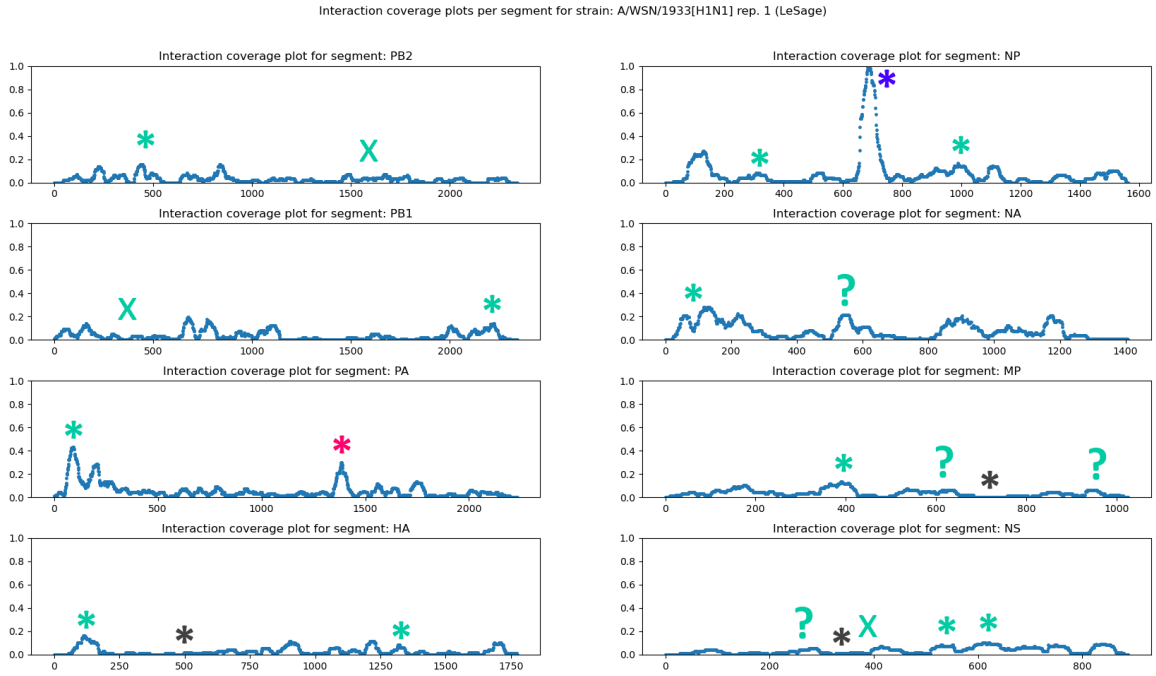


Figure 17: Interaction coverage plots per segment for strain A/WSN/1933[H1N1] (WSN), first replicate based on data from (Le Sage et al. 2020). For the description of the colours and markers, see Figure 14. ***Interpretation****: The NP hotspot is clearly visible as a high peak at around 700 nt in the bottom right (NP) plot, but other hotspots are suppressed compared to the Dadonaite et al. WSN results.*

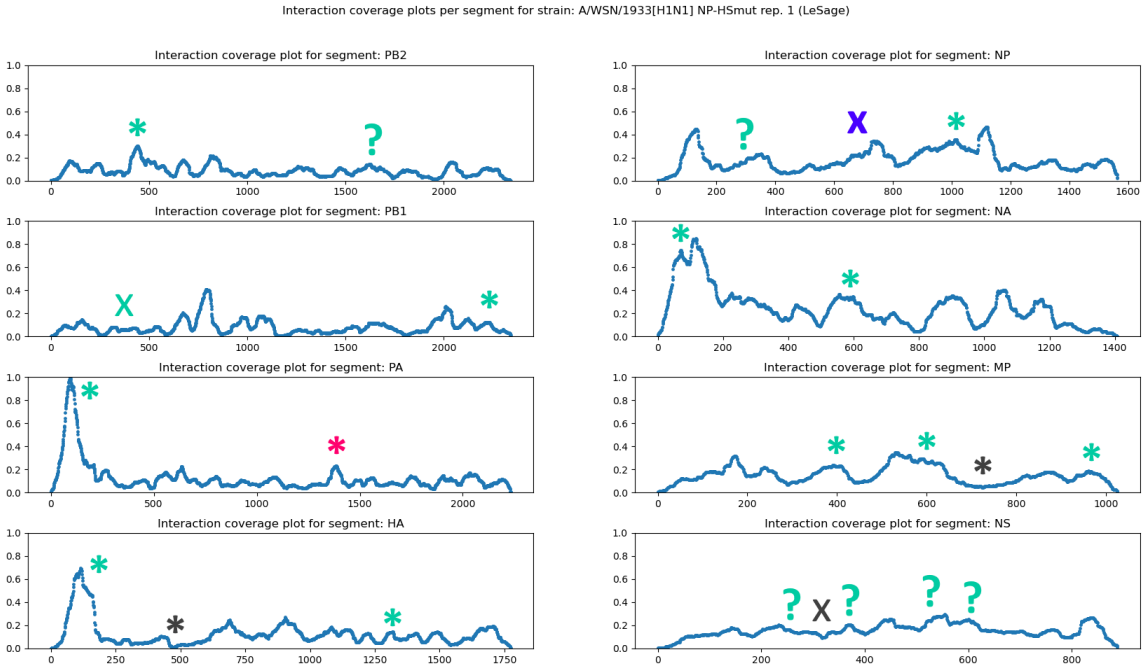


Figure 18: Interaction coverage plots per segment for strain A/WSN/1933[H1N1] (WSN) with disruption of NP hotspot nucleotides (HS-mut strain), first replicate based on data from (Le Sage et al. 2020). For the description of the colours and markers, see Figure 14. ***Interpretation****: the NP hotspot is now absent, but new hotspots have appeared, and hotspots in the non-HS-mut strain are more prominent.*

From these results, it is apparent that the degree of involvement of different vRNA segments, and of RNA regions within those segments, in inter-RNA interactions differs vastly. Some segments seem to be more involved in inter-RNA interactions than others, including when measured in average number of interactions per nucleotide, and patterns of hotspots and silent regions appear in the distributions of interactions over whole segments. These differences are strain-specific and seem to be influenced significantly by the method used to find and map the inter-RNA interactions, but large similarities also existed in the two genetically similar strains WSN and PR8 in the Dadonaite et al. dataset. It is unclear what kind of virion vRNP organisation is most fitting given these results, but the results do seem to underline the involvement of each segment in every strain in inter-RNA interactions. Again, lack of a static central organising segment seems most likely.

SEQUENCE CONSERVATION AND INTER-RNA INTERACTIONS

It is interesting to check whether there is evidence of increased conservation at the sequence level in regions which are highly involved in inter-RNA interactions (hotspots) across multiple strains. It is known that packaging signals in influenza A are conserved at the sequence level (Gog et al. 2007). Increased conservation at the sequence level can indicate an evolutionary constraint imposed by other factors such as protein conservation and perhaps even *intra*-RNA structure, and since a large number of the inter-RNA interactions take place inside coding regions, this complicates such sequence consensus analysis. Nevertheless, I constructed the consensus plots per segment based on the method of Gog et al., cf. (Gog et al. 2007, Fig. 1), with ‘consensus’ meaning the degree to which nucleotides are conserved (including deletion/insertion) at each position in the viral genome. To this end, whole viral genome sequences were gathered for WSN replicate 1 from Dadonaite et al. and 10 H1N1 vaccine strains, which were aligned (using multiple sequence alignment) and used to compute the (inverse) mean pairwise Hamming distance, a type of distance metric that was used by Gog et al. as a sequence consensus scoring metric. The 10 selected vaccine strains are shown in Box 1, excluding the incomplete A/Bayern/7/1995 and A/Solomon Islands/3/2006 strains. Bilateral sliding window averaging was applied in a 20 nt window for smoothing. The rolling consensus score per segment is overlaid with the interaction density plot of WSN as described in Figure 14. The result is shown in Figure 19.

Although the method used to determine sequence conservation is quite simplistic here, it is apparent from the results that there is variability of consensus at different vRNA genomic positions for all segments. As mentioned before, increased sequence consensus may be evidence of evolutionary constraints due to a variety of factors including codon conservation and possibly intra-RNA structure conservation. Comparing to the WSN interaction density overlay, it is unclear whether regions with high inter-RNA interaction density also constitute an evolutionary constraint, i.e. correlate with higher sequence consensus/conservation. The gold markers identify regions with high interaction density where the sequence consensus score is locally higher, supporting this hypothesis, while black markers show the opposite: regions with high interaction density where the sequence consensus score is locally lower. For the PA hotspot discussed before, displayed in blue, no significant local enhancement of consensus is seen. For the NP hotspot, a minor local increase in consensus score is observed, in support of the hotspot manipulation findings of (Le Sage et al. 2020). The significance of

this finding is low given the variability of the consensus score across every segment at this scale. It is possible that the degree of evolutionary constraint imposed, and therefore the degree of sequence-level conservation in RNA sites involved in inter-RNA interactions, is affected by the importance of the interactions at that RNA site in the overall genome packaging process.

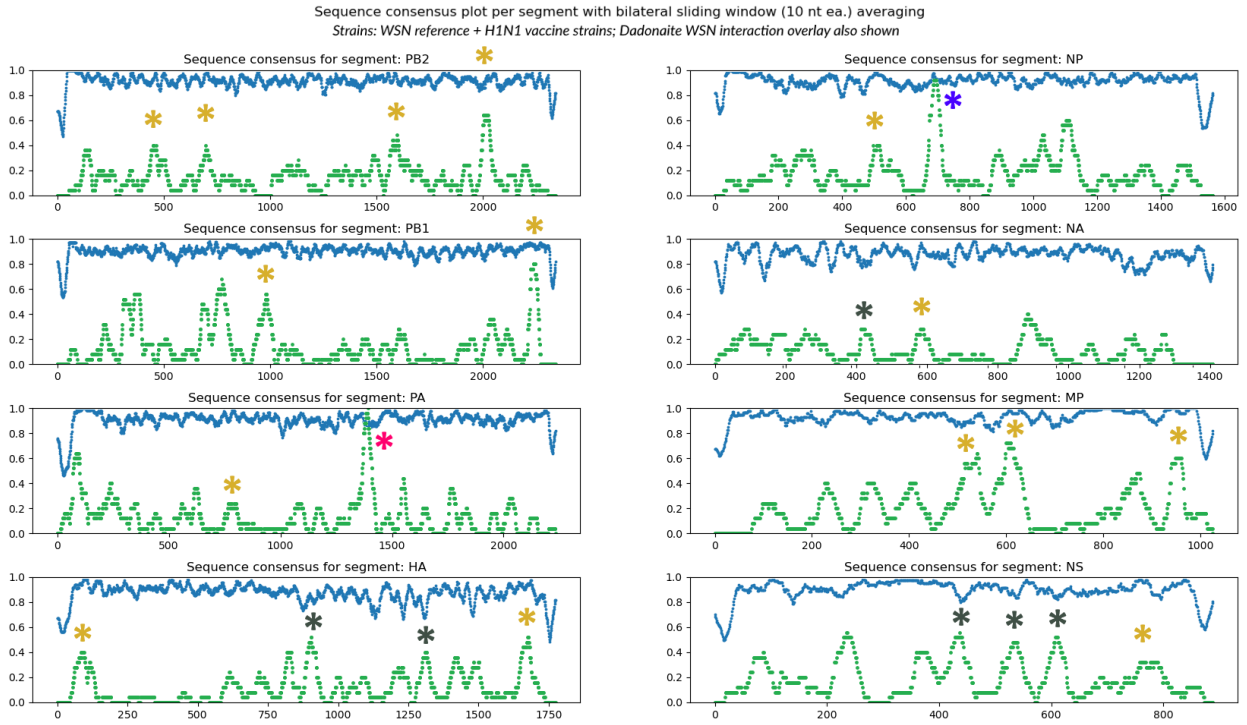


*Figure 19: Sequence consensus score per nucleotide over influenza genome segments in WSN and 10 vaccine-associated H1N1 strains using the inverse mean pairwise Hamming distance metric, averaged using a 20 nt (10 nt each side) bilateral sliding window, shown in **blue**. A higher value indicates greater consensus. The interaction density, normalised to all-segment highest value, of inter-RNA interactions in the average of replicates data on WSN from (Dadonaite et al. 2019) is also shown in **green** for comparison. **Markers:** blue: NP hotspot; red: PA hotspot; gold: several sites where a peak in the interaction density coincides with (locally) higher consensus; black: several sites where a peak in the interaction density coincides with (locally) lower consensus. **Interpretation:** there is variability in sequence conservation across the various gene segments. It is unclear whether regions with high inter-RNA interaction density display increased sequence conservation, with some regions supporting this hypothesis, and other contradicting it. The low consensus regions around segment ends are mostly due to sequencing issues and not due to genuine lack of consensus in those segment ends.*

Extrapolation of A/WSN/1933 RNA-RNA interactions to other strains

So far, I have only looked at existing data on inter-RNA interactions in influenza A from the (Dadonaite et al. 2019) and (Le Sage et al. 2020) studies. These experiments were able to derive regions of the vRNA between which inter-RNA interactions may be occurring. They did this for several well-known strains of influenza A of the most common (in humans) serotypes, H1N1 and H3N2. For this thesis, I extrapolated the data on these interacting vRNA regions to other strains that have not yet been analysed in this way, with the aim of deriving information about inter-RNA interactions in these strains. Especially interesting would be to look at conservation at the sequence- and structure level of these interactions in various influenza A strains, including different serotypes such as H5Nx strains, which have garnered attention due to their possible risk in causing dangerous outbreaks. The purpose of this is to gain further insights into similarities and differences in inter-RNA interactions in various strains of influenza A, with a view of acquiring knowledge relevant to the open questions of the influenza genome packaging process and reassortment.

CURSORY ANALYSIS OF WSN-H1N1 VACCINE STRAIN EXTRAPOLATED INTERACTIONS

The interaction extrapolation pipeline was performed initially for the Dadonaite et al. WSN reference strain with the H1N1 vaccine strains as the extrapolation set (cf. Box 1). Here, I will show some insights gained from the analysis of the extrapolation results.

Noting that there are over 600 discrete inter-RNA interactions in the Dadonaite et al. average of replicates WSN dataset, with widely differing RPM values (cf. Figure 13), a global overview of the interactions and their extrapolations is prudent. Shown in Figure 20 and Figure 22 are a histogram of the distribution of predicted MFE values for the reference strain (WSN) and a plot of the reference strain predicted MFE values versus RPM for each interaction, respectively. Somewhat surprisingly, the predicted MFE values seem to be almost normally distributed, with the mean of the distribution situated somewhere between -11 kcal/mol and -13 kcal/mol. The strongest interaction in terms of predicted MFE is the *2 MP 382-420 NS 605-631* interaction with -27.14 kcal/mol. The interaction notation can be broken down as follows: ranked *2 + 1 = 3rd* in terms of RPM in Dadonaite et al., between the MP segment at 382 to 420 and NS segment at 605 to 631, measured from the 5’ end. This is followed by: *0 PA 101-135 HA 831-863* with -24.46 kcal/mol and *31 NA 981-1011 MP 937-967* with -23.80 kcal/mol. The weakest predicted interaction in terms of MFE is: *187 PB1 2225-2269 PA 778-812* with just -0.62 kcal/mol, barely below the +0.00 kcal/mol threshold. Oddly, eight interactions did not have any predicted significant interaction, contrary to the result in the Dadonaite et al. intaRNA run. Note that those authors most likely used an older version of the algorithm, with differing parameters and including accessibility constraints. The lack of significant interaction structures found in those eight interactions is therefore probably due to technical (algorithmic) differences, but it should be noted that these interactions had poor MFE and low RPM values in the Dadonaite et al. data as well. Significant differences between the Dadonaite et al. *intaRNA* results and the results here were also observed for interactions for which significant structures did exist, but

this is not a large concern as long as the same method and configuration is employed within the scope of this thesis for all interactions and all strains.

From Figure 22 (compare to Figure 13) it is again apparent that (the logarithm of) RPM values is not a good predictor for the MFE values of interactions, but there may be some overall correlation of predicted MFE value with the overall bulk of low-RPM interactions versus the rarer high-RPM interactions. In particular, the very lowest (best) MFE interactions do seem to be associated with higher RPM values, although there are counterexamples outside the top five or so. This probably indicates that the (predicted) stability of inter-RNA interactions is far from the only factor influencing how likely that interaction is to arise and be maintained in the influenza genome, assuming RPM is a good predictor of interaction occurrence. Interaction site length was considered as a factor of influence in interaction RPM/MFE, but no significant correlation was found with either.

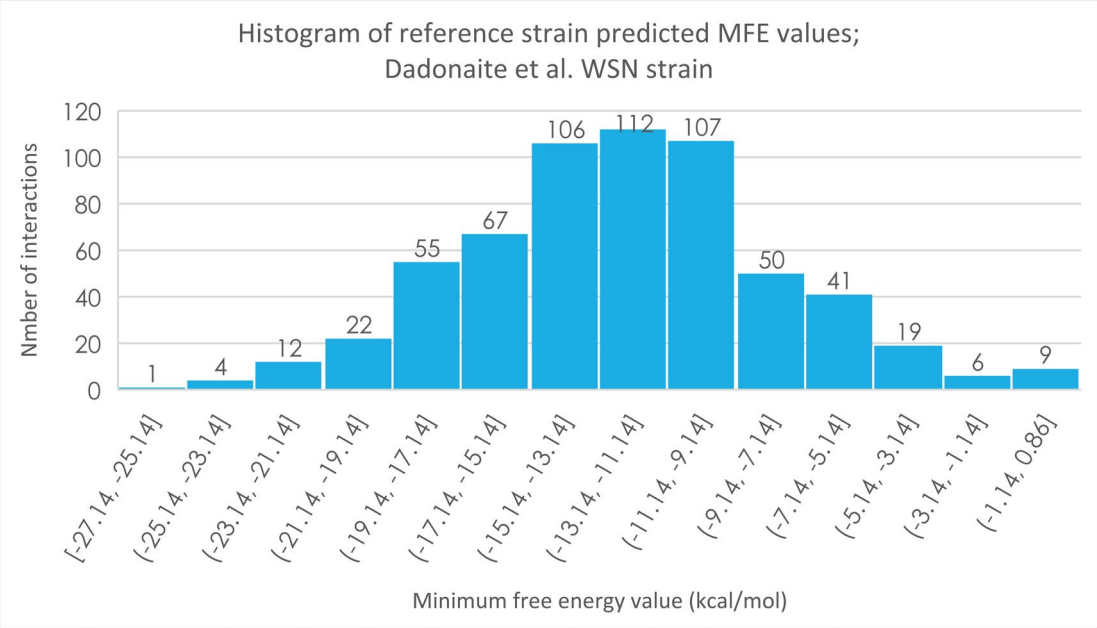


Figure 20: Distribution of intaRNA predicted minimum free energy (MFE) values for inter-RNA interactions of Dadonaite et al. WSN strain (average of replicates). ***Interpretation:*** *the MFE values follow a broad, approximately normal distribution.*

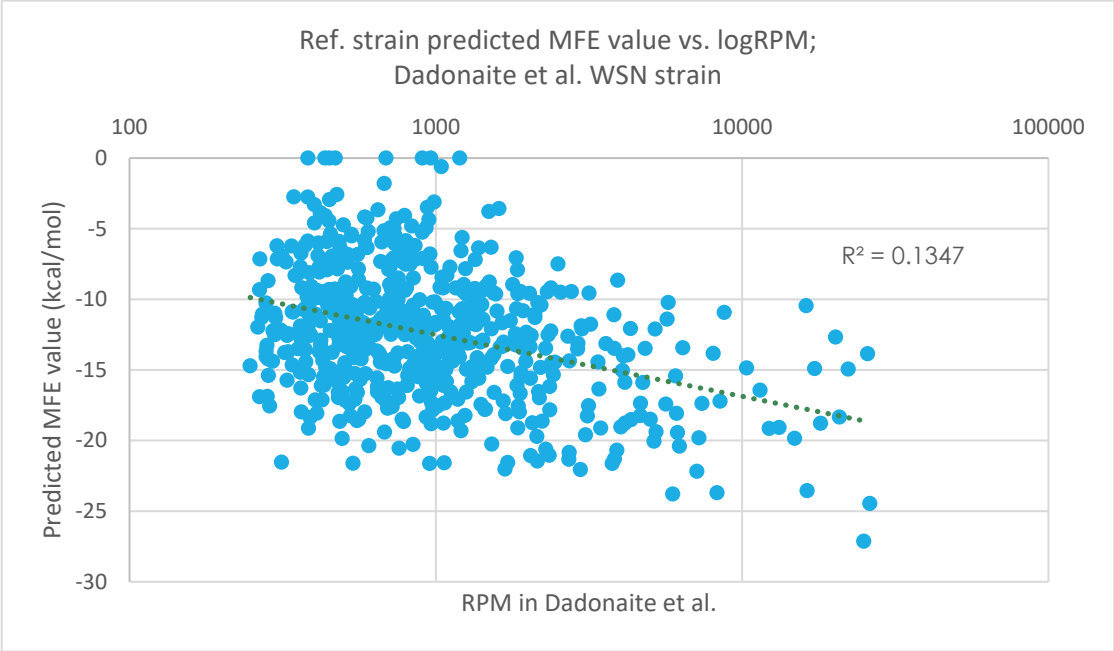


Figure 22: Predicted minimum free energy (MFE) values vs. (base 10) logarithm of reads-per-million values for inter-RNA interactions of Dadonaite et al. WSN strain (average of replicates). (Logarithmic) trendline shown with R^2 value. ***Interpretation:*** *the predicted MFE values are poorly correlated with RPM, but some correlation may exist for high-RPM interactions vs. low-RPM interactions in general.* ***Note:*** *this is a recreation of Figure 13 using predicted MFE values from the pipeline of this thesis.*

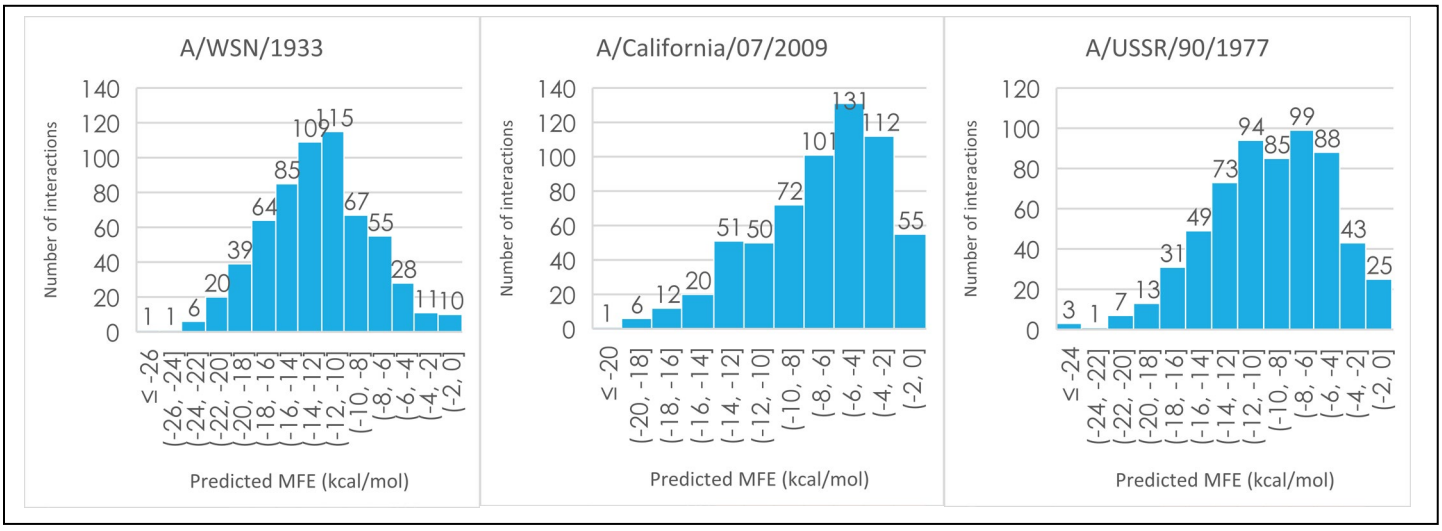


Figure 21: Predicted minimum free energy (MFE) distributions for the WSN reference strain (cf. Figure 20) and two H1N1 vaccine strains. California/2009 is associated with the 2009 H1N1 pandemic. Interactions for which no significant structure was found are mapped to the bin including zero. ***Interpretation:*** *the free energy distribution clearly shows that most interactions are predicted to be most stable in WSN, with an overall moderate positive-ward stability shift in USSR/1977 and a bigger shift in California/2009.*

So far, only interactions in the reference strain (WSN) have been analysed, with no extrapolation to new strains. In Figure 21, the distributions of predicted MFE values are shown for all WSN-based interactions for the reference strain and two extrapolation strains. The figures clearly show a shift in stability distributions in the extrapolation strains, with an especially strong shift in the 2009-pandemic-associated A/California/07/2009 strain. Assuming that inter-RNA interactions are important in maintaining genome packaging in influenza A, this is probably due to changes in the interaction

network, instead of a genuine overall weakening of interactions. In Figure 23, the distributions of predicted MFE shifts with respect to the reference strain are shown. It is apparent that most interactions are predicted to be less stable in the extrapolation strains, especially for A/California/07/2009. A significant portion of interactions is predicted to have a higher MFE in the extrapolation strains however, even though such negative shifts are usually comparatively small.

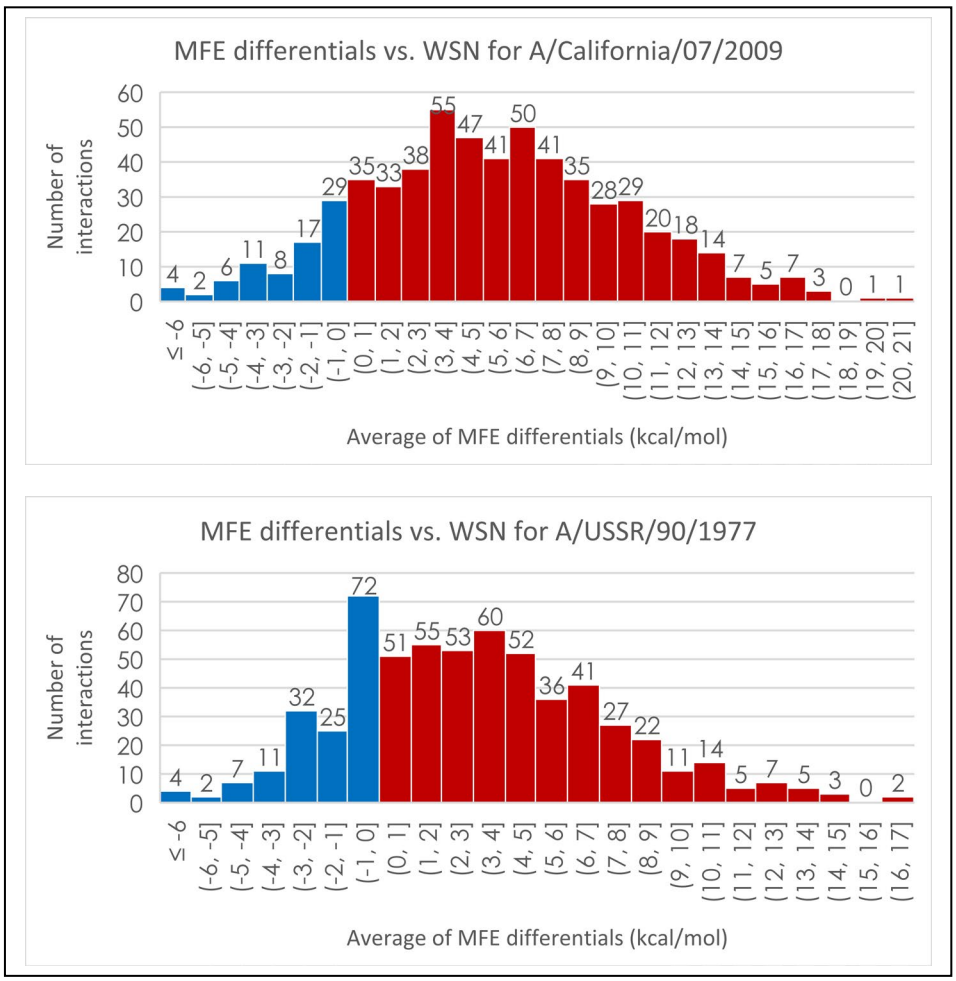


*Figure 23: Distributions of predicted minimum free energy (MFE) differences with the reference strain (WSN) for A/California/07/2009 and A/USSR/90/1977. Negative shifts (lower predicted MFE, including zero shift) in **blue**, positive shifts in **red**. Interactions for which no significant structure was found in WSN and/or the extrapolation strain are not binned. **Interpretation:** most interactions are predicted to become weaker in the two extrapolation strains. The shift is stronger for California/2009, with a mean shift of +5.18±4.83 kcal/mol (1σ), vs. +2.98±4.04 kcal/mol for USSR/1977. Not all interactions are predicted to become less strong, for California/2009, 77 out of 585 significant interactions (13.2%) are predicted to decrease in MFE. For USSR/1977, this is true for 138 out of 582 significant interactions (23.7%). Here 'significant interaction' means that a stable interaction structure below +0.00 kcal/mol was found for that strain. The dataset contains 611 interactions in total. **Note:** shifts of +0.00 kcal/mol in the [-1, 0] bin are common due to cases of no mutations in the interaction sites.*

Due to the large number of discrete interactions in the dataset, it is difficult to discuss and display information on all of them. A selection of interactions was made for the purpose of closer analysis: the twenty interactions that were conserved in WSN, PR8 and Udorn in the Dadonaite et al. data, according to their method of determining interaction overlap across strains. These interactions are shown in Table 4.

*Table 4: Interactions found in all three of WSN, PR8 and Udorn in the Dadonaite et al. dataset. The interaction 'index' is the position of that interaction in the WSN average of replicates dataset, sorted by RPM value, starting at 0. NP hotspot sites are marked in **green**. **Plot index** is the index of that particular interaction in Figure 24. Also given are the MFE values for WSN and Udorn, calculated using the intaRNA pipeline. The WSN values were calculated using Dadonaite et al. WSN average of replicates, the Udorn values using Dadonaite et al. Udorn replicate 1.*

Interaction (WSN RPM indexed)	**Plot index**	**WSN predicted MFE** (kcal/mol)	**Udorn predicted MFE** (kcal/mol)
2 MP 382-420 NS 605-631	0	-27.14	-28.10
13 PB2 1463-1513 NP 656-698	1	-14.87	-10.38
28 NP 670-712 MP 355-395	2	-12.12	-14.17
36 NP 675-717 MP 400-434	3	-13.94	-14.17
45 NP 1319-1365 MP 941-983	4	-20.69	-24.76
53 MP 116-148 NS 510-542	5	-17.37	-12.49
92 PB2 112-152 PA 1540-1566	6	-16.67	-14.75
98 PA 612-652 NP 689-731	7	-9.48	-12.43
103 PA 112-164 MP 932-964	8	-15.49	-13.95
149 PB1 1925-1959 NP 672-706	9	-14.78	-18.33
153 PA 611-645 NP 782-812	10	-16.60	-14.87
183 PB1 345-393 MP 399-445	11	-11.67	-11.85
197 NA 561-605 MP 601-641	12	-11.80	-18.59
241 HA 583-625 NP 667-701	13	-11.65	-21.70
304 PB2 464-508 PA 1453-1489	14	-12.40	-18.86
323 NP 1087-1123 NS 504-546	15	-1.81	-9.51
345 NP 350-384 NS 600-630	16	-16.27	-17.78
352 NP 953-1003 MP 761-817	17	-4.99	-8.59
396 PB2 2082-2154 NA 582-616	18	-10.24	-16.82
539 PB2 1057-1105 HA 1461-1505	19	-5.89	-7.72

Figure 24 shows an overview of the predicted MFE value differences for the extrapolated strains with respect to the WSN reference strain for these interactions. This plot shows that the patterns of change of MFE of predicted inter-RNA structures are not uniform across interactions, with some interactions becoming more stable in terms of predicted MFE and some becoming less stable. For some interactions, the variance of predicted MFE is high across strains, perhaps indicating decreased sequence conservation at these sites, while for some, the variance is lower. A combination of low variance and increased stability in all extrapolation strains is found for the interaction at index 2, which corresponds to *28 NP 670-712 MP 355-395*. The combination of low variance and increased stability might indicate that this interaction became more important in various strains than it was in WSN. However, the low correlation of predicted MFE with RPM values as shown earlier means caution must be exercised in making such statements. Other interactions, such as *92 PB2 112-152 PA 1540-1566* (plot index 6) are not conserved at all in terms of free energy, having an MFE of around -16.67 kcal/mol in WSN, increasing by more than +5 kcal/mol in all extrapolated strains (with an average increase of +7.53 kcal/mol). This indicates that this interaction may have become less stable in the diverse set of strains analysed, which may mean that this interaction is less important for the overall packaging process, and/or that the RNA regions involved in this interaction are subject to other evolutionary pressures. There are other interactions for which the same logic applies, e.g. *304 PB2 464-508 PA 1453-1489* (plot index 14), which surprisingly is in

the PA ~1400 nt hotspot region. Another interesting interaction is *2 MP 382-420 NS 605-631* (plot index 0), coinciding with triple-conserved peaks in both segments in Figure 14, which has one of the lowest overall MFE values in the reference strain WSN: -27.14 kcal/mol. In the extrapolated strains, there is a large spread of MFE values, including some strains (A/Singapore/6/1986 and A/USSR/90/1977) for which this interaction becomes even more stable, and some interactions for which it becomes much less stable, most notably A/California/07/2009 and A/Michigan/45/2015, both of which are 2009-pandemic- or post-pandemic strains found in North America. In conclusion, even patterns in predicted MFE values can bring about interesting discussions concerning inter-RNA interactions and their conservation, but other techniques of analysis are required for a more in-depth study.

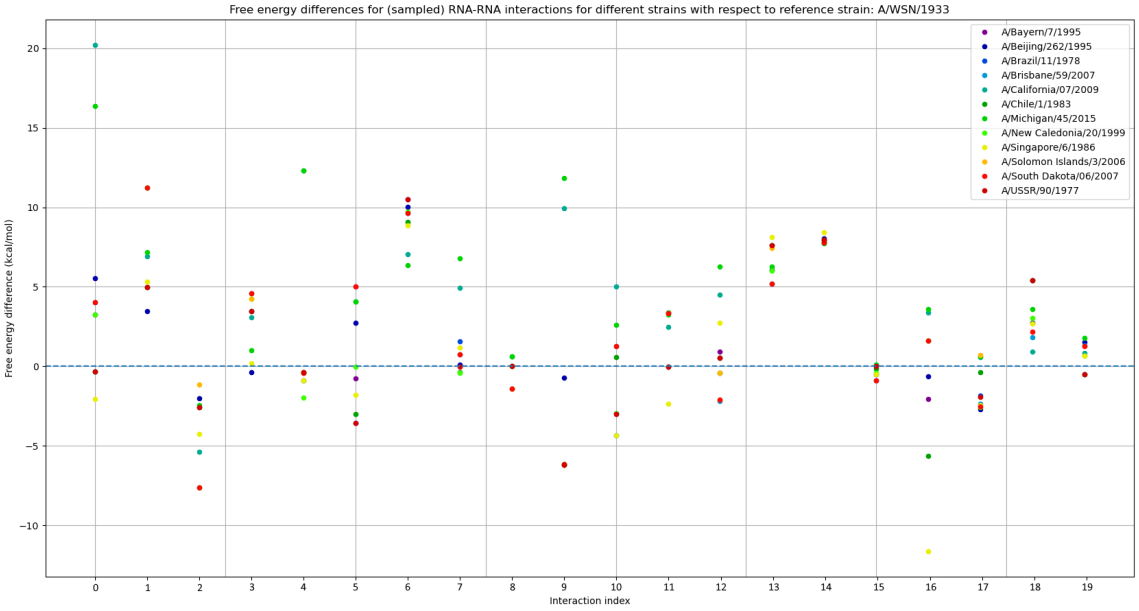


Figure 24: Plot of predicted mean free energy (MFE) value differences from the reference strain (Dadonaite et al. WSN) for H1N1 vaccine strains for 20 triple-conserved interactions. The blue line at +0.00 kcal/mol indicates the no-change reference strain baseline. ***Interpretation:*** *for some strains, the predicted MFE values in the new strains seem to increase, while for others, it decreases. The variance also differs greatly between discrete interactions. This indicates that some interactions may be conserved in terms of structure stability preservation, while others may even become more stable, but yet others are not conserved and/or become less stable.* ***Note:*** *the interaction indices do not correspond directly to Dadonaite et al. supplement dataset numbering, please refer to Table 4.*

INTER-RNA STRUCTURES

Zooming in on some of these interactions can provide interesting insights. Shown in Figure 25 is a selection of predicted structures for the *28 NP 670-712 MP 355-395* interaction for the WSN reference and several extrapolated H1N1 strains. This interaction seems to have a form of structural conservation in that a central loop (containing ‘UUACCA’ 3’ to 5’ in WSN on the MP segment) is surrounded by short regions of largely uninterrupted binding on both sides. However, one feature is most striking: the existence of a repeating pattern (‘CAGG’ 5’ to 3’) in the NP interaction site, which is most apparent as a quadruple tandem repeat in the WSN structure, covering the entire predicted interaction (WSN NP 685-697 5’ to 3’). This repeating pattern is somewhat conserved in the other selected strains. Its complementary pattern (‘GUCC’ 3’ to 5’) occurs several times in the MP

interaction site for most strains, but usually not in tandem repeat. Very interesting is the fact that this interaction involves the NP hotspot region. It may be the case that this tandem 'CAGG' repeat is involved in the stability and mutational robustness of this hotspot region.

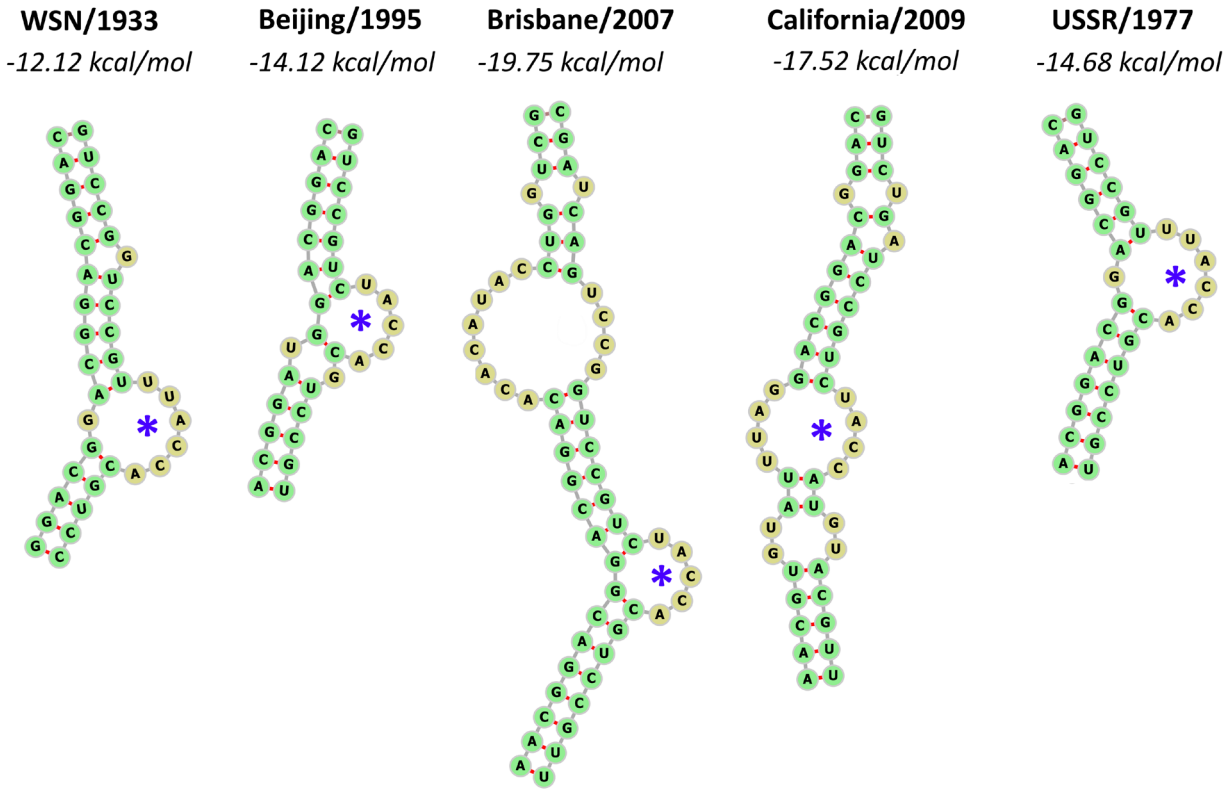


*Figure 25: A selection of predicted inter-RNA structures for the 28 NP 670-712 MP 355-395 interaction in the A/WSN/1933 reference strain (leftmost) and several extrapolated H1N1 strains. The NP interaction site (5' to 3' top to bottom) is the strand on the **left**, the MP interaction site (3' to 5' top to bottom) is the strand on the **right**. Visualisation was done using the ViennaRNA forna tool. Also shown are the free energy values of the predicted interaction structures. The **blue marker** shows the site of a conserved loop as a point of orientation. **Interpretation**: there seems to be some form of conservation of the 'UUACCA'-containing (in WSN) loop across the selected strains, with binding sequences on both sides of that loop. In the more modern strains (A/Brisbane/59/2007 and 2009-pandemic A/California/07/2009), the overall stability of the interaction seems to have increased. In the NP segment (left) interaction site for WSN, a quadruple repeat of 'CAGG' (top to bottom, here 5' to 3') occurs, covering the entire predicted interaction site sub-sequence. This pattern is somewhat, but not completely, conserved among the other strains. The complementary pattern of 'GUCC' (3' to 5') also appears several times, but not in unbroken tandem. The 'CAGG'-repeats could allow for shifting of the interaction along the repeating region, and this could confer some form of robustness of the interaction to mutations in the region. Also note that this repeating pattern is in the NP hotspot region. **Note:** alternative 'G-U' base pairs are not displayed.*

In Figure 26, the same selection of strains for the *2 MP 382-420 NS 605-631* interaction is shown. This interaction is ranked third highest in terms of RPM in the Dadonaite et al. WSN average of replicates dataset. As discusses earlier, it occurs between two small peaks in the interaction distribution of WSN (cf. Figure 14). In WSN, the MFE of the predicted optimal interaction structure is one of the lowest. Combined with the long uninterrupted binding region marked with an orange box, this could indicate that this interaction is stable and common in WSN, underlining its importance. The long binding region and low overall MFE are conserved in most other strains (including ones not shown), except for the 2009-pandemic A/California/07/2009 strain. It is also interrupted in the post-pandemic A/Michigan/45/2015 strain. This seems to be due mostly to mutations in the NS segment,

possibly attesting to a different evolutionary path or origin of this segment compared to the other H1N1 strains. As shown in Figure 13, predicted MFE for an interaction structure is in general not a good predictor of interaction RPM, but the large +20.23 kcal/mol jump between WSN and A/California/07/2009 most likely does have an effect on the stability and occurrence of this interaction in the 2009 pandemic and post-pandemic strains. It is possible that the mutations in the NS segment are due to other evolutionary pressures, and that the interaction network shifted to accommodate the resulting loss of stability within this 2 MP NS interaction that is so stable in other H1N1 strains. A comprehensive study on the origins of 2009 pandemic strains indicates that the NS segment in 2009 pandemic precursor strains has its origins in H3N2 strains, whilst the MP segment likely comes from avian H1N1 strains, supporting the analysis presented here (Smith et al. 2009, Fig. 8).

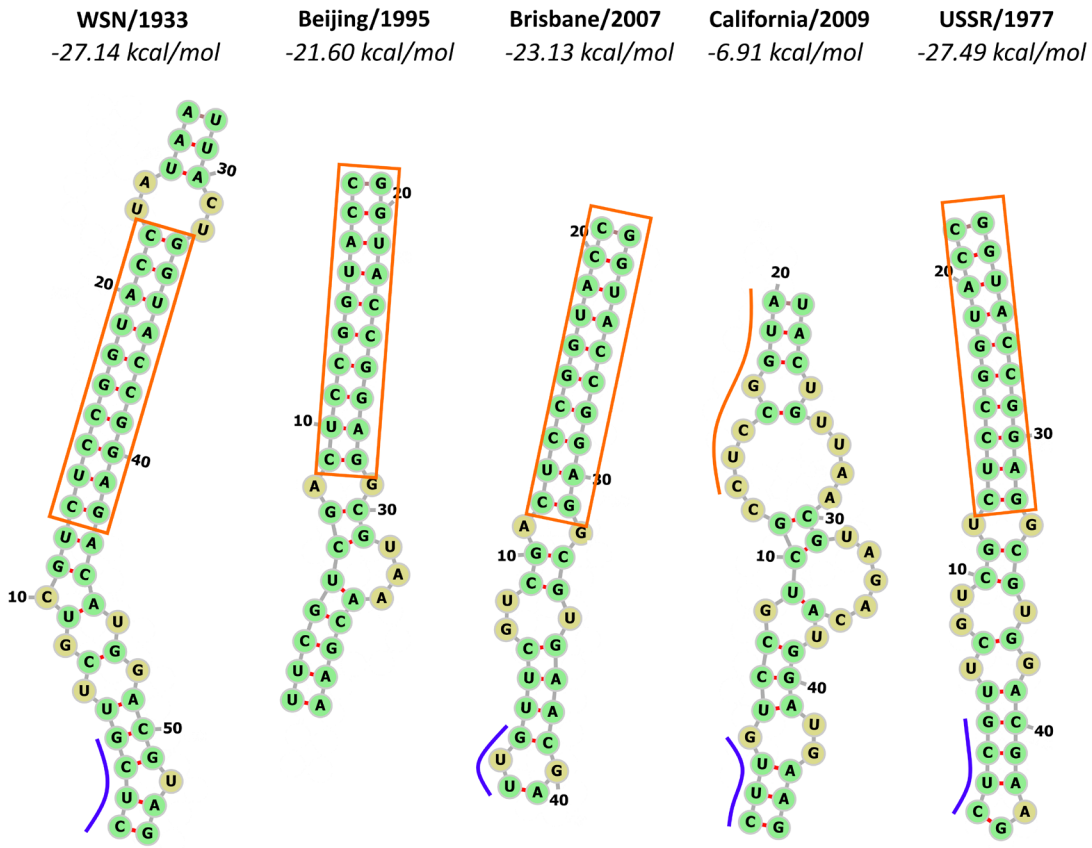


*Figure 26: A selection of predicted inter-RNA interaction structures for the 2 MP 382-420 NS 605-631 interaction in the A/WSN/1933 reference strain (leftmost) and several extrapolated H1N1 strains. The **left** strand is MP (5' to 3' top to bottom), the right **strand** is NS (3' to 5' top to bottom). Visualised using the ViennaRNA forna tool. The **blue** line gives an orientation point for the MP strand. The **orange** box/line marks a mostly conserved uninterrupted binding region. **Interpretation:** for one of the strongest interactions in terms of RPM- and MFE values, the long binding motif marked by the orange box is highly conserved among selected strains, except in the 2009-pandemic A/California/07/2009 strain due to heavy mutations in the NS interaction site. This may indicate that the NS segment in that strain has a different origin, and that this interaction is no longer as relevant in the 2009-pandemic and post-pandemic H1N1 strains. **Note:** alternative 'G-U' base pairs are not displayed.*

Sequence- and structure conservation

As described in the Sequence- and structure consensus section of the Methods, a method is available to determine nucleotide-resolution consensus profiles in the sets of reference- plus extrapolation strains, both at the sequence level as well as at the structure level. Using this method, a closer look can be taken at some of the interesting patterns seen in the previous section.

The *28 NP 670-712 MP 355-395* interaction, which overlaps the NP hotspot region, now abbreviated as *28 NP MP*, seemingly had an open loop (of varying sizes) in the MP segment surrounded by binding regions in the structures in Figure 25. In the middle plot (per-nucleotide structural function consensus) of Figure 27, the open loop in the MP segment (bottom) as seen in the predicted structure for WSN is formed by nucleotides 11 to 16 ('ACCAUU') from the left, which corresponds to a highly conserved structure (orange line) except for the last nucleotide, which engages in an alternative 'G-U' base-pair. In the structure-aligned consensus profile on the bottom, this is not the case, probably due to shifts in the interaction alignments (which can be due to changes in loops in the interactions, or due to interaction frame shifts). In the top plot, which gives the sequence consensus profiles for the interaction sites on both segments, one pattern is most apparent: tandem repeats of triplets, where the third nucleotide has a lower consensus score than the other two. This is due to the fact that mutations in the third nucleotide in gene codons are more likely to be silent (not result in an amino acid change), so this effect will be seen in any interaction site overlapping with a protein-coding region. The lower conservation pressure on the third nucleotides could also mean that these nucleotides are influenced more by other evolutionary constraints, such as inter-RNA interaction conservation/changes. Nucleotide 14 on the NP segment and nucleotide 16 on the MP segment both have very low consensus scores. For the NP 14 nucleotide, this seems to result in a loss of an interaction nucleotide binding, but the MP 16 nucleotide mutations result in an upgrade of a G-U alternative base pair to a full base pair in most structures. Manual analysis in this manner of mutations at the nucleotide scale and their effects on predicted interaction structures is difficult, as changes in the predicted structures from a reference structure are conferred by the whole ensemble of mutations in a complex manner. Looking at larger patterns is therefore more prudent.

The most interesting feature of the *28 NP MP* interaction as discussed in the previous section was the existence of the quadruple tandem repeat of 'CAGG' nucleotides (5' to 3') in the NP hotspot predicted binding region. This is shown in the top plot in Figure 27 as nucleotides 13 to 28 (incl.) in Sequence 1, coinciding with the blue background frame. In this region, there are few mutations even in the codon third nucleotides, except at index 14. This suggests that the 'CAGG'-pattern in the NP hotspot is at least somewhat conserved among H1N1 strains, underlining its potential importance in maintaining inter-RNA interactions. The only really common mutation in this region is G699A (index 25) in the last repeat on the 3' end, which occurs in most other H1N1 vaccine strains. The role of the 'CAGG'-pattern in NP hotspot inter-RNA structure is further underlined by its predicted involvement in nucleotide binding in various strains for this interaction, as well as in the *36 NP 675-717 MP 400-434* and *98 PA 612-652 NP 689-731* interactions. Another remarkably conserved region lies just outside this tandem repeat (to the right, in the 5' direction) of the NP segment, stretching for about ~10 nt.

Perhaps this region is also involved in the NP hotspot interaction(s), although *intaRNA* does not include it in the predicted optimal structure for the *28 NP MP* interaction specifically, or for other strong hotspot interactions such as the highly similar (in terms of localisation) *36 NP 675-717 MP 400-434* interaction.

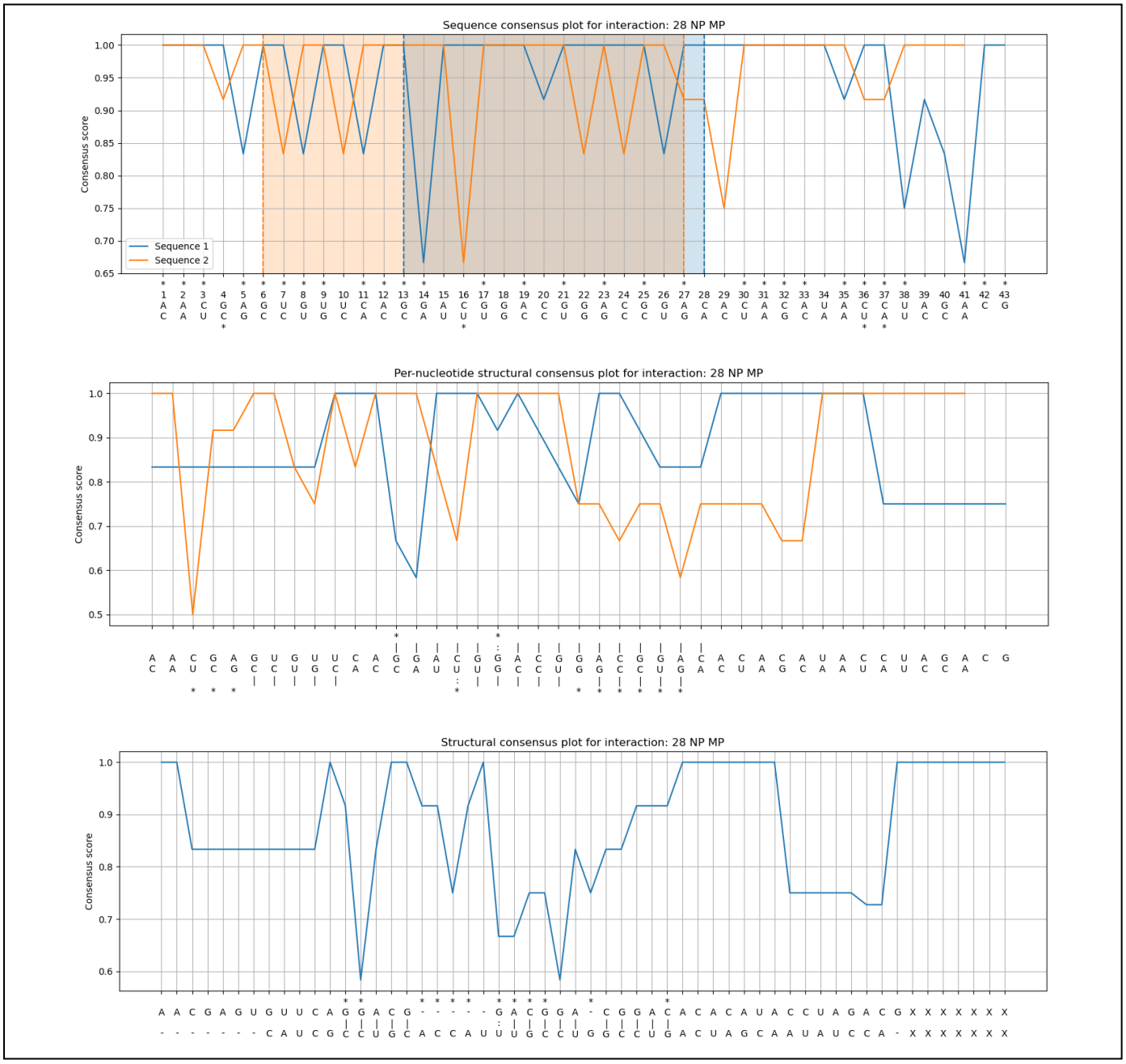


Figure 27: Consensus profiles for the WSN 28 NP 670-712 MP 355-395 interaction with WSN + H1N1 vaccine strains. ***Top:*** *sequence consensus profiles for both interacting regions.* ***Middle:*** *per-nucleotide structural function consensus profiles for both interacting regions. The blue- and orange windows denote the locations of the predicted interacting sub-sequences. NP segment in* ***blue****; MP segment in* ***orange****.* ***Bottom:*** *structure-oriented consensus profile. Gaps are denoted by a '-', and characters outside the interaction window for WSN are marked using an 'x'. The* ***x-axis*** *shows the sequence or (predicted) structure in WSN, with an asterisk marking positions where a different variant is more frequent in the other strains, to distinguish high-consensus positions that disagree with WSN. On the* ***x-axes****, the NP strand is on top (in reverse direction, i.e. 3' to 5'), and the MP strand (5' to 3') is at the bottom. The consensus score in each plot at each position was computed using the entropy method. Sliding window averaging was not used here.*

For the *2 MP 382-420 NS 605-631* interaction, from now on abbreviated as *2 MP NS*, the most notable conserved feature in Figure 26 was the long binding region (~10 nt, orange box). In Figure 28, this region lies between positions 16 and 25 in the structure consensus profile (bottom plot), corresponding to nucleotides 17 to 28 (measured within the plot bounds) in the sequence consensus profile (top plot) for MP, and nucleotides 11 to 22 for NS. The central part of this region seems to be well-conserved in the bottom structure consensus profile, although shifts in the interaction alignment may influence this. In the sequence consensus plots, the 17 to 28 region is relatively well conserved, except at position 18. In the NS segment 11 to 22 region, there is a little more variation, almost exclusively at codon third positions. Except for the MP 18 U – NS 12 A base pair, all base pairs in this region are relatively well conserved in the structure consensus plot in the middle. In conclusion, the region of strong uninterrupted base pairing seen in the predicted structure for WSN and several other strains in Figure 26 is generally well conserved among H1N1 strains, further highlighting the potential importance of this interaction and of this structural element in the stability of this interaction. However, some sequence mutations and corresponding structural variations do occur in this region, indicating some level of robustness to such mutations.

Overall, analysis of sequence- and structure consensus at the nucleotide level is possible, but it is very difficult to interpret the small-scale variations seen here in a justified manner, taking care so as to avoid over-interpreting such patterns.

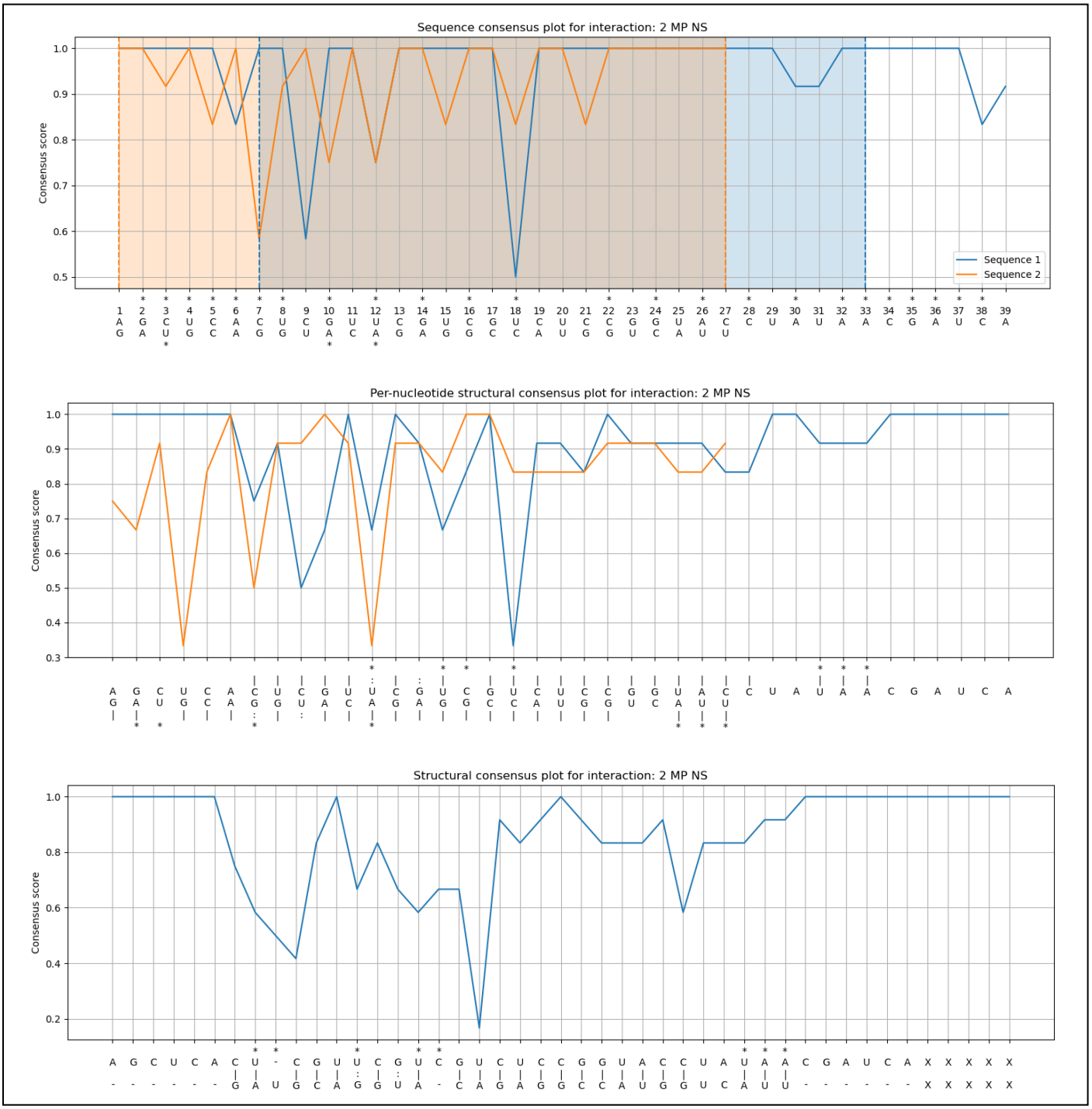


Figure 28: Consensus profiles for the WSN 2 MP 382-420 NS 605-631 interaction with WSN + H1N1 vaccine strains. ***Top:*** *sequence consensus profiles.* ***Middle:*** *per-nucleotide structural function consensus profiles for both interacting regions. MP segment in* ***blue****; NS segment in* ***orange****.* ***Bottom:*** *structure-oriented consensus profile. On the* ***x-axes****, the MP strand is on top (3' to 5'), and the NS strand (5' to 3') is at the bottom. The consensus score in each plot at each position was computed using the entropy method.*

Mutational analysis of inter-RNA interaction sites

The proposed existence of conserved inter-RNA interactions in influenza A genomes poses an interesting question: are inter-RNA interaction sites more conserved relative to the rest of the genome? In the section Sequence conservation and inter-RNA interactions, very limited evidence (and counterevidence) was found of some form of increased sequence consensus in regions corresponding to interaction hotspots or ‘peaks’, but most regions of consensus variation seemed unrelated to the distribution of inter-RNA interactions. A simpler analysis is possible by taking a reference strain such as WSN, which sourced the inter-RNA interactions in the analyses so far, and comparing its genome to the extrapolation strains. A comparison is then possible by calculating the average mutation rate (the sum of the substitution-, deletion- and insertion rates) over the whole genome versus within interaction sites, for several strains. For the H1N1 vaccine strain set with WSN as the reference, this is shown in Figure 29. From this figure, it is apparent that as expected, not all strains are equally similar to WSN. Especially the 2009-pandemic A/California/07/2009 and post-pandemic A/Michigan/45/2015 strains are dissimilar: somewhat surprisingly this dissimilarity is most notable in the HA and NA segments, even though these strains are still classified as H1N1.

Looking at the plot showing the same calculation restricted only to the Dadonaite et al. WSN interaction sites, it is apparent that there may indeed be some level of increased conservation in the WSN-sourced interaction sites, but not by a large margin. The A/Michigan/45/2015 MP segment is an example, with a whole-genome conservation rate of 0.85 versus 0.90±0.01 within interaction sites (one standard deviation, *t-test* statistic ~69.0, two-sided $p < 0.00001$, sample size 128 WSN MP interactions). Care must be taken in interpreting these values however as the differences are not very large, and they may be subject to the multiple testing problem. Overall though, several such instances for which the interaction site conservation rate is higher can be found. Another problem is the existence of sequencing issues in segment end regions, which may influence especially the whole-segment calculation. This was corrected for by excluding such regions from the calculation. Lastly, the same interaction site (or overlapping sites) may be included multiple times in the calculation if it is involved in multiple discrete interactions.

A more in-depth study into this phenomenon could be undertaken by further restricting the calculation to predicted interaction nucleotides only, and by taking more- and larger sets of genomes.

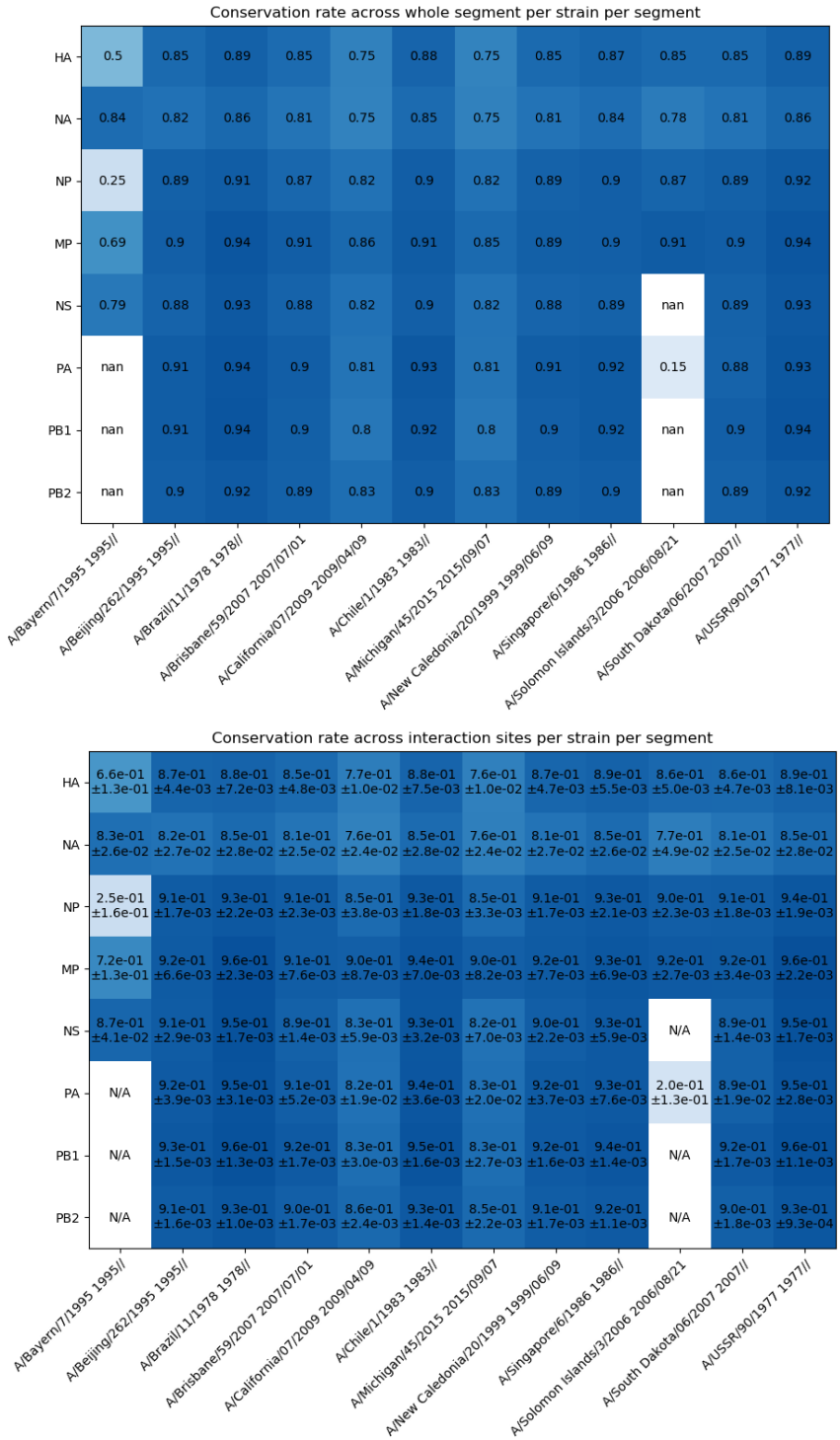


Figure 29: Conservation rates (one minus the mutation rate) for WSN as the reference strain with H1N1 vaccine strains as the variant set. ***Top:*** *conservation rates across whole segments.* ***Bottom:*** *conservation rates inside WSN Dadonaite et al. interaction sites within segments.* ***N/A*** *and extremely low values are usually indicative of absent segments and sequencing errors respectively, cf. Box 1.* ***Interpretation:*** *strains such as Brazil/1978, Chile/1983 and USSR/1997 are more similar to WSN than others. Especially the 2009-pandemic and post-pandemic California/2009 and Michigan/2015 strains are quite dissimilar, most notably in the NA and HA segments. Across most strains and most segments, the conservation rate within the Dadonaite et al. interaction sites may be slightly higher than the whole-segment conservation rate, the Michigan/2015 MP segment being an example (0.85 vs. 0.90±0.01 (1σ), p<0.00001), but the differences are not always significant, and family-wise errors should be taken into account.*

Structural imposition

The structural imposition experiment aims to provide another method of assessing the similarity of two strains for a given inter-RNA interaction. By substituting nucleotides (for both segments) from one strain (the sequence donor strain) into a structure predicted for another (the structure donor strain), removing non-complementary base pairs and introducing new base pairings between directly opposing complementary nucleotides, and finally using ViennaRNA *RNAeval* to determine the predicted free energy of the corresponding structure, the following question is answered: how well do the nucleotides from the sequence donor strain fit into the predicted inter-RNA structure for the structure donor strain?

In Figure 30, a heatmap of free energy values for structural impositions of the H1N1 vaccine strains (and the WSN reference strain) is shown for the *2 MP NS* interaction. Note that the self-imposed structures, where the sequence donor strain is the structure donor strain, are also evaluated using *RNAeval*. Figure 31 can be used as a reference figure, showing a heatmap of the mutation rates within the donor structure interaction sites between the sequence donor strain and the structure donor strain. The pattern of sequence donors with the WSN predicted interaction structure is similar to the sequence conservation pattern for the MP and NS segments in Figure 29. Overall, the predicted imposition FE values follow the mutation patterns in Figure 31 quite closely. In case of structure-preserving or structure-strengthening mutations between two strains, an unexpected deviation from this mutation patterns would be expected, but such a phenomenon does not jump out at first glance. A weaker example of this is that the structures of A/Brisbane/59/2007 and A/South Dakota/06/2007 imposed on the WSN sequence result in highly stable structures even though a few mutations occurred in these strains' interaction sites relative to WSN.

One pattern is most apparent: the low similarity of the 2009-pandemic A/California/07/2009 and post-pandemic A/Michigan/45/2015 strains with virtually all other strains, both when acting as the sequence donor and as the structure donor. However, the predicted self-imposed structures for these strains are not very stable either. The imposition of non-2009-pandemic-associated strains on A/California/07/2009 seems to result in the least stable structures, especially the A/Beijing/262/1995 structure. Somewhat surprisingly, the imposition of A/California/07/2009 structure on A/Michigan/45/2015 and vice versa does not yield better results than with other strains, even though they are highly similar in terms of mutations inside the interaction sites in Figure 31. Interestingly, WSN nucleotides are relatively stable in the A/Michigan/45/2015 structure, more so than self-imposed, possibly due to the generation of new base pairs using WSN nucleotides. This may indicate that some element of the WSN structure may persist in the A/Michigan/45/2015 predicted structure, but that nucleotide bonds have been lost due to mutations.

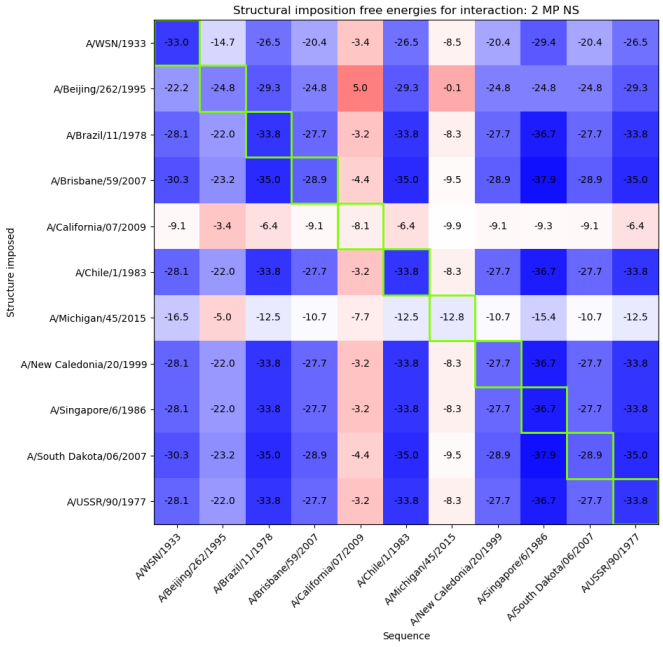


*Figure 30: Structural imposition for the 2 MP NS interaction with H1N1 vaccine strains, with WSN as the reference strain. The sequence donor strain is on the **x-axis**, the structure donor strain is on the **y-axis**. The free energy value was predicted for each single-molecule fused interaction structure with internal open-loop 'GGGGG' nucleotides insert using RNAeval. Therefore, the self-impositions along the diagonal are not necessarily equal in predicted FE to the intaRNA predictions. **Interpretation:** a clear pattern of low similarity between 2009-pandemic and post-pandemic strains with other strains emerges, both for sequence donation as well as for structure donation.*

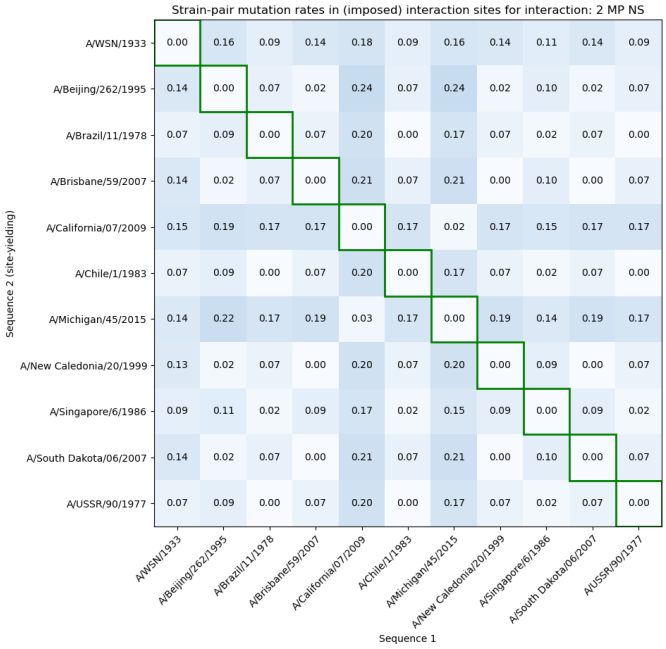


*Figure 31: Structural imposition experiment mutation rate heatmap for the 2 MP NS interaction for H1N1 vaccine strains, with WSN as the reference strain. The sequence donor strain is on the **x-axis**, the structure donor strain is on the **y-axis**. Shown here are the average mutation rates (sum of substitution-, deletion- and insertion rates) between sequence donor strain and structure donor strain within the predicted interacting sub-sequences of the donor structures only, averaged over both segments. **Interpretation:** the same pattern of higher dissimilarity between 2009-pandemic associated strains with other strains is visible here. One notable feature is that the low mutation rate between California/2009 and Michigan/2015 does not result in higher free energy values for cross-imposed structures between the two, given that the Michigan/2015 structure has a better predicted MFE value.*

Cross-strain interaction

The cross-strain interaction experiment evaluates the compatibility of two strains for a specific inter-RNA interaction, by taking the interaction site on one segment from the first strain, and using *intaRNA* to predict a potential interaction with the interaction site on the other segment from the second strain. For each pair of strains, two such combinations can be made.

For the H1N1 vaccine strains including WSN, the resulting predicted MFE values are shown in Figure 32 for the *2 MP NS* interaction. Additionally, Figure 33 shows the cross-strain interaction heatmap for the mixed serotype strain set (see Box 3), again including the reference strain WSN. In the Inter-RNA structures section, it was determined that mutations in the NS segment interaction site are most likely responsible for the loss of structure stability of this interaction in 2009-pandemic associated strains, i.e. A/California/07/2009 and A/Michigan/45/2015. Figure 32 provides further confirmation of this hypothesis by showing that stable structures are possible using the MP segment interaction sites from these strains with non-2009-pandemic-associated NS segment interaction sites, but that the reverse leads to much more unstable structures. Also interesting is the observation that for some combinations of strain crossings, the predicted MFE value is lower than for the non-crossed interactions (along the diagonal), indicating that a cross-strain interaction for these pairs of strains could result in a more stable structure than within the strains themselves. This could have implications for the potential for cross-strain co-segregation of these segments in reassortment scenarios. Examples are: A/California/07/2009 MP with A/Brazil/1/1978, A/Chile/1/1973, A/Singapore/6/1986 or A/USSR/90/1977 NS, and to a lesser extent A/South Dakota/06/2007 NS with A/Chile/1/1983 or WSN MP. Care must be taken in interpreting these results directly in this way on multiple levels: these structure predictions are not necessarily reflective of the *in virio* situation, interaction sites can shift between strains, the MFE value does not say everything about the probability of an interaction occurring, and the integrated effect of other interactions must be considered as well. Nevertheless, some results are strong enough to warrant further investigation, especially the incompatibility of the 2009-pandemic associated NS interaction site with non-2009-pandemic associated strains.

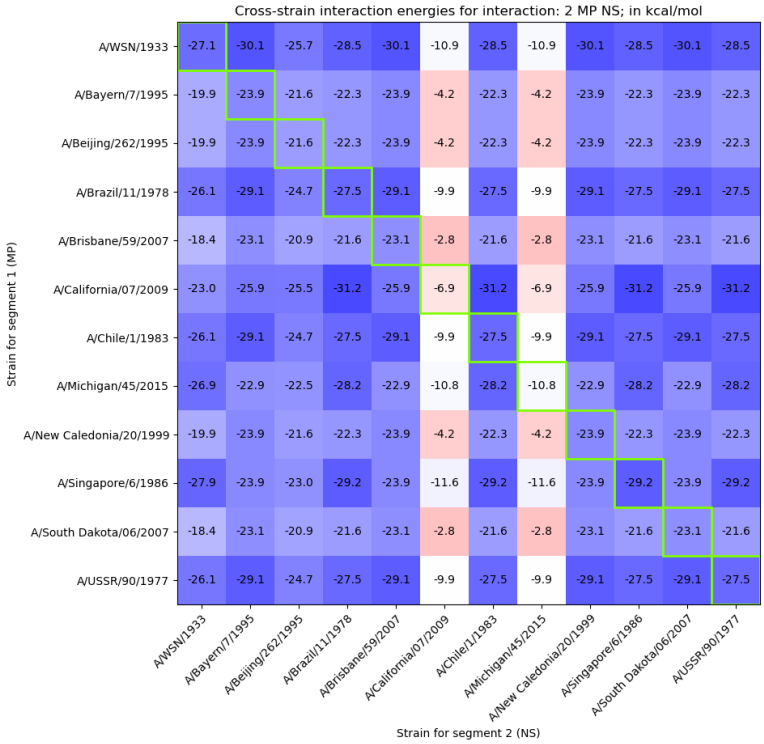


*Figure 32: Cross-strain interaction MFE heatmap for the 2 MP NS interaction for the H1N1 vaccine strains and WSN. MP segment donor strain on the **x-axis**, NS segment donor strain on the **y-axis**. In-strain interactions are shown along the diagonal marked in **green**. **Interpretation:** the pattern here clearly shows that the 2009-pandemic-associated California/2009 and Michigan/2015 strains' NS segment is incompatible for this interaction with MP segments from non-pandemic-associated strains, and the in-strain values indicate incompatibility as well. The MP segment of these 2009-pandemic-associated strains however is compatible with the NS segment of non-2009-pandemic associated strains. Some cross-strain interactions are predicted to be more stable than the interactions within the 'parent' strains.*

In Figure 33, the same heatmap of cross-strain interactions is shown for *2 MP NS* interaction for the mixed serotype strain set, including WSN. These strains are much more dissimilar between themselves compared to the H1N1-only vaccine strains set. This results in increased spread of predicted MFE values for both the in-strain- as well as cross-strain interactions. More caution should be taken in interpreting these results as there is even less guarantee that WSN *in virio* interaction sites are a genuine reflection of the interaction network in these distant influenza strains. Nevertheless, some interesting patterns are visible. The 1968-pandemic associated H3N2 strain A/Hong Kong/01/1968 has a highly stable predicted structure for this interaction, even more so than WSN. The *2 MP NS* interaction was also found *in virio* in the A/Udorn/1972[H3N2] (Udorn) strain, which is also temporally associated with the Hong Kong pandemic strain, by Dadonaite et al. For both possible cross-strain interactions with WSN, a highly stable interaction is predicted, indicating compatibility between these strains even though they are of different serotypes. Since only the MP and NS segments are directly involved in this interaction, this is not as unusual as it seems at first glance. It instead highlights the restrictive nature of the *HxNy* serotype notation for describing influenza A genomes, and suggests that more focus may be needed in analysing reassortment processes not just involving HA and NA segments, but also involving other segments. In the same vein, A/Puerto Rico/8/1934-Korea/426/1968[H2N2] and A/Brisbane/59/2007[H1N1] (which is also in

the H1N1 vaccine strains set) are also predicted to have stable in-strain interactions, and stable cross-strain interactions with WSN and A/Hong Kong/01/1968. Also note the 2009-pandemic-associated A/Belgium/145-MA/2009[H1N1], which, like the A/California/07/2009 and A/Michigan/45/2015 strains, is associated with the 2009 H1N1 pandemic. The pattern for this strain is the same as in Figure 32: its in-strain predicted MFE is high, and its NS segment is predicted to interact poorly with WSN and other H1N1, H2N2 and H3N2 MP segments within the *2 MP NS* interaction sites. However, like the other 2009-pandemic-associated strains, its MP segment does seem to be compatible with these H1N1, H2N2 and H3N2 strains. Such patterns are especially interesting because the potential for a more stable cross-strain interaction could indicate a cross-strain co-segregation preference in reassortment scenarios, if this pattern holds for other cross-strain interactions between the two segments in question.

The more serotypically varied strains included in this analysis are: A/Anhui/1-BALF_RG1/2013[H7N9], A/duck/Hokkaido/Vac-3/2007[H5N1], A/duck/Zhejiang/6DK19-MA/2013[H5N2] and A/mallard/Alberta/70/2017[H7N3], of which only the first was isolated from a human, having been the cause of a significant outbreak of novel influenza in 2013 (Watanabe et al. 2013). Each of these strains except the last have relatively high predicted MFE values for the *2 MP NS* interaction. A straightforward explanation is that A/mallard/Alberta/70/2017[H7N3] has relatively few mutations in both the MP and NS sites with respect to WSN. This seems to be true on average over the whole lengths of the MP and NS segments, as shown in Figure 34, possibly pointing to a WSN-proximal origin of the MP and NS segments in this strain. Notably, these segments are more conserved in all strains than the HA and NA segments (and to a lesser extent the PB1 segment). As a result, interactions between these segments may remain more stable throughout the evolution of influenza strains than interactions between highly variable segments. However, even relatively few mutations in interaction sites are predicted to be able to lead to large variations in interaction stability. This is the case for example for the 2009-pandemic-associated strains for the *2 MP NS* interaction, which have only four extra mutations in the NS interaction site compared to e.g. the stable A/Brazil/11/1978 strain. In general, the MP and NS segment conservation rates w.r.t. WSN correlate well with the predicted MFE values of the *2 MP NS* interaction in the extrapolation strains, with lower conservation rates generally resulting in decreased predicted MFE. In the case of A/duck/Zhejiang/6DK19-MA/2013[H5N2], a 15-nucleotide deletion in the NS interaction site results in a radically changed, weaker interaction structure, and incompatibility of this NS interaction site with other strains in cross-interaction. Oddly enough, the MP interaction site of that strain is quite compatible with NS segments of all other strains (except itself).

Here, it is important again to stress the shifting nature of inter-RNA interactions in influenza A genomes, meaning that it is quite possible that any important functions of this interaction have been taken over by other interactions, or that the interaction sites have gradually shifted away from the homologous genomic location in WSN. This is assuming this interaction even existed in an ancestor strain in the first place, which is not guaranteed especially for the strains with segments more distantly related to WSN.

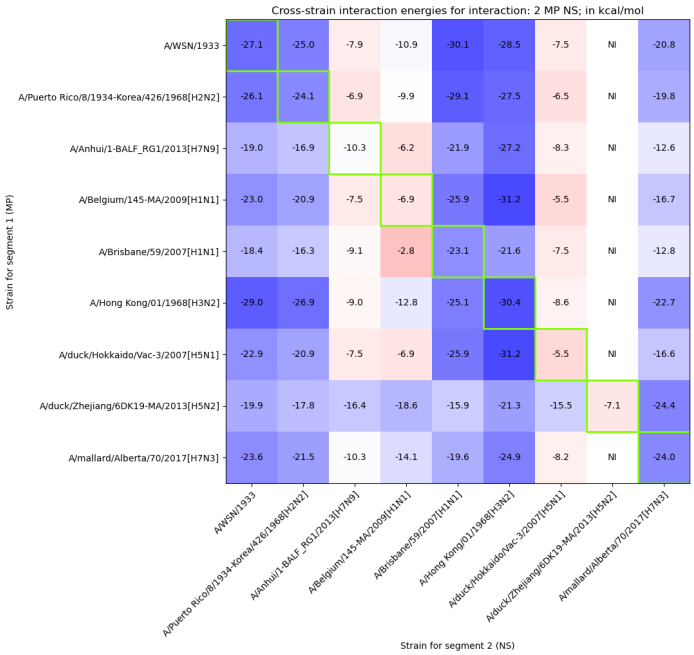


*Figure 33: Cross-strain interaction MFE heatmap for the 2 MP NS interaction for the mixed serotype strain set and WSN. **NI** means no predicted stable interaction (i.e. no interaction structure below 0.00 kcal/mol). **Interpretation:** in-strain stable predicted structures exist for this interaction for WSN, the Puerto Rico…[H2N2] strain, the Brisbane/2007[H1N1] strain and Hong Kong/1968[H3N2]. These strains are also predicted to form stable cross-strain interactions with each other, for all possible segment combinations. Other strains generally have lower predicted in-strain MFE values, including the 2009-pandemic associated Belgium/2009 strain, whose MP segment does form a stable cross-interaction with WSN and the aforementioned stable strains. Other surprising results are seen as well, such as the stability of the Alberta/H7N3 predicted in-strain- and cross-strain interaction structures.*

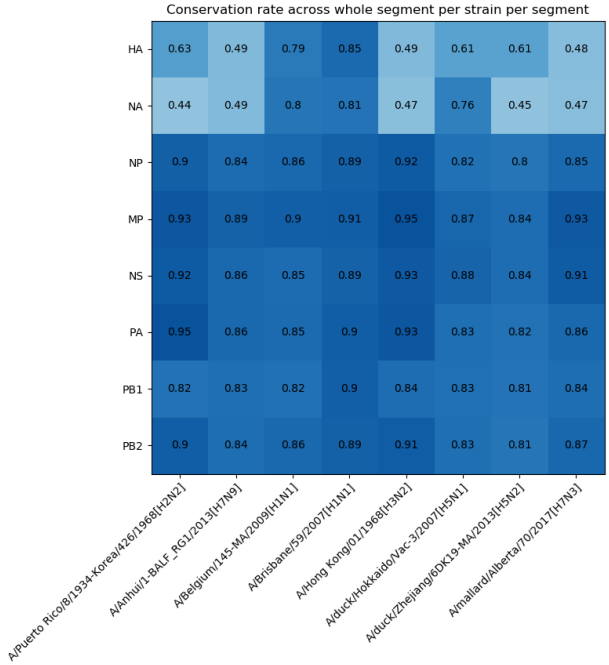


*Figure 34: Whole-segment average conservation rates (one minus the mutation rate) for the mixed serotype strain set with WSN reference strain. **Interpretation:** the H1N1, H2N2 and H3N2 are closest to WSN in terms of conservation for most segments. As expected, the strains not of serotype H1N1 carry dissimilar HA and NA segments, including H3N2 and H2N2. The Alberta/2017 H7N3 strain non-antigenic segments are surprisingly similar to WSN, especially in the MP and NS segments, perhaps pointing to a shared origin or ancestral relationship.*

Although such caveats remain, it is worthwhile to note some significant results. Especially interesting is the predicted enhanced stability of the *2 MP NS* interaction in the A/Hong Kong/1/1968[H3N2] strain, of which the MP and NS segments are highly similar to those of WSN. Figure 35 shows a closer analysis of the predicted structures of this interaction in WSN, A/Hong Kong/1/1968[H3N2] and A/mallard/Alberta/70/2017[H7N3] and the mutations that cause variations in these structures, as well as an overview of mutations in the interaction sites in the mixed serotype strain set. The conclusion here is that several mutations commonly occur in these interaction sites across strains of likely different origins, but that these mutations are not predicted to result in obvious gains or conservation of the (stability of the) interaction structure. It seems more likely that these mutations are common due to other evolutionary pressures or were introduced randomly in a shared ancestor (although this is less likely given the variety of the strain set). Given the centrality of the *2 MP NS* interaction in terms of both RPM and predicted MFE in the WSN reference strain, this is a somewhat surprising result. It may be the case that the interaction network is sufficiently plastic to accommodate most random- and non-random mutations through interaction site shifts and the formation of new interactions. Minor variations in the stability of inter-RNA interactions could also simply not affect the prevalence of that interaction by much, although not all mutations are predicted to result in only minor variations.

(A) 2 MP NS interaction structures

```
>A/WSN/1933
AGCUCACU-CGUUCGUCGUCUCCGGUACCUAUAACGAUCA
      || |||:|:| ||||||||||||  |||             -27.14 kcal/mol
------GAUGCAGGUA-CAGAGGCCAUGGUCAUU------

>A/Hong Kong/01/1968[H3N2]
AGCUCAC-UCGUCCGUCGUCUCCGGUACCUCCAACGAUCA
      | ||||||: ||:||||||||||                  -30.38 kcal/mol
------GAAGCAGGU-GCGGAGGCCAUGGUCAUU------

>A/mallard/Alberta/70/2017[H7N3]
AGUUCAC-UCGUCCGUCGUCUCCGAUACCUCCAACGAUCA
      | |||||||| |||||||                       -24.02 kcal/mol
------GAAGCAGGCA-CAGAGGCAAUAGUCAUU------
```

(B) 2 MP NS interaction sites mutation table

```
  A/WSN/1933 2 MP 382-420 NS 605-631
  v = genomic start position; $ = genomic end position
                 $          *                 **       v v                         $|
  A/WSN/1933     AGCUCACUCGUUCGUCGUCUCCGGUACCUAUAACGAUCA&GAUGCAGGUACAGAGGCCAUGGUCAUU|kcal/mol
> A/WSN/1933                                            &                           |-27.14
  A/Puerto Rico..                             CC        &  C                        |-24.05
  A/Anhui/1-BAL..  U     U  C     C  U        UC    U   & GA  U  C  U  A  A  AC     |-10.35
  A/Belgium/145..        U  C     C           CC      U &  A U         U  A  U      |-6.91
  A/Brisbane/59..     G  U        A           CC        &  A   A  G G               |-23.13
> A/Hong Kong/0..           C                 CC        &  A      G G               |-30.38
  A/duck/Hokkai..        U  C     C           CC        &  A  C  C  U  A  A  A      |-5.48
  A/duck/Zhejia..  U        C        U  U     CC G  G U &  A UC ---------------G    |-7.07
> A/mallard/Alb..  U        C           A     CC        &  A     C        A  A      |-24.02
```

Figure 35: Mutation analysis for the 2 MP NS interaction with WSN as the reference strain and the mixed serotype strain set for extrapolation. ***(A):*** *three selected (re-integrated) predicted interaction structures and corresponding predicted MFE values. Mutations in the interaction sites w.r.t. WSN are marked in* ***yellow****. G-U alternative base pairs are shown using the ':' character.* ***(B):*** *mutation table for the mixed serotype strain set w.r.t. WSN in the MP and NS interaction sites. Deletions are shown using the '-' character.* ***Interpretation:*** *mutations in the Hong Kong/1968 and Alberta/2017 interaction sites are predicted to result in some consolidation of alternative base pairs into standard complementary base pairs, and loss of three bonds in the last part of the structure. For Alberta/2017, mutations in the latter half of interaction sites on both segments result in the loss of several bonds in the main uninterrupted binding part, also leading to slightly decreased stability. The mutations occurring in these strains are usually commonly found in other strains as well, see (B). The MP U390C and A391C dual mutation (****blue*** *marker), resulting in the loss of the three lone base pairs on the right side of the structure, is especially universal. The MP U409C (****red*** *marker) mutation resulting in conversion of a U-G base pair to C-G is also common. Overall, most mutations do not seem to have an obvious benefit for interaction stability, including mutations in Hong Kong/1968.*

Covariation

As discussed in Covariation in the Methods section, nucleotide covariations can reveal the presence of evolutionary constraints imposed on a sequence. Although evidence presented so far points toward mutations in interaction sites generally not being associated with conservation or increased stability of inter-RNA interactions, the detection of covariation events in predicted base pairing positions in such interactions may provide another perspective.

The covariation detection pipeline was performed for the *2 MP NS* interaction with the mixed serotype, H3N2 and H5Nx strain sets for interaction extrapolation. No covariation events were found in the H1N1 vaccine strains set for this interaction. The two discrete covariations seen for the *2 MP NS* interaction across these strains are shown in Figure 36. These covariations occur in separate base pairing positions in the WSN uninterrupted binding region of the interaction structure, cf. Figure 26. Both covariations are found in multiple strains, with the red-marked covariation occurring mostly in H5Nx strains and the blue-marked covariation occurring mostly in H3N2 strains. Unfortunately, the serotypic and genomic similarity of these strain groups means that these covariation events likely occurred once in shared ancestor strains and were subsequently inherited into the strains shown here. The lack of multiple separate covariations occurring in the same coupled positions also represents a lack of certainty that these covariation events reflect a genuine evolutionary constraint imposed by inter-RNA interaction conservation, instead of random variation. It is interesting however that two separate covariations occur in the WSN uninterrupted binding region, whereas covariations in interaction structures are observed to be relatively rare even when searching in all four extrapolation strain sets (many interactions do not have any detected covariation events in any extrapolated structure). For example, the *28 MP NP* interaction covered earlier does not have any detected covariation events across any of the strains analysed.

Looking back at the sequence- and structure consensus profiles in Figure 28, it is apparent that the bond involved in the covariation event marked in blue is very poorly conserved with low consensus in H1N1 strains at both the sequence- and structure level, which is further evidence that this covariation event may just be due to random variation. The red-marked covariation bond and its nucleotides especially are a lot more conserved.

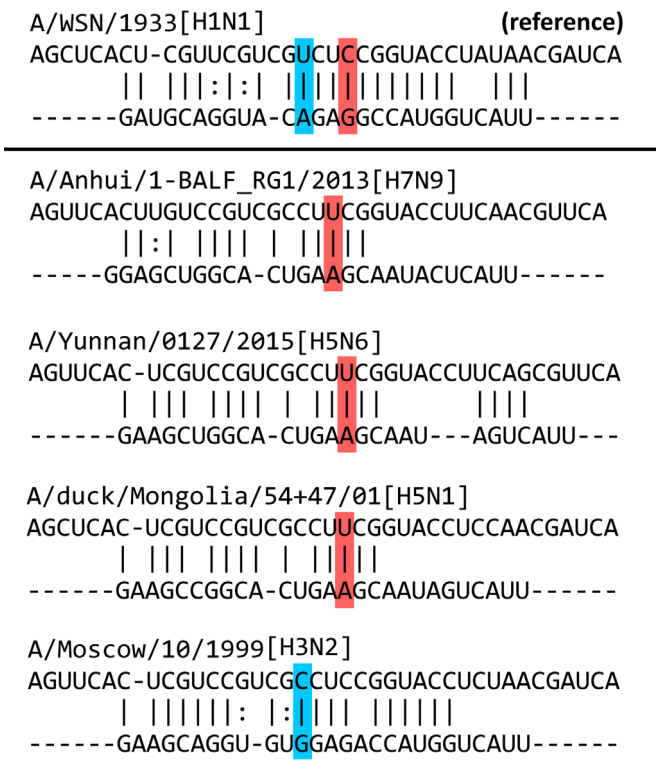


*Figure 36: Interaction structure covariations in the 2 MP NS interaction, with WSN as the reference strain. Two separate covariations are marked in **red** and **blue**. **Interpretation:** each of these covariation events is found in multiple strains, with the covariation event marked in blue being common in H3N2 strains (the H3N2 vaccine strain set), and the event marked in red being common in the H5Nx strain set. The similarity of these strains means that these events likely occurred in shared ancestor strains, and thus that there is little evidence of multiple separate occurrences of these covariation events.*

DISCUSSION

The aim of this thesis is to provide a bioinformatics-based perspective into the 'goldmine' of intersegmental RNA-RNA interaction data provided by the (Dadonaite et al. 2019) and (Le Sage et al. 2020) studies. A pipeline was built to extrapolate known inter-RNA interactions in reference strains to other strains and predict the structures of these interactions. Several experiments were developed to gain a better understanding of these interactions and of the sequence- and structure level similarities and differences in these interactions between strains of influenza A.

In the Known inter-RNA interactions in influenza A section, the Dadonaite et al. and Le Sage et al. data on inter-RNA interactions was discussed in detail. The data mainly concerned interactions in the A/WSN/1933[H1N1] (WSN), A/Puerto Rico/8/1934[H1N1] (PR8) and A/Udorn/1972[H3N2] (Udorn) strains. It was concluded that most genomic segments in each of these strains engage in multiple interactions with other segments, usually with multiple discrete interactions between pairs of segments. Some segments engage in more interactions than others, both in absolute numbers and when measured relative to segment length. However, it does not seem to be the case that one segment is central to the interaction network as a mediator of inter-RNA interactions. The overall interaction networks looked very different when comparing WSN to Udorn, underlining the effects of genomic variation on inter-RNA interactions.

A '7+1'-organisation of viral ribonucleoproteins (vRNPs) in influenza virions has been observed in several studies, in which one vRNP in the centre is surrounded by the other seven (Hutchinson et al. 2010). It is possible that this central vRNP is not static, i.e. that different segments can fulfil this central vRNP role, and/or that this organisation is simply due to spatial packing constraints within the viral particles. The evidence does not completely rule out static segment (vRNP) positioning in the 7+1-organisation in a given strain, since there are many unknowns about influenza RNA interaction networks. However, the presence of multiple interactions between most segments in each strain analysed by Dadonaite et al. and Le Sage et al. does seem to refute the idea of a static vRNP organisation, as this seems difficult to achieve if some pairs of vRNPs do not come into close contact, unless the interaction networks are dynamic at different stages of replication and/or in different viral particles.

By analysing the distributions of discrete inter-RNA interactions over the genome segments in the various strains for which data was available, very interesting observations were made. The distribution of interactions is not uniform for any strain or for any segment. Instead, there exist regions that are involved in many discrete inter-RNA interactions (*hotspots* or *peaks*), alternating with regions engaging in few to none (*silent regions*). The patterns of hotspots and silent regions were similar for the closely related WSN and PR8 strains, but not when compared to Udorn. There were also significant differences between the distributions for WSN in Dadonaite et al. (derived using the *SPLASH* method) and in Le Sage et al. (derived using the *2CIMPL* method) due to as of yet unknown reasons, as also discussed in (Le Sage et al. 2020, p. 6).

Contrary to expectations from research on the selective packaging model of influenza genome packaging, cf. e.g. (Hutchinson et al. 2010), hotspots in inter-RNA interaction profiles are not just concentrated around segment ends, where packaging signals are known to reside. Instead, strong hotspots are frequently located in the middle of coding regions. One of these coding-region hotspots, the hotspot located around 680-720 nt 5' to 3' in the NP segment of WSN, PR8 and Udorn, was analysed in depth in (Le Sage et al. 2020). They showed that synonymous mutations in this hotspot region result in rearrangements in the overall interaction network in WSN, indicating the importance of this hotspot in the interaction network. The lack of disruption of genome packaging in this mutated strain was unexpected, and probably indicates a form of *robustness* of the interaction network to mutations even in important interaction sites. Besides the NP hotspot, other hotspots were found by re-analysing data from mainly the Dadonaite et al. dataset. Particularly strong peaks were found in the PA ~1400 nt and PB1 ~2400 nt regions for WSN, PR8 and perhaps even Udorn. Segments such as MP in WSN had an oscillating pattern of peaks interspersed with silent regions, whilst other segments displayed more 'disordered' profiles.

Additionally, a basic genome-wide sequence consensus study was performed, inspired by (Gog et al. 2007), to determine whether genomic regions with high involvement in inter-RNA interactions are more conserved at the sequence level. With WSN as the reference strain, several H1N1 vaccine strains were selected for this consensus study, cf. Extrapolation strains in the Methods section. It was inferred from overlaying the inter-RNA interaction distribution with the sequence consensus profile that there is little evidence of significantly increased conservation of hotspot regions, including for the strong NP and PA hotspots. For some peak regions, such as the peak around PB1 ~2400 nt, there appeared to be some form of local increase in sequence consensus, but other regions contradicted these findings. More research is needed in order to confirm these findings, but it may be the case that inter-RNA interactions do not necessarily impose strong evolutionary constraints at the sequence level. A more in-depth study should consider the existence of other constraints on sequence evolution, such as those imposed by *intra*-RNA structure and protein conservation.

So far, only results based directly on the Dadonaite et al. and Le Sage et al. data have been discussed. The more 'experimental' results of this thesis concern the extrapolation of this data to new strains in order to support a series of experiments analysing the inter-RNA interactions in depth in a wider context. The bulk of extrapolation experiments were performed with Dadonaite et al. WSN as the reference strain providing the data on inter-RNA interaction locations, which were extrapolated mainly to the set of selected H1N1 vaccine strains using multiple sequence alignment. The Freiburg RNA Tools *intaRNA* algorithm was then used for inter-RNA structure- and minimum free energy (MFE) prediction. For WSN, the predicted MFE values were broadly and roughly normally distributed. It was noted that the predicted MFE values correlate only weakly to the reads-per-million (RPM) values of inter-RNA interactions in the Dadonaite et al. WSN dataset. This most likely indicates that the predicted MFE value does not tell us everything about the occurrence and stability of an interaction *in virio*. A global or interaction-specific comparison of MFE values could still provide useful insights. It was found that the MFE distributions of two extrapolation strains, A/California/07/2009 and A/USSR/90/1977, tended towards less stable interactions. This skew was especially notable for the

more dissimilar 2009-pandemic-associated A/California/07/2009 strain. The overall positive-ward shift of these distributions in extrapolation strains is another possible indication of a lack of conservation of inter-RNA interactions across strains of influenza.

Twenty interactions were selected for further analysis out of the set of 611 found by Dadonaite et al. for WSN. These interactions were the ones that were found to be conserved across WSN, PR8 and Udorn in the Dadonaite et al. study, hopefully reflecting some form of cross-strain conservation of these interactions, even though their RPM and predicted MFE values vary greatly. Of these, two interactions in particular were analysed in depth: *28 NP 670-712 MP 355-395* and *2 MP 382-420 NS 605-631*. The *28 NP MP* interaction was ranked relatively high in RPM and predicted MFE for WSN, overlaps the NP hotspot, and is predicted to be more stable in terms of MFE in most extrapolation strains. The *2 MP NS* interaction is interesting because it was ranked third highest in terms of RPM in WSN, and because its predicted MFE in WSN and other strains indicated that it should form one of the most stable interaction structures. Visual inspection of the predicted interaction structures indicated that some conserved structural elements may be present, including a region of relatively uninterrupted binding in the *2 MP NS* interaction. However, a degree of variation was also observed in the predicted structures.

The existence of these conserved structural elements was examined by analysing sequence- and structure consensus at the single nucleotide scale. The results were inconclusive, with some apparent increased consensus at key nucleotide binding positions in the *2 MP NS* binding region, and perhaps also in an open loop in the *28 NP MP* interaction. However, there was no clear overall pattern of increased conservation of inter-RNA-binding nucleotide positions and inter-RNA bonds in these two interactions.

A very interesting sequence pattern in the NP hotspot region involved in e.g. *28 NP MP* was noted however, which was most apparent in WSN: a quadruple tandem repeat of four nucleotides (‘CAGG’) in the 5’ to 3’ direction. This unusual sequence feature was moderately conserved in other strains: usually one or more, but not all, of the repeats were partially mutated, cf. Figure 37. It is possible that this tandem repeat pattern is involved in inter-RNA interactions, for example by decreasing the chance that each ‘CAGG’ repeat in this region is broken in case of mutation. The perfect complementary pattern, ‘GUCC’ 3’ to 5’ (antiparallel), was not as common in interaction sites interacting with this NP hotspot, but it did occur thrice within the (WSN) MP interaction site for *28 NP MP*. Most interactions in the NP hotspot however were predicted to involve bonds with the ‘CAGG’-repeat region, including the *28 NP MP* interaction, as well as the *36 NP 675-717 MP 400-434* and *98 PA 612-652 NP 689-731* interactions. Another hypothesis is that the ‘CAGG’-repeats allow for shifting and ‘slipping’ of interaction structures, e.g. multiple stable interaction structures with this region could be possible. What the effect of this will be *in virio* is hard to hypothesise. Perhaps one of the ‘CAGG’-motifs could act as the initiator of an interaction by means of a ‘kissing-loop’ interaction, in which free nucleotides (in this case, ‘CAGG’ nucleotides) in a stem-loop structure bind to free nucleotides elsewhere, in this case to the opposing interaction site. This initial interaction could then act as a seed for consolidation into the full, stable inter-RNA interaction. It is feasible that a repeating pattern of multiple ‘CAGG’ nucleotides could engage in such an interaction more easily, or more

likely: that it could more easily receive such a ‘kissing’ interaction, perhaps from a ‘GUCC’ 3’ to 5’ complementary stem-loop. The existence of stem-loops in the influenza genome has been discussed before (Gultyaev, Fouchier, and Olsthoorn 2010, p. 17-18).

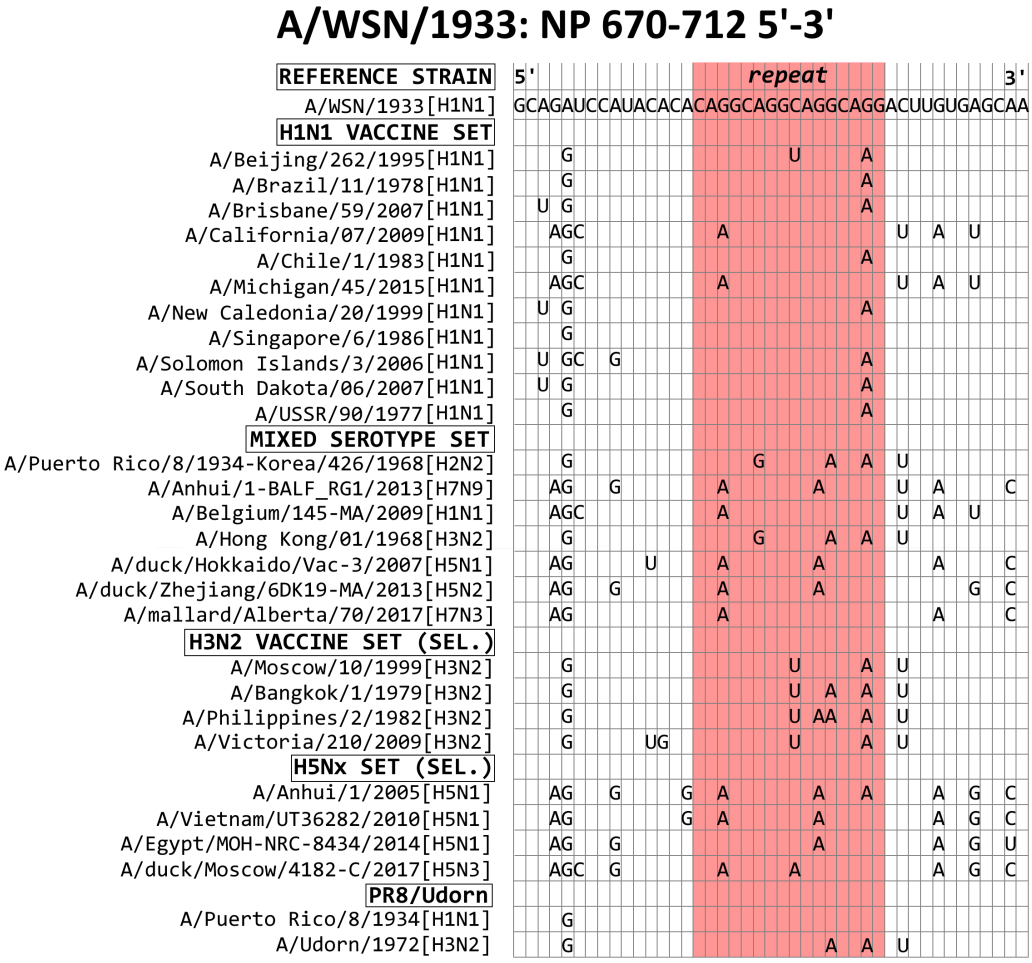


*Figure 37: The NP hotspot (here NP 670-712 5’ to 3’ measured in WSN) and tandem repeat region for various strains of influenza A. The tandem repeat region is shaded **pink**. (Selections from) all strain sets used in the pipeline of this thesis are shown, including the two additional interaction reference strains PR8 and Udorn. **Interpretation:** the tandem repeat region is well conserved in H1N1 strains especially, with no mutations at all in Singapore/1986 and PR8. At least one ‘CAGG’ pattern is conserved in all strains shown. Except perhaps for the ‘G’ to ‘A’ mutation in the last repeat, there is little consensus for mutations in the tandem repeat region. Note that nucleotides outside the tandem repeat may also be involved in NP hotspot interactions. The 2009-pandemic associated strains California/2009, Michigan/2015 and Belgium/2009 have a unique pattern of mutations in/around the NP hotspot with only one mutation inside the tandem repeat.*

Another way of computing possible sequence-level conservation of interaction sites was devised by calculating whole-segment conservation rates (a form of sequence identity, the inverse of mutation rate) for all segments and all selected extrapolation strains, with WSN as the reference genome. These values were then compared to the conservation rates within interaction sites inside each segment. These results were more promising, showing possible increases in sequence conservation inside interaction sites for multiple segment-strain pairs. It was difficult to establish the statistical significance of these findings, and the quality of the sequences and of the alignments can have a major effect on such computations. Especially concerning was the potential for skew of the

conservation rates due to sequencing issues in segment extremities. Manual inspection of the alignments however showed that these issues were of limited effect on the quality of the alignments, and the segment ends were excluded from computations as a precaution. A cursory t-test calculation did yield a promising indication of statistical significance of this finding for at least one segment in one strain.

The structural imposition experiment aimed to investigate the similarity of predicted inter-RNA structures for two strains in a different way. By imposing the predicted interaction structure for one strain onto the interaction site sequences of another and measuring the free energy of that imposed structure, a measure of interaction similarity for two strains is constituted. For the *2 MP NS* interaction, it was found that the 2009-pandemic-associated strain A/California/07/2009 and the post-pandemic strain A/Michigan/45/2015 are not compatible with non-pandemic strains, neither when their structure is imposed nor when their nucleotides are substituted into another structure. Also note that these two strains had relatively unstable predicted structures in the first place. Taken together, this indicates that the *2 MP NS* interaction has likely weakened significantly, or even been lost, in the 2009-pandemic-associated strains due to mutations in key binding positions in the interaction.

In a similar vein, the cross-strain interaction experiment was used to compare pairs of strains on a given interaction. However, the cross-strain interaction experiment was intended to check the compatibility of two strains in a reassortment scenario, by predicting the interaction in case of an exchange of segments between strains. For the *2 MP NS* interaction, the conclusion was that the NS segment of the 2009-pandemic-associated strains drives the weakening of that interaction and the incompatibility with non-pandemic-associated H1N1 strains, whilst its MP segment is predicted to be compatible with all other H1N1 strains tested. The cross-strain interaction analysis was expanded to a mixed serotype strain set for this interaction. The main observation here was the formation of a compatible cluster of H1N1 (non-2009-pandemic) strains, and one strain each of the H3N2, H2N2 and H7N3 serotypes. The unexpected serotypic variety of this cluster is an interesting avenue for future research on the potential for cross-serotype reassortment. A larger scale computational study of cross-strain interaction potentials would be a first step to that end.

Finally, one last attempt was made to investigate conservation in inter-RNA interactions: by looking for nucleotide covariations in inter-RNA binding positions. This yielded no results for the *28 NP MP* interaction, but two discrete covariation events were found in the long binding region of the *2 MP NS* interaction. These events were seen in several strains of H5Nx and H3N2 serotypes respectively. Unfortunately, due to the genetic similarity of these strains and in the absence of good phylogenetic information on the evolution and reassortment of these strains, the conclusion must be that the evidence points to single occurrences of these two discrete covariation events in ancestor strains. Although covariation events in inter-RNA binding positions were observed to be relatively rare in inter-RNA interactions even when analysing relatively large numbers of strains, there is therefore no strong evidence that these covariations are indicative of an evolutionary constraint imposed on conservation of these binding positions. If a covariation event is shown to occur in multiple discrete lineages at the same position, this would constitute more significant evidence of an evolutionary constraint.

Due to space constraints, few results on the H3N2 and H5Nx strain sets were covered in this thesis. The role of H3N2 strains as one of the foremost causes of seasonal- and epidemic influenza, and the concern about potential human outbreaks of highly pathogenic H5Nx strains, warrants further investigation into the patterns of inter-RNA interactions and consequences for reassortment in these strains. Some insights into H3N2 and H5Nx strains were covered mostly for the WSN-based *2 MP NS* and *28 NP MP* interactions. For H3N2, the use of the Dadonaite et al. Udorn interactions datasets is probably more prudent, given the serotypic connection of Udorn to other H3N2 strains. For H5Nx unfortunately no such reference strain/dataset is currently available. Preliminary results indicate that H5Nx strains are most likely more dissimilar to WSN than the H1N1 strains in terms of interaction site sequences and predicted interaction structures, as expected. The defining feature of H5Nx strains is the H5 segment type, which is highly dissimilar to H1, with around ~60% sequence identity/conservation rate in Figure 34. Other segments in H5Nx strains are usually also less similar to WSN when compared to e.g. non-2009-pandemic H1N1 strains. This resulted in more pronounced predicted MFE shift distributions from the WSN reference for H5Nx strains compared to A/USSR/90/1977[H1N1], but not compared to 2009-pandemic-associated A/California/9/2009[H1N1] (comparing Figure 23 to Figure A 2 in Appendix A: H5Nx figures). As with the H1N1 strain set, not all interactions were predicted to have less stable structures in H5Nx strains: for a surprisingly large number of interactions, significant negative shifts were predicted across numerous strains, cf. Table A 1 in Appendix A: H5Nx figures. Lastly, Figure A 1 shows the predicted MFE values for a selection of H5Nx strains for the triple conserved interactions (cf. Table 4 and Figure 24). The pattern of MFE shifts here does not seem similar to the pattern for H1N1 extrapolation strains in Figure 24: virtually all triple-conserved (WSN, PR8, Udorn) interactions are predicted to become less stable in H5Nx strains.

Limitations

The wealth and complexity of the inter-RNA interaction datasets means that there are various approaches to deriving new information from them. The approach chosen for this thesis has its advantages, but it also has drawbacks. The most fundamental limitation lies in the concept of interaction extrapolation: given the apparent volatile nature of inter-RNA interaction networks, there is no guarantee that extrapolation of interaction sites from one reference strain to homologous RNA regions in another yields a real interaction occurring in the extrapolation strain. The interaction sites may have shifted due to mutations, or the interaction may have substantially weakened or disappeared altogether. Another option is that the interaction never existed in ancestor strains of the extrapolation strain, having arisen in a distinct phylogenetic lineage for the reference strain, or that it disappeared due to reassortment with other genetic material. In this case, the entire concept of ‘conservation’ of interactions is not valid, as only convergent evolution or multiple reassortment could result in similar interactions arising in the extrapolation strain. In an attempt to alleviate this, I focused mostly on triple conserved interactions (in WSN, PR8 and Udorn) from the Dadonaite et al. dataset for the extrapolation analyses. That should mean that interactions in strains with segments derived from the lineages of these three reference strains should be valid targets for conservation analysis, but it is difficult to determine when this is the case, as high mutation rates and reassortment make

influenza phylogeny complex. Still, in cases where the concept of interaction conservation is not sensible, *apparent* conservation or lack thereof may still provide useful perspectives into interaction network plasticity, reassortment or even convergent evolution. A more phylogenetically informed study may prove useful however in determining the scope of this plasticity.

A related issue is that the extrapolation method can never find new interactions (unless they perfectly overlap a known interaction) in the extrapolation strains. This likely resulted in unrealistic free energy distributions showing an apparent overall weakening of interactions in most extrapolation strains with respect to the reference strain.

In order to make a solid comparison of inter-RNA interactions between strains, it would be prudent to analyse a large set of interactions between various segments. However, only 20 triple-conserved inter-RNA interactions were available from the Dadonaite et al. dataset, and of these, only few were analysed in any depth due to space- and time constraints. These individual interaction analyses probably do not tell us the full story of interaction conservation and cross-strain potentials. An extended analysis including more interactions and perhaps a more statistics-driven interpretation could provide additional perspectives and help make the conclusions more concrete.

The *intaRNA* algorithm is a powerful and modern tool for inter-RNA structure prediction, but it can also confer drawbacks in the context of this thesis. Firstly, it only yields the minimum free energy predicted structure, but this may not always be the most likely structure to occur *in virio* due to environmental factors, the effect of *intra*-RNA structure (no accessibility constraints were available), protein binding, etc. It is also possible that multiple structures occur in the real situation, i.e. some sort of thermodynamic ensemble of interaction structures is more realistic. Another issue is that the predicted MFE value does not tell us everything about interaction stability or the probability of that interaction occurring *in virio*, as partially supported by the weak correlation of interaction RPM with predicted MFE, cf. Figure 22. Moreover, the *intaRNA* algorithm can only predict structures that are two-dimensional and *linear* in some sense, i.e. three-dimensional RNA structure is neglected and more complicated pseudoknot-like inter-RNA structures will not be found. A three-dimensional and non-linear structure prediction algorithm would likely carry much heavier computational costs, especially in the pipeline of this thesis, where the procedure would have to be called hundreds of times. Another concern is the fundamental assumption that inter-RNA interaction only occurs in the antiparallel direction, i.e. 5’ to 3’ for one strand, 3’ to 5’ for the other. This assumption may not be correct in all cases, as parallel RNA duplexes are thought to be possible (Szabat and Kierzek 2017). A cursory look for a few low MFE interactions did not reveal more stable interaction in the parallel direction, but a deeper study would be necessary to confirm this. To verify the inter-RNA structures predicted by algorithmic approaches, a nucleotide-resolution inter-RNA structure elucidation technique would have to be developed. Perhaps developments in the (single molecule) nucleic acid structure determination field could aid such an effort (Weeks 2010). The *SHAPE-MaP* technique used to elucidate the intra-RNA structure of influenza genome segments and derive accessibility constraints for inter-RNA structure prediction by (Dadonaite et al. 2019) could conceivably also help in determining the RNA structure of crosslinked inter-RNA interaction fragments.

Outlook

RESOLVE THE ROLE OF INTER-RNA INTERACTIONS IN INFLUENZA GENOME PACKAGING

Although evidence abounds in favour of the occurrence of selective packaging in influenza, the precise mechanism has remained a mystery. Evidence for the existence of so-called packaging signals mainly near segment extremities has been found, but the role of these regions is unclear. Literature on influenza genome packaging has noted the potential involvement of direct inter-RNA interactions in maintaining selective packaging, cf. e.g. (Hutchinson et al. 2010), (Gerber et al. 2014) and (Shafiuddin and Boon 2019), the last also discussing the seminal findings of the (Dadonaite et al. 2019) study. Although these inter-RNA interactions are noted to be important in maintaining selective packaging e.g. in (Gavazzi, Yver, et al. 2013), (Gavazzi, Isel, et al. 2013) and (Le Sage et al. 2020), it is not clear how they are involved. At what stage of replication do inter-RNA interactions arise? Is the interaction network dynamic, with interactions arising and disappearing at different time points? How does the interaction network itself arise, does any segment play a bigger role in this regard than others? What is the interplay between vRNA association into vRNP complexes and inter-RNA interaction, is inter-RNA interaction hindered or assisted by nucleoprotein binding? A logical hypothesis is that the direct inter-RNA interactions cause vRNPs to stick together, resulting in co-segregation into new virions. Is this a realistic model of the (contribution of inter-RNA interactions to the) mechanism of selective packaging? These are just a few of the questions posed that we may want to see answered in the coming years.

ELUCIDATE INTERACTION NETWORKS IN VIRIO

The *SPLASH* and *2CIMPL* techniques used/developed by (Dadonaite et al. 2019) and (Le Sage et al. 2020) respectively are able to resolve influenza intersegmental RNA-RNA interactions *in virio*. The application of these techniques has so far been limited mostly to a few common strains of influenza. In order to study the conservation and plasticity of inter-RNA interaction networks, a larger-scale study should be performed on strains with diverse genomic- and serotypic backgrounds. Especially interesting is tracking changes in interaction networks due to reassortment events, this could for example be done using artificial reassortant strains combining genetic material from several ancestor strains, as (Dadonaite et al. 2019) already showed. Based on notable similarities in the WSN and PR8 interaction networks, it is also interesting to compare closely related strains. This could perhaps allow us to track the effects of specific mutations on the interaction network, in the same manner as the NP hotspot mutagenesis experiment in (Le Sage et al. 2020). Perhaps the analysis of interaction networks could be expanded to influenza B, which does engage in reassortment (Dudas et al. 2015) even though it is less common in animal reservoirs, or even to other segmented single-stranded RNA viruses. In the end, elucidating the structure and evolution of inter-RNA interaction networks could provide us with a greater understanding of reassortment potentials between strains, which could prove invaluable in predicting dangerous reassortment events.

ALGORITHMIC APPROACH: WHOLE GENOME INTERACTION SCAN

The core contribution of this thesis is to introduce a method to extrapolate known inter-RNA interactions found for one strain to other strains. Influenza genome databases contain thousands of relatively complete influenza genomes, and this approach can easily be extended to other interesting strains. Although several useful experiments in this thesis build upon the data generated using this method, it has its limitations. Besides the fact that *intaRNA*-predicted interaction structures may not be a genuine reflection of the *in virio* interactions, a more fundamental limitation is the fact that inter-RNA interactions are not static. From the results in (Dadonaite et al. 2019) and (Le Sage et al. 2020), we know that the genomic locations of interactions can shift due to genetic variation between strains. One major concern for the method in this thesis is the fact that new interactions may arise in a variant strain, which can never be found using a dataset that only provides interactions derived in some reference strain.

An approach that may alleviate this issue is to perform a whole-genome evaluation of potential inter-RNA interactions using an *intaRNA*-like structure prediction method. This method would take all pairs of segments (28 combinations) given a particular influenza genome and calculate potential interactions between each pair of segments. As has been discussed throughout this thesis, e.g. in Distributions of inter-RNA interactions in Influenza A, most segments interact through multiple discrete interactions with each other segment. This is a problem, because current inter-RNA interaction structure prediction algorithms usually only return the strongest predicted interaction between a pair of sequences, i.e. only one interaction would be found using a naïve approach of querying to *intaRNA* using full sequences for each pair of segments.

Instead, the algorithm should predict all discrete interactions possible between two sequences (two segments) above some free energy threshold, e.g. 0.00 kcal/mol as the absolute maximum. Given two sequences of length n and m, such an approach would involve performing multiple dynamic programming-based (pairwise alignment-like) procedures on a pairing table of size $n \times m$, similar to other inter-RNA interaction finding algorithms such as *intaRNA* or *RNAup*. This would be computationally expensive, as even efficient algorithms like *intaRNA* have trouble scaling up to whole-segment scales, and that is just for one interaction.

To ease the computational complexity, a seed scan step could be implemented first. In this procedure, each sequence is subdivided into its constituent consecutive k-mers of nucleotides. For each k-mer found, *complementary k-mers* can be computed based on its antiparallel perfect nucleotide-complementary k-mer, and then allowing for some number of mismatches and alternative base pairs (i.e. G-U base pairs). Then, *seed matches* can be found by matching complementary k-mers between the two query sequences. This procedure is similar to the *word matching* step used as a starting point in the various versions of BLAST algorithms (Wheeler and Bhagwat 2007, Section 1.3). In analogy to BLAST algorithms, these seed matches can then be used as a starting point for an *intaRNA*-like dynamic programming procedure which expands the interaction region around the seed region in order to find the local minimum free energy interaction. Several inter-RNA interaction algorithms already have a similar seed-finding and -starting step built in, including *intaRNA*. The only

difference is that each seed region found should be expanded separately, i.e. result in discrete interactions with different predicted free energy values. In cases where two discrete interactions are close in the pairing diagram, they could be merged if the resulting interaction is more stable. This merging procedure could become complicated in cases where there is disagreement on which nucleotides bind to which other nucleotides, which is something to keep in mind. The resulting list of interactions and corresponding free energy values then constitute all predicted possible inter-RNA interactions between the two segments. See Figure 38 for a graphical overview of the proposed method.

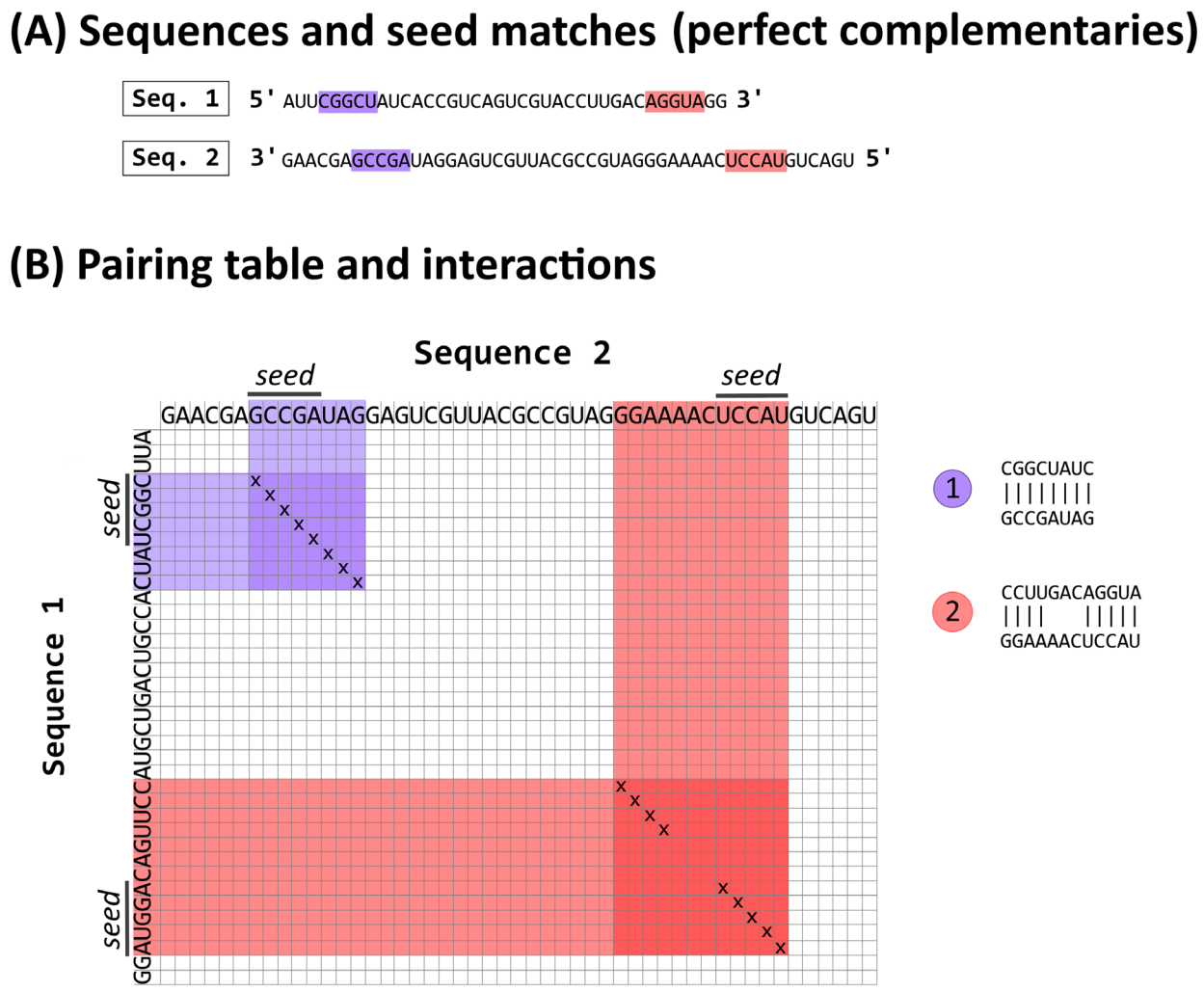


Figure 38: An illustration of the proposed whole-sequence inter-RNA interactions profiling method. ***(A):*** *the seed matching step, showing perfect complementary k-mers of length 5.* ***(B):*** *the pairing diagram for the two sequences with two interactions extended from the seed matches shown. The first interaction (blue) has an uninterrupted binding region of 8 nucleotides, meaning that three length-5 k-mers seed matches would be merged into this single interaction. The second interaction (red) has an internal loop of 3 nucleotides on both sequences.*

This approach itself comes with limitations. By relying only on sequence information, important considerations such as RNA accessibility are left out. A more complicated algorithm could be devised to include such information, as *intaRNA* already does (Busch, Richter, and Backofen 2008). The

most fundamental limitation is that this approach would rely fully on an *in silico* approach of finding interaction sites and corresponding interaction structures. There is no guarantee that a profile of inter-RNA interactions generated by such a pipeline would be a genuine reflection of the interaction network found *in virio* in the particular strain of influenza to analyse. The key to resolving this is by testing the approach on real *in virio* interaction network datasets, such as those provided already by (Dadonaite et al. 2019) and (Le Sage et al. 2020). In this way, an assessment can be made as to whether this approach can produce a realistic profile of the inter-RNA interaction network in a particular influenza genome. The current state of conflicting results between the (Dadonaite et al. 2019) and (Le Sage et al. 2020), cf. (Le Sage et al. 2020, p. 6), may prove to be problematic in this regard, but this situation may change in the future if more data becomes available.

Conclusions

This thesis aimed to investigate intersegmental RNA-RNA interactions (inter-RNA interactions) in the influenza A genome from a bioinformatics perspective. Tools were developed to answer a few key questions, such as: are inter-RNA interactions in influenza A conserved among different strains and serotypes on the sequence- and/or structure level? Are some interactions more or less conserved than others? Based on the results in this thesis, I would say that most inter-RNA interactions are likely not well conserved across strains. This is especially true for genomically more distant strains. Influenza genomics is complex, and serotype certainly does not tell us everything about the interrelatedness of strains due to the potential for ‘silent’ reassortment of other segments, i.e. where no antigenic shift is involved. Some evidence points to the existence of partially conserved interaction-site sequence elements, structural elements, and structure stability for interactions that were studied in depth, even between more distant strains. Crucially, the vast majority of mutations within existing (extrapolated) interaction sites seemed to result in less stable predicted structures; consolidation of interaction structures as a result of mutational processes seemed rare to non-existent.

Such results should be confirmed by extending to new strains the highly innovative *in virio* inter-RNA interaction resolution procedures that were utilised by (Dadonaite et al. 2019) and (Le Sage et al. 2020), whose data forms the foundation of this thesis. This thesis by no means describes all the interesting information and patterns that can be extracted from these datasets, and myriad fundamental questions remain. Is it true that inter-RNA interaction networks are highly plastic, changing quickly even when relatively few mutations occur? How does the interaction network respond to reassortment events? If inter-RNA interactions are important in maintaining selective packaging, how come the interaction network is so volatile? Is there a degree of robustness built in, ensuring that selective packaging is maintained even though inter-RNA interactions and the interaction networks overall may shift frequently and rapidly? By expanding our knowledge of inter-RNA interaction networks, we may elucidate the role they play in influenza genome packaging and pave the way towards solving key mysteries of the influenza reassortment process.

APPENDIX A: H5NX FIGURES

This appendix contains figures concerning H5Nx strains referred to in the Discussion.

Table A 1: Dadonaite et al. WSN dataset interactions for which the predicted MFE values in the H5Nx strain set are on average shifted negatively (more stable) w.r.t. WSN by more than -1.00 kcal/mol. Sorted by segment-pair. Predicted MFE values for WSN and three H5Nx strains also shown. Triple conserved (WSN, PR8, Udorn) interactions in the Dadonaite et al. datasets are marked ***green****.*

Interaction	Mean H5Nx MFE shift	WSN	A/Anhui/1/2005[H5N1]	A/duck/Mongolia/54+47/01[H5N1]	A/swine/Banten/UT2071/2005[H5N1]
150 PB2 1996-2030 PB1 343-401	-2.15	-3.79	-8.65	-9.28	-9.32
294 PB2 1984-2042 PB1 2038-2116	-1.57	-3.49	-4.79	-7.91	-6.35
564 PB2 478-532 PB1 2224-2272	-1.07	-6.75	-12.19	-6.71	-12.92
506 PB2 1556-1606 HA 1398-1454	-1.39	-4.08	-5.07	-5.04	-5.16
303 PB2 185-239 NP 674-716	-1.44	-12.06	-16.72	-17.33	-16.28
359 PB2 1971-2035 NP 1452-1516	-1.11	-4.31	-4.52	-4.37	-4.47
187 PB1 2225-2269 PA 778-812	-2.88	-0.62	-3.38	-4.48	-3.22
258 PB1 917-961 PA 1364-1412	-3.27	-6.20	-10.24	-16.22	-8.56
366 PB1 340-398 PA 203-271	-3.66	-4.19	-7.75	-6.14	-7.11
383 PB1 1883-1925 PA 749-805	-2.25	-5.16	-9.20	-7.97	-7.50
269 PB1 749-799 HA 819-865	-1.45	-7.24	-9.94	-10.16	-9.83
271 PB1 2211-2265 HA 85-135	-1.59	-6.01	-9.38	-7.97	-6.04
572 PB1 1037-1073 HA 103-133	-5.68	-2.76	-8.33	-6.09	-11.52
419 PB1 307-347 NA 285-335	-1.02	-7.20	-9.69	-11.58	-14.45
80 PB1 730-772 NS 527-565	-2.90	-9.21	-13.09	-12.22	-13.09
483 PB1 293-353 NS 227-267	-1.53	-4.46	-6.46	-4.21	-6.74
536 PA 1362-1398 HA 1400-1450	-5.04	-4.61	-7.98	-10.43	-9.61
103 PA 112-164 MP 932-964	-1.63	-15.49	-16.91	-19.37	-17.26
253 PA 1351-1411 MP 373-419	-4.36	-7.52	-10.25	-14.51	-12.97
309 PA 1357-1409 MP 848-898	-1.91	-4.07	-9.19	-7.95	-5.42
311 PA 1353-1403 MP 456-500	-2.19	-5.50	-8.58	-12.21	-7.46
27 PA 1539-1561 NS 207-261	-4.21	-10.23	-16.59	-11.51	-15.72
487 PA 139-187 NS 528-570	-3.24	-5.92	-10.70	-7.97	-7.83
45 NP 1319-1365 MP 941-983	-2.21	-20.69	-25.61	-17.56	-26.43
343 NP 450-500 MP 840-882	-3.59	-3.68	-5.16	-6.16	-9.65
323 NP 1087-1123 NS 504-546	-3.77	-1.81	-4.83	-3.69	-4.83
488 NP 937-979 NS 360-396	-2.04	-7.42	-9.67	-8.75	-10.62
544 NP 681-715 NS 597-627	-3.21	-2.77	-8.18	-7.96	-5.33
609 NP 108-156 NS 730-770	-1.47	-7.15	-9.50	-11.12	-9.63
458 NA 869-919 MP 601-647	-2.03	-5.91	-13.17	-11.68	-6.50

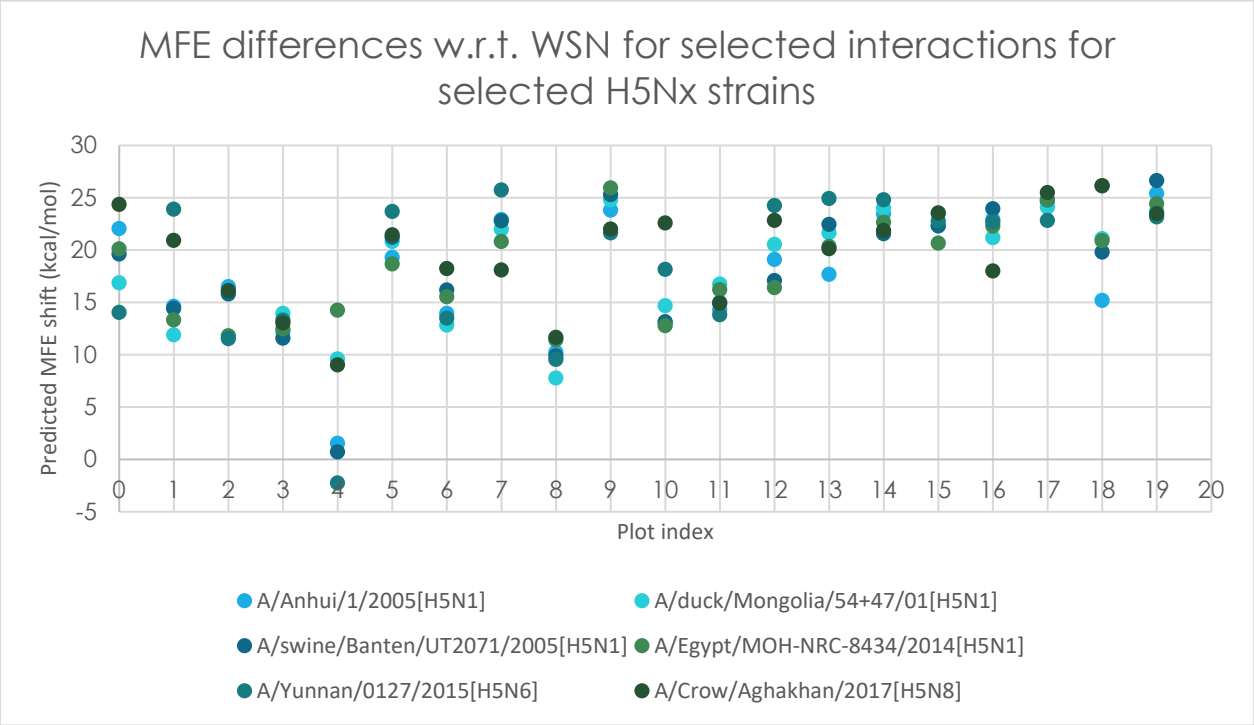


*Figure A 1: Recreation of Figure 24 for the H5Nx strain set: predicted MFE differences w.r.t. WSN for the triple conserved (WSN, PR8, Udorn) interactions in the Dadonaite et al. datasets, for several strains from the H5Nx strain set. Please refer to Table 4 for the interactions corresponding to each **plot index**.*

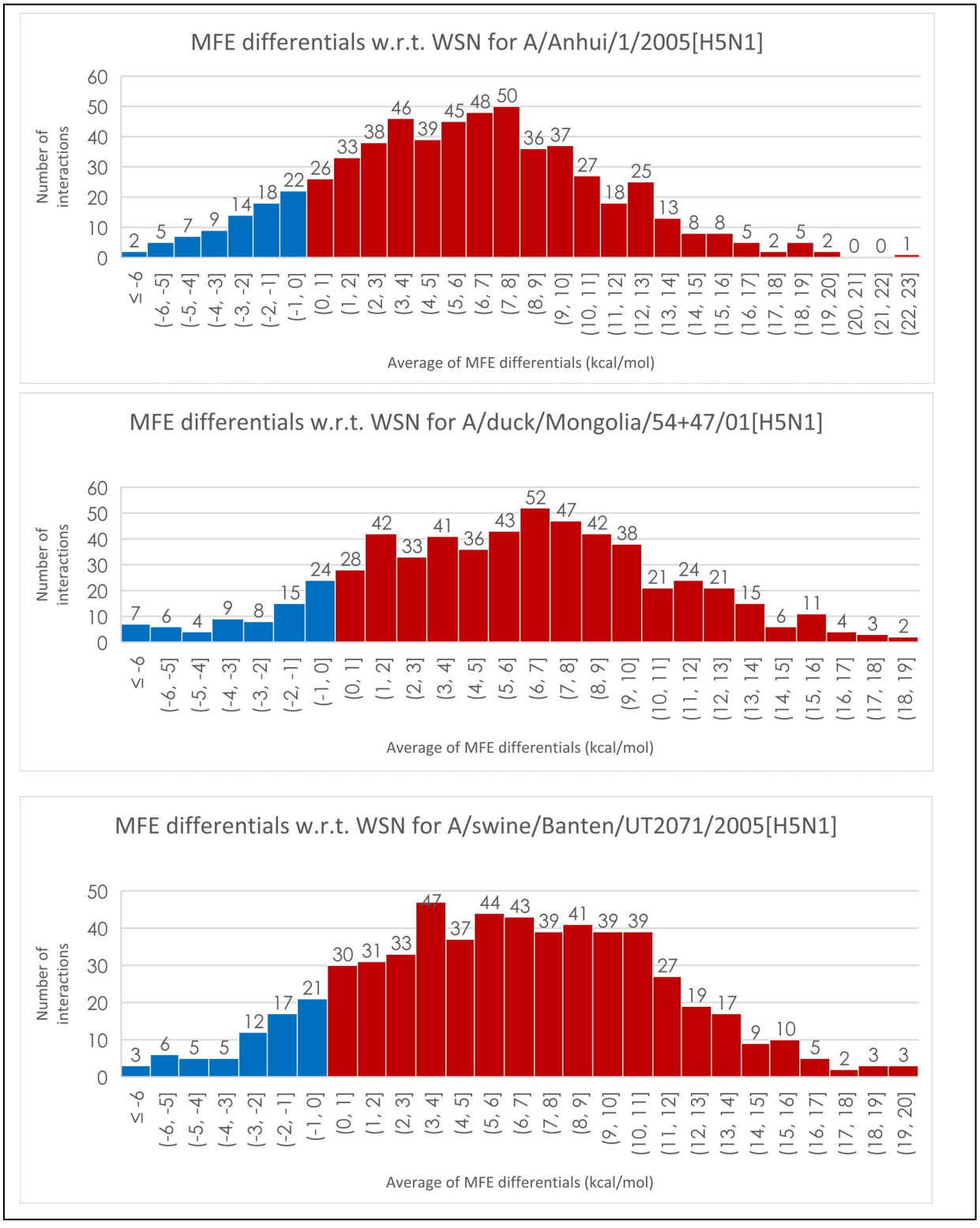


*Figure A 2: Histograms of predicted MFE differences for three H5Nx strains w.r.t. WSN for the full Dadonaite et al. WSN (average of replicates) interactions dataset. Positive shifts in **red**, negative shifts (including zero) in **blue**. The negative shift ratios for the respective strains are: 77 out of 589 significant interactions (13.1%), 71 out of 588 significant interactions (12.1%), 69 out of 587 significant interactions (11.8%). The dataset contains 611 interactions in total. The mean shifts per strain are respectively (±1σ): +5.77±5.05 kcal/mol; +5.65±4.99 kcal/mol; +6.05±5.05 kcal/mol.*

REFERENCES

Chicago Manual of Style, 17th edition (author-date)

Alders, Robyn, Joseph Adongo Awuni, Brigitte Bagnol, Penny Farrell, and Nicolene de Haan. 2014. “Impact of Avian Influenza on Village Poultry Production Globally.” *EcoHealth* 11 (1): 63–72. https://doi.org/10.1007/s10393-013-0867-x.

Aw, Jong Ghut Ashley, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum-Loong Boon, Sidika Tapsin, et al. 2016. “In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation.” *Molecular Cell* 62 (4): 603–17. https://doi.org/10.1016/j.molcel.2016.04.028.

Bao, Yiming, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, and David Lipman. 2008. “The Influenza Virus Resource at the National Center for Biotechnology Information.” *Journal of Virology* 82 (2): 596–601. https://doi.org/10.1128/JVI.02005-07.

Bavagnoli, Laura, Stefano Cucuzza, Giulia Campanini, Francesca Rovida, Stefania Paolucci, Fausto Baldanti, and Giovanni Maga. 2015. “The Novel Influenza A Virus Protein PA-X and Its Naturally Deleted Variant Show Different Enzymatic Properties in Comparison to the Viral Endonuclease PA.” *Nucleic Acids Research* 43 (19): 9405–17. https://doi.org/10.1093/nar/gkv926.

Beigel, John H. 2008. “Concise Definitive Review: Influenza.” *Critical Care Medicine* 36 (9): 2660–66. https://doi.org/10.1097/CCM.0b013e318180b039.

Belshe, Robert B. 2010. “The Need for Quadrivalent Vaccine against Seasonal Influenza.” *Vaccine*, Seasonal Influenza Vaccines: An update for health care professionals, 28 (September): D45–53. https://doi.org/10.1016/j.vaccine.2010.08.028.

Boone, Stephanie A., and Charles P. Gerba. 2007. “Significance of Fomites in the Spread of Respiratory and Enteric Viral Disease.” *Applied and Environmental Microbiology* 73 (6): 1687–96. https://doi.org/10.1128/AEM.02051-06.

Bouvier, Nicole M., and Peter Palese. 2008. “THE BIOLOGY OF INFLUENZA VIRUSES.” *Vaccine* 26 (Suppl 4): D49–53.

Brierley, Ian, Simon Pennell, and Robert J. C. Gilbert. 2007. “Viral RNA Pseudoknots: Versatile Motifs in Gene Expression and Replication.” *Nature Reviews Microbiology* 5 (8): 598–610. https://doi.org/10.1038/nrmicro1704.

Bui, Matthew, Elizabeth G. Wills, Ari Helenius, and Gary R. Whittaker. 2000. “Role of the Influenza Virus M1 Protein in Nuclear Export of Viral Ribonucleoproteins.” *Journal of Virology* 74 (4): 1781–86.

Busch, Anke, Andreas S. Richter, and Rolf Backofen. 2008. “IntaRNA: Efficient Prediction of Bacterial SRNA Targets Incorporating Target Site Accessibility and Seed Regions.” *Bioinformatics (Oxford, England)* 24 (24): 2849–56. https://doi.org/10.1093/bioinformatics/btn544.

CDC. 2020. “Burden of Influenza.” Centers for Disease Control and Prevention. October 5, 2020. https://www.cdc.gov/flu/about/burden/index.html.

Chaitanya, K. V. 2019. “Structure and Organization of Virus Genomes.” *Genome and Genomics*, November, 1–30. https://doi.org/10.1007/978-981-15-0702-1_1.

Chow, Eric J., Melissa A. Rolfes, Alissa O’Halloran, Nisha B. Alden, Evan J. Anderson, Nancy M. Bennett, Laurie Billing, et al. 2020. “Respiratory and Nonrespiratory Diagnoses Associated With Influenza in Hospitalized Adults.” *JAMA Network Open* 3 (3). https://doi.org/10.1001/jamanetworkopen.2020.1323.

Coffin, John M., Stephen H. Hughes, and Harold E. Varmus. 1997. *Viral RNA Packaging*. *Retroviruses*. Cold Spring Harbor Laboratory Press. https://www.ncbi.nlm.nih.gov/books/NBK19418/.

Cowling, Benjamin J., Dennis K. M. Ip, Vicky J. Fang, Piyarat Suntarattiwong, Sonja J. Olsen, Jens Levy, Timothy M. Uyeki, et al. 2013. “Aerosol Transmission Is an Important Mode of Influenza A Virus Spread.” *Nature Communications* 4: 1935. https://doi.org/10.1038/ncomms2922.

Dadonaite, Bernadeta, Brad Gilbertson, Michael L. Knight, Sanja Trifkovic, Steven Rockman, Alain Laederach, Lorena E. Brown, Ervin Fodor, and David L. V. Bauer. 2019. “The Structure of the Influenza A Virus Genome.” *Nature Microbiology* 4 (11): 1781–89. https://doi.org/10.1038/s41564-019-0513-7.

Dobson, Joanna, Richard J. Whitley, Stuart Pocock, and Arnold S. Monto. 2015. “Oseltamivir Treatment for Influenza in Adults: A Meta-Analysis of Randomised Controlled Trials.” *Lancet (London, England)* 385 (9979): 1729–37. https://doi.org/10.1016/S0140-6736(14)62449-1.

Dudas, Gytis, Trevor Bedford, Samantha Lycett, and Andrew Rambaut. 2015. “Reassortment between Influenza B Lineages and the Emergence of a Coadapted PB1-PB2-HA Gene Complex.” *Molecular Biology and Evolution* 32 (1): 162–72. https://doi.org/10.1093/molbev/msu287.

Dutheil, Julien Y. 2012. “Detecting Coevolving Positions in a Molecule: Why and How to Account for Phylogeny.” *Briefings in Bioinformatics* 13 (2): 228–43. https://doi.org/10.1093/bib/bbr048.

Eisfeld, Amie J., Gabriele Neumann, and Yoshihiro Kawaoka. 2015. “At the Centre: Influenza A Virus Ribonucleoproteins.” *Nature Reviews. Microbiology* 13 (1): 28–41. https://doi.org/10.1038/nrmicro3367.

El Ramahi, Razan, and Alison Freifeld. 2019. “Epidemiology, Diagnosis, Treatment, and Prevention of Influenza Infection in Oncology Patients.” *Journal of Oncology Practice* 15 (4): 177–84. https://doi.org/10.1200/JOP.18.00567.

Essere, Boris, Matthieu Yver, Cyrille Gavazzi, Olivier Terrier, Catherine Isel, Emilie Fournier, Fabienne Giroux, et al. 2013. “Critical Role of Segment-Specific Packaging Signals in Genetic Reassortment of Influenza A Viruses.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (40): E3840-3848. https://doi.org/10.1073/pnas.1308649110.

Ferhadian, Damien, Maud Contrant, Anne Printz-Schweigert, Redmond P. Smyth, Jean-Christophe Paillart, and Roland Marquet. 2018. “Structural and Functional Motifs in Influenza Virus RNAs.” *Frontiers in Microbiology* 9 (March). https://doi.org/10.3389/fmicb.2018.00559.

Fournier, Emilie, Vincent Moules, Boris Essere, Jean-Christophe Paillart, Jean-Daniel Sirbat, Catherine Isel, Annie Cavalier, et al. 2012. “A Supramolecular Assembly Formed by Influenza A Virus Genomic RNA Segments.” *Nucleic Acids Research* 40 (5): 2197–2209. https://doi.org/10.1093/nar/gkr985.

Gavazzi, Cyrille, Catherine Isel, Emilie Fournier, Vincent Moules, Annie Cavalier, Daniel Thomas, Bruno Lina, and Roland Marquet. 2013. “An in Vitro Network of Intermolecular Interactions between Viral RNA Segments of an Avian H5N2 Influenza A Virus: Comparison with a Human H3N2 Virus.” *Nucleic Acids Research* 41 (2): 1241–54. https://doi.org/10.1093/nar/gks1181.

Gavazzi, Cyrille, Matthieu Yver, Catherine Isel, Redmond P. Smyth, Manuel Rosa-Calatrava, Bruno Lina, Vincent Moulès, and Roland Marquet. 2013. “A Functional Sequence-Specific Interaction between Influenza A Virus Genomic RNA Segments.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (41): 16604–9. https://doi.org/10.1073/pnas.1314419110.

Gerber, Marie, Catherine Isel, Vincent Moules, and Roland Marquet. 2014. “Selective Packaging of the Influenza A Genome and Consequences for Genetic Reassortment.” *Trends in Microbiology* 22 (8): 446–55. https://doi.org/10.1016/j.tim.2014.04.001.

Gog, Julia R., Emmanuel Dos Santos Afonso, Rosa M. Dalton, India Leclercq, Laurence Tiley, Debra Elton, Johann C. von Kirchbach, Nadia Naffakh, Nicolas Escriou, and Paul Digard. 2007. “Codon Conservation in the Influenza A Virus Genome Defines RNA Packaging Signals.” *Nucleic Acids Research* 35 (6): 1897–1907. https://doi.org/10.1093/nar/gkm087.

Gultyaev, Alexander P., Ron A. M. Fouchier, and René C. L. Olsthoorn. 2010. “Influenza Virus RNA Structure: Unique and Common Features.” *International Reviews of Immunology*, November. http://www.tandfonline.com/doi/abs/10.3109/08830185.2010.507828.

Hale, Benjamin G., Richard E. Randall, Juan Ortín, and David Jackson. 2008. “The Multifunctional NS1 Protein of Influenza A Viruses.” *The Journal of General Virology* 89 (Pt 10): 2359–76. https://doi.org/10.1099/vir.0.2008/004606-0.

Hayashi, Tsuyoshi, Leslie A. MacDonald, and Toru Takimoto. 2015. “Influenza A Virus Protein PA-X Contributes to Viral Growth and Suppression of the Host Antiviral and Immune Responses.” *Journal of Virology* 89 (12): 6442–52. https://doi.org/10.1128/JVI.00319-15.

Houser, Katherine, and Kanta Subbarao. 2015. “Influenza Vaccines: Challenges and Solutions.” *Cell Host & Microbe* 17 (3): 295–300. https://doi.org/10.1016/j.chom.2015.02.012.

Hsu, Jonathan, Nancy Santesso, Reem Mustafa, Jan Brozek, Yao Long Chen, Jessica P. Hopkins, Adrienne Cheung, et al. 2012. “Antivirals for Treatment of Influenza.” *Annals of Internal Medicine* 156 (7): 512–24. https://doi.org/10.7326/0003-4819-156-7-201204030-00411.

Hu, Yanmei, Hannah Sneyd, Raphael Dekant, and Jun Wang. 2017. “Influenza A Virus Nucleoprotein: A Highly Conserved Multi-Functional Viral Protein as a Hot Antiviral Drug Target.” *Current Topics in Medicinal Chemistry* 17 (20): 2271–85. https://doi.org/10.2174/1568026617666170224122508.

Hutchinson, Edward C., Johann C. von Kirchbach, Julia R. Gog, and Paul Digard. 2010. “Genome Packaging in Influenza A Virus.” *The Journal of General Virology* 91 (Pt 2): 313–28. https://doi.org/10.1099/vir.0.017608-0.

Iuliano, A Danielle, Katherine M Roguski, Howard H Chang, David J Muscatello, Rakhee Palekar, Stefano Tempia, Cheryl Cohen, et al. 2018. “Estimates of Global Seasonal Influenza-Associated Respiratory Mortality: A Modelling Study.” *Lancet (London, England)* 391 (10127): 1285–1300. https://doi.org/10.1016/S0140-6736(17)33293-2.

Jabbari, Hosna, Ian Wark, and Carlo Montemagno. 2018. “RNA Secondary Structure Prediction with Pseudoknots: Contribution of Algorithm versus Energy Model.” *PLOS ONE* 13 (4): e0194583. https://doi.org/10.1371/journal.pone.0194583.

Kosik, Ivan, and Jonathan W. Yewdell. 2019. “Influenza Hemagglutinin and Neuraminidase: Yin–Yang Proteins Coevolving to Thwart Immunity.” *Viruses* 11 (4). https://doi.org/10.3390/v11040346.

Labaronne, Alice, Christopher Swale, Alexandre Monod, Guy Schoehn, Thibaut Crépin, and Rob W. H. Ruigrok. 2016. “Binding of RNA by the Nucleoproteins of Influenza Viruses A and B.” *Viruses* 8 (9). https://doi.org/10.3390/v8090247.

Lampejo, Temi. 2020. “Influenza and Antiviral Resistance: An Overview.” *European Journal of Clinical Microbiology & Infectious Diseases* 39 (7): 1201–8. https://doi.org/10.1007/s10096-020-03840-9.

Layne, Scott P., Arnold S. Monto, and Jeffery K. Taubenberger. 2009. “Pandemic Influenza: An Inconvenient Mutation.” *Science* 323 (5921): 1560–61. https://doi.org/10.1126/science.323.5921.1560.
Le Sage, Valerie, Jack P. Kanarek, Dan J. Snyder, Vaughn S. Cooper, Seema S. Lakdawala, and Nara Lee. 2020. “Mapping of Influenza Virus RNA-RNA Interactions Reveals a Flexible Network.” *Cell Reports* 31 (13): 107823. https://doi.org/10.1016/j.celrep.2020.107823.
Le Sage, Valerie, Adalena V. Nanni, Amar R. Bhagwat, Dan J. Snyder, Vaughn S. Cooper, Seema S. Lakdawala, and Nara Lee. 2018. “Non-Uniform and Non-Random Binding of Nucleoprotein to Influenza A and B Viral RNA.” *Viruses* 10 (10). https://doi.org/10.3390/v10100522.
Lemieux, Camille, Gabrielle Brankston, Leah Gitterman, Zahir Hirji, and Michael Gardam. 2007. “Questioning Aerosol Transmission of Influenza.” *Emerging Infectious Diseases* 13 (1): 173–75. https://doi.org/10.3201/eid1301.061202.
Li, Xiuli, Min Gu, Qinmei Zheng, Ruyi Gao, and Xiufan Liu. 2021. “Packaging Signal of Influenza A Virus.” *Virology Journal* 18 (February). https://doi.org/10.1186/s12985-021-01504-4.
Lowen, Anice C. 2018. “It’s in the Mix: Reassortment of Segmented Viral Genomes.” *PLoS Pathogens* 14 (9). https://doi.org/10.1371/journal.ppat.1007200.
Lyons, Daniel M., and Adam S. Lauring. 2018. “Mutation and Epistasis in Influenza Virus Evolution.” *Viruses* 10 (8): 407. https://doi.org/10.3390/v10080407.
Ma, Wenjun, Robert E Kahn, and Juergen A Richt. 2008. “The Pig as a Mixing Vessel for Influenza Viruses: Human and Veterinary Implications.” *Journal of Molecular and Genetic Medicine : An International Journal of Biomedical Research* 3 (1): 158–66.
Mann, Martin, Patrick R. Wright, and Rolf Backofen. 2017. “IntaRNA 2.0: Enhanced and Customizable Prediction of RNA–RNA Interactions.” *Nucleic Acids Research* 45 (Web Server issue): W435–39. https://doi.org/10.1093/nar/gkx279.
Masters, Paul S. 2019. “Coronavirus Genomic RNA Packaging.” *Virology* 537 (November): 198–207. https://doi.org/10.1016/j.virol.2019.08.031.
McDonald, Sarah M., Martha I. Nelson, Paul E. Turner, and John T. Patton. 2016. “Reassortment in Segmented RNA Viruses: Mechanisms and Outcomes.” *Nature Reviews. Microbiology* 14 (7): 448–60. https://doi.org/10.1038/nrmicro.2016.46.
Morris, Denise E., David W. Cleary, and Stuart C. Clarke. 2017. “Secondary Bacterial Infections Associated with Influenza Pandemics.” *Frontiers in Microbiology* 8 (June). https://doi.org/10.3389/fmicb.2017.01041.
Moura, Fernanda E. A. 2010. “Influenza in the Tropics.” *Current Opinion in Infectious Diseases* 23 (5): 415–20. https://doi.org/10.1097/QCO.0b013e32833cc955.
Neumann, Gabriele, Mark T. Hughes, and Yoshihiro Kawaoka. 2000. “Influenza A Virus NS2 Protein Mediates VRNP Nuclear Export through NES-Independent Interaction with HCRM1.” *The EMBO Journal* 19 (24): 6751–58. https://doi.org/10.1093/emboj/19.24.6751.
Newburn, Laura R., and K. Andrew White. 2019. “Trans-Acting RNA–RNA Interactions in Segmented RNA Viruses.” *Viruses* 11 (8). https://doi.org/10.3390/v11080751.
Nicholson, Beth L., and K. Andrew White. 2014. “Functional Long-Range RNA–RNA Interactions in Positive-Strand RNA Viruses.” *Nature Reviews Microbiology* 12 (7): 493–504. https://doi.org/10.1038/nrmicro3288.
Noda, Takeshi, Yukihiko Sugita, Kazuhiro Aoyama, Ai Hirase, Eiryo Kawakami, Atsuo Miyazawa, Hiroshi Sagara, and Yoshihiro Kawaoka. 2012. “Three-Dimensional Analysis of Ribonucleoprotein Complexes in Influenza A Virus.” *Nature Communications* 3 (1): 639. https://doi.org/10.1038/ncomms1647.
Nuñez, Ivette A., and Ted M. Ross. 2019. “A Review of H5Nx Avian Influenza Viruses.” *Therapeutic Advances in Vaccines and Immunotherapy* 7 (February). https://doi.org/10.1177/2515135518821625.
Park, Chin-Ju, Sung-Hun Bae, Mi-Kyung Lee, Gabriele Varani, and Byong-Seok Choi. 2003. “Solution Structure of the Influenza A Virus CRNA Promoter: Implications for Differential Recognition of Viral Promoter Structures by RNA-Dependent RNA Polymerase.” *Nucleic Acids Research* 31 (11): 2824–32.
Peselis, Alla, and Alexander Serganov. 2014. “Structure and Function of Pseudoknots Involved in Gene Expression Control.” *Wiley Interdisciplinary Reviews. RNA* 5 (6): 803–22. https://doi.org/10.1002/wrna.1247.
Pielak, Rafal M., and James J. Chou. 2011. “Influenza M2 Proton Channels.” *Biochimica et Biophysica Acta* 1808 (2): 522–29. https://doi.org/10.1016/j.bbamem.2010.04.015.
Ríos, Fernando G., Elisa Estenssoro, Fernando Villarejo, Ricardo Valentini, Liliana Aguilar, Daniel Pezzola, Pascual Valdez, et al. 2011. “Lung Function and Organ Dysfunctions in 178 Patients Requiring Mechanical Ventilation During The 2009 Influenza A (H1N1) Pandemic.” *Critical Care* 15 (4): R201. https://doi.org/10.1186/cc10369.
Ruigrok, Rob W. H., Annie Barge, Peter Durrer, Josef Brunner, Kai Ma, and Gary R. Whittaker. 2000. “Membrane Interaction of Influenza Virus M1 Protein.” *Virology* 267 (2): 289–98. https://doi.org/10.1006/viro.1999.0134.
Shafiuddin, Md, and Adrianus C. M. Boon. 2019. “RNA Sequence Features Are at the Core of Influenza A Virus Genome Packaging.” *Journal of Molecular Biology* 431 (21): 4217–28. https://doi.org/10.1016/j.jmb.2019.03.018.
Shao, Wenhan, Xinxin Li, Mohsan Ullah Goraya, Song Wang, and Ji-Long Chen. 2017. “Evolution of Influenza A Virus by Mutation and Re-Assortment.” *International Journal of Molecular Sciences* 18 (8): 1650. https://doi.org/10.3390/ijms18081650.

Short, Kirsty R., Edwin J. B. Veldhuis Kroeze, Ron A. M. Fouchier, and Thijs Kuiken. 2014. “Pathogenesis of Influenza-Induced Acute Respiratory Distress Syndrome.” *The Lancet. Infectious Diseases* 14 (1): 57–69. https://doi.org/10.1016/S1473-3099(13)70286-X.

Smith, Gavin J. D., Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu Kit Ma, et al. 2009. “Origins and Evolutionary Genomics of the 2009 Swine-Origin H1N1 Influenza A Epidemic.” *Nature* 459 (7250): 1122–25. https://doi.org/10.1038/nature08182.

Smyth, Redmond P., Matteo Negroni, Andrew M. Lever, Johnson Mak, and Julia C. Kenyon. 2018. “RNA Structure—A Neglected Puppet Master for the Evolution of Virus and Host Immunity.” *Frontiers in Immunology* 9 (September). https://doi.org/10.3389/fimmu.2018.02097.

Su, Shuo, Xinliang Fu, Gairu Li, Fiona Kerlin, and Michael Veit. 2017. “Novel Influenza D Virus: Epidemiology, Pathology, Evolution and Biological Characteristics.” *Virulence* 8 (8): 1580–91. https://doi.org/10.1080/21505594.2017.1365216.

Sun, Siyang, Venigalla B. Rao, and Michael G. Rossmann. 2010. “Genome Packaging in Viruses.” *Current Opinion in Structural Biology* 20 (1): 114–20. https://doi.org/10.1016/j.sbi.2009.12.006.

Szabat, Marta, and Ryszard Kierzek. 2017. “Parallel-Stranded DNA and RNA Duplexes – Structural Features and Potential Applications.” *The FEBS Journal* 284 (23): 3986–98. https://doi.org/10.1111/febs.14187.

Tao, Hui, John Steel, and Anice C. Lowen. 2014. “Intrahost Dynamics of Influenza Virus Reassortment.” *Journal of Virology* 88 (13): 7485–92. https://doi.org/10.1128/JVI.00715-14.

Tarus, Bogdan, Christophe Chevalier, Charles-Adrien Richard, Bernard Delmas, Carmelo Di Primo, and Anny Slama-Schwok. 2012. “Molecular Dynamics Studies of the Nucleoprotein of Influenza A Virus: Role of the Protein Flexibility in RNA Binding.” *PLOS ONE* 7 (1): e30038. https://doi.org/10.1371/journal.pone.0030038.

Taubenberger, Jeffery K., and David M. Morens. 2006. “1918 Influenza: The Mother of All Pandemics.” *Emerging Infectious Diseases* 12 (1): 15–22. https://doi.org/10.3201/eid1201.050979.

Tellier, Raymond. 2006. “Review of Aerosol Transmission of Influenza A Virus.” *Emerging Infectious Diseases* 12 (11): 1657–62. https://doi.org/10.3201/eid1211.060426.

Trifkovic, Sanja, Brad Gilbertson, Emily Fairmaid, Joanna Cobbin, Steven Rockman, and Lorena E. Brown. 2021. “The Role of Gene Segment Interactions in Driving the Emergence of Dominant Gene Constellations during Influenza Virus Reassortment.” *BioRxiv*, February, 2021.02.10.430697. https://doi.org/10.1101/2021.02.10.430697.

Turner, Douglas H., and David H. Mathews. 2010. “NNDB: The Nearest Neighbor Parameter Database for Predicting Stability of Nucleic Acid Secondary Structure.” *Nucleic Acids Research* 38 (Database issue): D280-282. https://doi.org/10.1093/nar/gkp892.

Umu, Sinan Ugur, and Paul P. Gardner. 2017. “A Comprehensive Benchmark of RNA-RNA Interaction Prediction Tools for All Domains of Life.” *Bioinformatics (Oxford, England)* 33 (7): 988–96. https://doi.org/10.1093/bioinformatics/btw728.

Varga, Zsuzsanna T., Alesha Grant, Balaji Manicassamy, and Peter Palese. 2012. “Influenza Virus Protein PB1-F2 Inhibits the Induction of Type I Interferon by Binding to MAVS and Decreasing Mitochondrial Membrane Potential.” *Journal of Virology* 86 (16): 8359–66. https://doi.org/10.1128/JVI.01122-12.

Velthuis, Aartjan J. W. te, and Ervin Fodor. 2016. “Influenza Virus RNA Polymerase: Insights into the Mechanisms of Viral RNA Synthesis.” *Nature Reviews Microbiology* 14 (8): 479–93. https://doi.org/10.1038/nrmicro.2016.87.

Venev, Sergey V., and Konstantin B. Zeldovich. 2013. “Segment Self-Repulsion Is the Major Driving Force of Influenza Genome Packaging.” *Physical Review Letters* 110 (9): 098104. https://doi.org/10.1103/PhysRevLett.110.098104.

Viboud, Cécile, Wladimir J Alonso, and Lone Simonsen. 2006. “Influenza in Tropical Regions.” *PLoS Medicine* 3 (4). https://doi.org/10.1371/journal.pmed.0030089.

Vijaykrishna, Dhanasekaran, Reshmi Mukerji, and Gavin J. D. Smith. 2015. “RNA Virus Reassortment: An Evolutionary Mechanism for Host Jumps and Immune Evasion.” *PLoS Pathogens* 11 (7). https://doi.org/10.1371/journal.ppat.1004902.

Watanabe, Tokiko, Maki Kiso, Satoshi Fukuyama, Noriko Nakajima, Masaki Imai, Shinya Yamada, Shin Murakami, et al. 2013. “Characterization of H7N9 Influenza A Viruses Isolated from Humans.” *Nature* 501 (7468): 551–55. https://doi.org/10.1038/nature12392.

Webster, Robert G., and Elena A. Govorkova. 2014. “Continuing Challenges in Influenza.” *Annals of the New York Academy of Sciences* 1323 (1): 115–39. https://doi.org/10.1111/nyas.12462.

Weeks, Kevin M. 2010. “Advances in RNA Secondary and Tertiary Structure Analysis by Chemical Probing.” *Current Opinion in Structural Biology* 20 (3): 295–304. https://doi.org/10.1016/j.sbi.2010.04.001.

Weinstein, Robert A., Carolyn Buxton Bridges, Matthew J. Kuehnert, and Caroline B. Hall. 2003. “Transmission of Influenza: Implications for Control in Health Care Settings.” *Clinical Infectious Diseases* 37 (8): 1094–1101. https://doi.org/10.1086/378292.

Wheeler, David, and Medha Bhagwat. 2007. *BLAST QuickStart*. *Comparative Genomics: Volumes 1 and 2*. Humana Press. https://www.ncbi.nlm.nih.gov/books/NBK1734/.

WHO. 2021. “Influenza - Surveillance and Monitoring.” Influenza - Surveillance and Monitoring. World Health Organization. 2021. http://www.who.int/influenza/surveillance_monitoring/en/.

Williams, Graham D., Dana Townsend, Kristine M. Wylie, Preston J. Kim, Gaya K. Amarasinghe, Sebla B. Kutluay, and Adrianus C. M. Boon. 2018. “Nucleotide Resolution Mapping of Influenza A Virus Nucleoprotein-RNA Interactions Reveals RNA Features Required for Replication.” *Nature Communications* 9 (1): 465. https://doi.org/10.1038/s41467-018-02886-w.

Ye, Qiaozhen, Robert M. Krug, and Yizhi Jane Tao. 2006. “The Mechanism by Which Influenza A Virus Nucleoprotein Forms Oligomers and Binds RNA.” *Nature* 444 (7122): 1078–82. https://doi.org/10.1038/nature05379.

Zuker, M. 1989. “On Finding All Suboptimal Foldings of an RNA Molecule.” *Science (New York, N.Y.)* 244 (4900): 48–52. https://doi.org/10.1126/science.2468181.