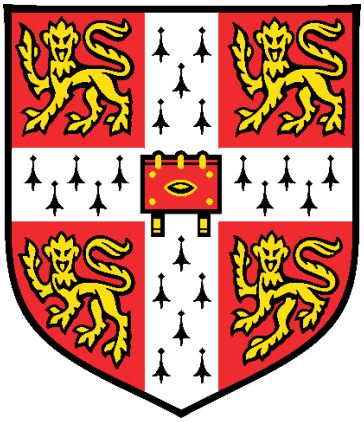
How codon choice determines evolvability and evolutionary robustness in short linear motifs



Peter Alexander Gunnarsson

MRC Laboratory of Molecular Biology

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Hughes Hall

October 2019

Till mormor

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed limit of 60,000 words as defined by the Department of Biology Degree Committee.

Peter Alexander Gunnarsson

October 2019

How codon choice determines evolvability and evolutionary robustness in short linear motifs

Short linear motifs, made up of 2-10 amino acids in linear sequence space, are a central component of cellular decision making through proteins. They form a modular system in cells where combinations of domains and motifs are used as basic functional building blocks through interactions. Functions mediated through these motifs include cellular localisation, post-translational modifications, degradation and general protein-protein interactions. Since motifs are made up of a small number of amino acids they have unusual evolutionary properties, for instance they can evolve *de novo*, or be lost, through a small number of substitutions. This is of particular importance in pathogens such as viruses. Many viruses evolve new host-like motifs to interact with the host and change the regulation and signalling landscape within host cells to mediate infection.

In this body of work, I have used influenza as a model to elucidate aspects of the evolutionary properties of motifs. I have been able to leverage recent progress made in determining nucleotide mutation rates and have developed a model for motif evolution that is defined from the nucleotide and codon levels. Simulations using this methodology suggested that different codons have varying propensities to evolve into amino acids within a linear motif. In other words, some sequences have higher motif evolvability. The simulations also indicated a fitness benefit to use some codons over others to encode linear motifs, due to the varying propensity to evolve. These findings suggest that motifs that are encoded by specific codons have higher motif evolutionary robustness, i.e. they can tolerate more mutations without affecting function.

I went on to investigate if these predicted properties have played a role in motif evolution in influenza. I found that conserved motifs in influenza use the codons inferred to have higher evolutionary robustness. This would lead to increased fitness, as motifs are less often lost through mutations. I also found that this mutational robustness acts on stop codon usage in influenza, suggesting an explanation for an old observation of predominant use of TAA in many organisms. Interestingly, it also appears that evolutionary robustness of a motif can be varied to tune the rate of motif change, which influenza utilises in glycosylation motifs that interface with the host immune system.

Finally, I investigated whether the codon choice and evolvability at early stages of viral host shifts could be used to predict the emergence of functional motifs. I have found that motif evolvability can aid the prediction of motif emergence. For influenza strains H1N1 and H3N2, which were introduced in the human population from birds during the 1900s, the sequence of the early strains could be used to predict the majority of the glycosylation sites that would emerge the following decades. The predictability of motif emergence could have important implications for vaccination efforts.

The methodologies developed here, and the observations made about how motif evolution is shaped by codon choices in a predictable way will be important for a better understanding of the evolution of complexity and regulation involving motifs. This may have implications for complex diseases such as cancers, and for our understanding of the evolution of pathogen innovations and functionality.

Peter Alexander Gunnarsson

Acknowledgements

This thesis would not have been possible without the inspiration, mentorship and support from a large number of people. I want to sincerely thank everyone who have listened, questioned, provided suggestions, and been generally supportive through the three years I have spent trying to make sense of the world. There are several people to whom I want to express my deepest gratitude and appreciation.

First and foremost to my supervisor Madan Babu, who has supported me throughout and who has been more understanding, inspiring and guiding than I could have ever wished for. Our discussions and exchange of ideas shaped this work from the beginning, and through his mentorship I have been given a strong scientific foundation for the future. It has been an incredibly valuable time.

To the past and present Babu group, who have endured my repeated questions, sometimes dull presentations and my general presence for over three years. And particularly to Greg Slodkowicz, Maria Marti Solano, Xiaohan Li, Duccio Malinverni, Yonathan Goldtzvik, Andal Murthy and Charles Ravarani for providing me with critical support and feedback during the writing of this thesis.

To all my past and present mentors, who have guided and supported me throughout my academic career and without whom I would not be where I am today; To Micke Kuwahara, Gunilla Naucler, Paul Keane and Brad Houston without whom I would have never made it to Cambridge in the first place; To Alexey Morgunov who fanned the spark that led me to pursuing molecular biology and who has been a cornerstone for my many opportunities and successes here in Cambridge. I am eternally grateful.

To my dear friends and colleagues throughout my time in Cambridge who have been an inexorable source of fun, stimulating conversations and grand adventures. Cheers.

To my housemates, Joe Thompson, Pauline Kiesow and Andrei Smid who have taken care of me and supported me through the last trying weeks of write up, providing hot meals, general cheering-up and respite during stressful times.

To my family I want to express my love and gratitude. I am forever indebted to my parents Ove and Katharina Gunnarsson for their love and boundless support and belief in me, from Sweden to Cambridge and beyond. And to my siblings, Josefine and Jakob Gunnarsson with whom I have shared many great adventures and late nights of conversation. Despite the distance, you are never far away.

And finally to my partner Helen Waters, who has been the wind in my sails through this pursuit; an unwavering force that helped me move forwards when I faltered. Thank you.

Table of contents

List of figures

List of tables

Abbreviations

ELM	Eukaryotic linear motif database
GSK3	Glycogen synthase kinase 3
HA	Haemagglutinin
IDP	Intrinsically disordered protein
IDR	Intrinsically disordered region
M1	Matrix protein 1
M2	Matrix protein 2
NA	Neuraminidase
NP	Nucleoprotein
NS1	Non-structural 1
NS2	Non-structural 2
PA	Polymerase acidic
PB1	Polymerase basic 1
PB2	Polymerase basic 2
PKA	Protein kinase A
PTM	Post-translational modification
SLiMs	Short linear motifs

Chapter 1
General introduction

1.1. Overview

A cell is fundamentally a chaotic, stochastic set of chemical reactions, yet at the same time an efficient, highly regulated and precise living system. To understand how both of these things can be true, it is essential to consider that biological systems operate in crowded and highly complex environments that rely on the dynamic properties of large numbers of molecules to process information. This perspective is increasingly defining current molecular biological research.

In molecular biology there has historically been an emphasis on categorising biological problems and thinking about them in isolation in a way that pertains to the questions we are interested in. Sometimes it is useful to think about molecular pathways as a series of discrete and isolated ordered events, such as a transcription factor binding in an orderly fashion to a promoter, recruiting the transcription machinery which in turn creates an mRNA that is translated into a new molecular machine. In reality it is rarely so orderly. In the crowded chemical soup of the cell, balancing forces of concentrations, gradients, relative activity levels, diffusion and temperature all orchestrate the properties that give a cell its functionality. Slight shifts in concentration equilibria or enzyme activity levels can have drastic changes on the cell environment. Chance encounters can trigger unintended modifications and reactions. Some proteins get misfolded more frequently than they fold correctly (Schubert *et al.* 2000). Some proteins get degraded before they even perform their function. Yet, despite all the chaos, 37 trillion cells in each human body can work in unison to make all of us more or less function as intended.

To address more complex biological questions, that integrate information from the molecular dynamics of the cell and the large numbers of molecules involved, we need a systems approach to biology (Aderem 2005). In recent years we have seen the emergence of this systems level biology, where large scale data is used to characterise the complex dynamics, interactions and reactions by looking at the whole picture (Tavassoly *et al.* 2018). The fields of genomics, proteomics, transcriptomics, interactomics and others all aim to characterise large datasets of the complex dynamics within cells. With this increasing focus on systems properties of cells, the importance of interaction dynamics, interaction networks and regulation at all levels of molecular biology has become even more evident.

Within this systems view of biology, understanding evolution and evolutionary forces becomes equally complex. Fundamental questions in the field involve the characterisation of the mechanisms of molecular evolution, and how they affect the systems properties of cells in evolution and disease. There are many ways mutations can affect interaction networks and information processing through a small number of sequence changes, and these properties have only just begun to be elucidated. Currently there is a big gap in our understanding of molecular evolution at the intersection between systems level biology and changes in individual molecules. In this thesis I have used a predominantly systems based approach to explore the evolutionary aspects of features within these networks, specifically short linear motifs, that mediate information processing and cellular decision making.

In the following introduction chapter, I first outline the fundamental basis for protein function through the chemical properties of the chain of amino acids. This is at the foundation of motif function and highlights the emergence of the systems properties of molecular biology from the ground up. This also establishes the relevance of this systems approach to the questions of the evolution of short linear motifs within viral proteins for the purposes of altering the complex regulation required during infections. These fundamental properties are key in thinking about both the driving forces behind the interactions driven by short linear motifs and the functions they mediate. I subsequently highlight what is currently well established within motif function and evolution, in particular with viral motif mimicry. I conclude this chapter with a discussion and overview of the key objectives of this PhD research that have been addressed in the thesis.

1.2. Protein structure and function

At their most fundamental level, living organisms are a product of the various chemical interactions and reactions driven by the basic molecules of life. One of the most important paradigm shifts in our understanding of the function of organisms was the structure-function paradigm of protein science. The realisation that protein function is defined by its structure in three dimensional space driven by early findings of protein shape and sequence by Linus Pauling and Frederick Sanger among others (Pauling *et al.* 1951; Sanger & Tuppy 1951). The structure is fundamentally a consequence of the chemistry of the amino acids. The charge, size, polarity, electronegativity and structure of the atoms within amino acids shape them and their chemistry, giving rise to the main biophysical properties that influence protein folding; mainly ionic interactions, hydrophobicity, pi-pi stacking, hydrogen bonding and van der Waals forces (Figure 1.1). As they are translated by the ribosome these forces create interactions that fold the protein. In the landmark experiments of Anfinsen on ribonuclease folding, he established that the primary sequence of an amino acids defines the protein structure as the fold constitutes an energy minimum where the protein is most stable in solution (Anfinsen 1973). It is also these

same properties of amino acids that shape the ultimate function of the folded protein. For example, the appropriate charge or electron-acceptor positioned in the right place allow chymotrypsin to catalyse the breakage of the peptide bond (Neurath *et al.* 1967); the shape and charge distribution of histones allow them to bind and package DNA closely in the nucleus (Luger *et al.* 1997); and in ribonuclease A, charge and positioning of both non-specific RNA backbone binding residues and acid-base catalysing histidines determine its function and specificity (Cuchillo *et al.* 2011; Findly *et al.* 1961) (Figure 1.1). The structure-function paradigm is a backbone of modern biology and has shaped our understanding of all the processes in the body, and the ability to view the structures through x-ray crystallography further cemented the importance of the three dimensional structure for many key proteins.

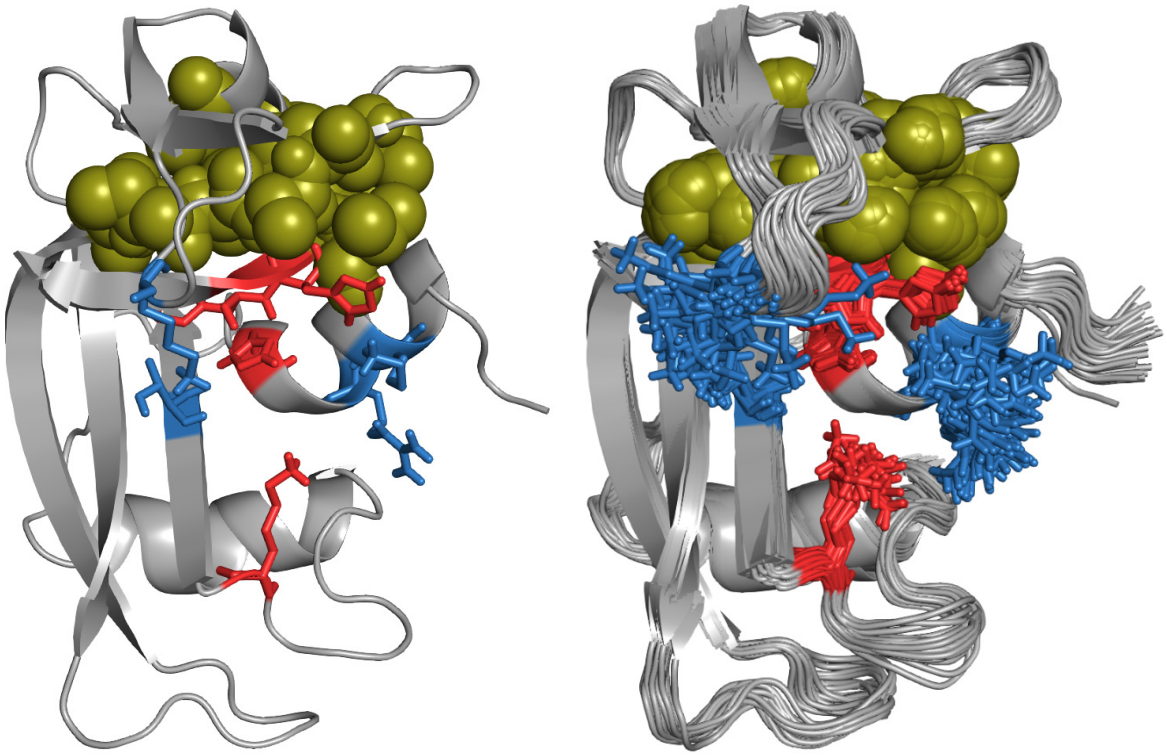


Figure 1.1. NMR Structure of Ribonuclease A. Some of the properties of protein folding and enzymatic function highlighted. In red are the residues involved in the catalytic activity. His12, His119 perform acid/base catalysis and the charge of Lys41 stabilises the transition state. In blue are Lys7, Lys66 and Arg10 that bind to the negatively charged backbone of DNA and Thr45 which forms a hydrogen bond specifically with pyrimidines. In gold/olive colour are highlighted in a space filling model residues involved in hydrophobic packing of parts of the globular core of the protein. On the right an NMR structure ensemble is shown to illustrate that even fully folded enzymes are dynamic structures where many parts move continuously, although some parts (e.g. the catalytic residues, red) move much less than others (e.g. RNA backbone binding residues in blue). PDB: 2AAS

However, the structure-function model is a simplification that overlooks the importance of many key processes in the cell that heavily contribute to life. Furthermore, it overemphasises proteins as rigid, defined shapes where in reality they are highly dynamic shapes that rely on movement to function (Figure 1.1). In recent years we have found increasingly that the classical structure-function paradigm is insufficient to further our understanding of important biological questions. Cells (and organisms)

can be argued to fundamentally be information processing units that respond to chemical and physical information input and produces a response with the goal of self-replication. The specific structure of proteins in three dimensions is an important contributor to these systems, but form only a part of the full system. To understand complex biology like diseases such as cancer, and viral infections – as well as organismal development and evolution – we need to look at all the contributors in the system and all the ways in which they interact and change. Protein structure can be thought of as the first order of information processing in the cell – the chemical executors. The higher order information processing includes the interactions, chemical modifications, encoded localisation and specific context of the proteins (and other molecules) in the cell that regulate and modulate the function of the executors. To a large extent these functions are driven by proteins that do not have a defined three dimensional structure but that still are defined by the same chemical properties of amino acids to process signals and information. These proteins are generally referred to as disordered proteins or unstructured proteins. To fully understand the properties and functions of these proteins a more systems based view is essential. The functions in these contexts rely in large part on weak interactions and balances between multiple states at different proportions in a population of molecules.

1.3. Protein disorder and function

Intrinsically disordered proteins (IDPs) or intrinsically disordered regions (IDRs) are proteins or parts of proteins without a defined three dimensional structure under normal physiological conditions (Babu 2016; van der Lee *et al.* 2014). The importance of IDRs has become increasingly clear over the last two decades as research has elucidated their functional importance in many cellular contexts (Tompa 2012). The importance of IDRs was initially significantly underappreciated due to their inability to crystallise and therefore to be elucidated structurally. Their functional roles, outside of being flexible linkers in proteins, were generally poorly understood (Wright & Dyson 1999). However, we now know that protein disorder is prevalent in eukaryotes, with ~40% of human proteins estimated to contain intrinsically disordered regions, and that they are functionally important in many cellular contexts, with the function mediated specifically by the disordered state (Dyson & Wright 2005; Oates *et al.* 2012; Wright & Dyson 2015).

Disordered proteins have very different functions and properties to structured proteins. Fundamentally however, it is still the same chemical properties of the linear peptide sequence that ensure that disordered proteins remain unfolded (Latysheva *et al.* 2015; Mittag *et al.* 2010). This is achieved through a skewed amino acid composition, favouring charged and polar amino acids and being depleted in small and hydrophobic amino acids (Dunker *et al.* 2001). These properties give the peptide a flatter energy landscape which enables it to sample a large conformational ensemble and it is not locked in a single

energy minimum in a folded state (Fisher & Stultz 2011). This means that the majority of amino acids in a disordered region are exposed to the surrounding solvent and are thus accessible to other molecules in the cell (Figure 1.2). This is a fundamental property that gives disordered proteins the ability to function in unique ways in comparison with structured proteins.

Figure 1.2. Conformational ensemble of intrinsically disordered protein. Left: a single conformation of CsgF. Right: 20 different conformations as captured by NMR. PDB: 5M1U

IDRs and IDPs contribute to function and complexity in multiple ways (see (van der Lee *et al.* 2014; Van Roey & Davey 2015)). Their function is derived from the flexible and solvent exposed properties of the peptides through conformational ensembles, multivalent interactions and the ability to be modified and regulate function in a complex manner (Fung *et al.* 2018). The ability to bind a large range of different partners is a crucial role in many cellular contexts, importantly as central hubs and bottlenecks in interaction and signalling networks. It has been shown that intrinsic disorder is significantly enriched in proteins that perform these roles (Dosztanyi *et al.* 2006; Haynes *et al.* 2006). They are also extensively post translationally modified to regulate the function of proteins or to combine many sources of interaction information and regulatory modifications. Interactions and modifications are predominantly mediated through short linear motifs, a linear sequence of amino acids that can be recognised by various other proteins and molecules (discussed in section 1.4) (Davey *et al.* 2012; Van Roey *et al.* 2014). One of the most important functional consequences of this is that local context, regulation of interactions and many levels of chemical modifications can be combined to achieve complex information processing (Van Roey *et al.* 2012). The large number of transient interactions also allows disordered proteins to play an important role in phase transitions in cells (Boeynaems *et al.*

2018). These phase transitions are important in the nucleus as part of the nucleolus and during stress conditions in the cell in the formation of stress granules (Harmon *et al.* 2017; Li *et al.* 2018a).

The increased complexity in regulation and information suggests we might find a correlation between disorder and organismal complexity as well. That turns out to be the case: more complex eukaryotes have significantly more disordered proteins than prokaryotes and single celled eukaryotes (Schad *et al.* 2011). Importantly, many viruses and other pathogens also have high levels of disorder, contrary to what might be expected given their relatively small genomes (Pushker *et al.* 2013). There seems to be a link between virus protein disorder and their adaptability and consequently the difficulty for us to treat or vaccinate against the virus, in particular with disorder in the coat proteins as in the case of HIV (Tokuriki *et al.* 2009; Xue *et al.* 2012). Many viruses also use disorder as a means to interact with and mimic properties of host proteins to rewire and subvert the cellular machinery (Davey *et al.* 2011; Hagai *et al.* 2014) (discussed in section 1.7).

The importance of disordered proteins and their roles for functional complexity can be seen in almost all cellular contexts and functions. Disordered proteins are key elements for the function and regulation of proteins including: p53, in regulating the DNA damage response; RNA Polymerase II, at the well-studied C-terminal domain; the nuclear pore complex, where they form a phase separated gel that ensures the integrity of the nucleus by restricting movement through the pore; and GPCRs, where they ensure the regulation and attenuation for signalling through the arrestin pathway (Denning *et al.* 2003; Hafner *et al.* 2019; Komarnitsky *et al.* 2000; Venkatakrishnan *et al.* 2014).

As mentioned previously, the histone performs its function based on the structure and biophysical properties (importantly charge) of its globular core subunits. However, the histone also contains long disordered tails that have been well studied and are extremely important for the higher order function and regulation of chromatin packing and gene expression. The histone tails contain motifs for a range of chemical modifications sites (Post-translational modifications or PTMs) including phosphorylation, acetylation, methylation, sumoylation and ubiquitylation and in addition contain a substantial amount of recognition motifs for binding other protein domains including important chromatin remodellers (Figure 1.3) (Bannister & Kouzarides 2011). The binding sites interact with the modifications to regulate timing and occurrence of binding events so that chromatin packing and gene expression can be tightly regulated.

Disordered proteins are also known to be centrally involved in – and the cause of – many diseases (Babu *et al.* 2011; Buljan *et al.* 2013). Due to the flat energy landscape of disordered protein conformation, they can be prone to aggregate and fold into amyloid which is thought to be a global energy minimum for most peptides (Perczel *et al.* 2007). The proteins that are thought to commonly fold into amyloid and aggregate include well known IDPs such as α-synuclein and beta-amyloid that are involved in neurodegenerative diseases such as Alzheimer's, Parkinson's and dementia (Uversky 2015).

Disordered proteins can also be part of diseases without aggregation, importantly in cancers. It has been established that changes to splicing and genetic reorganisation that cause protein fusions frequently and significantly include IDRs (Latysheva *et al.* 2016; Li *et al.* 2017). These changes to the regulation of protein activity and function are overrepresented in cancers and many of the cancer hallmarks are importantly related to these same mechanisms of change to regulation, expression and function (Sever & Brugge 2015). Cancer is predominantly a disease that results from changes in the higher order protein functionality rather than a defect in the primary function of individual proteins, and therefore changes to the disordered segments have a disproportionate impact on those kinds of diseases (Iakoucheva *et al.* 2002).

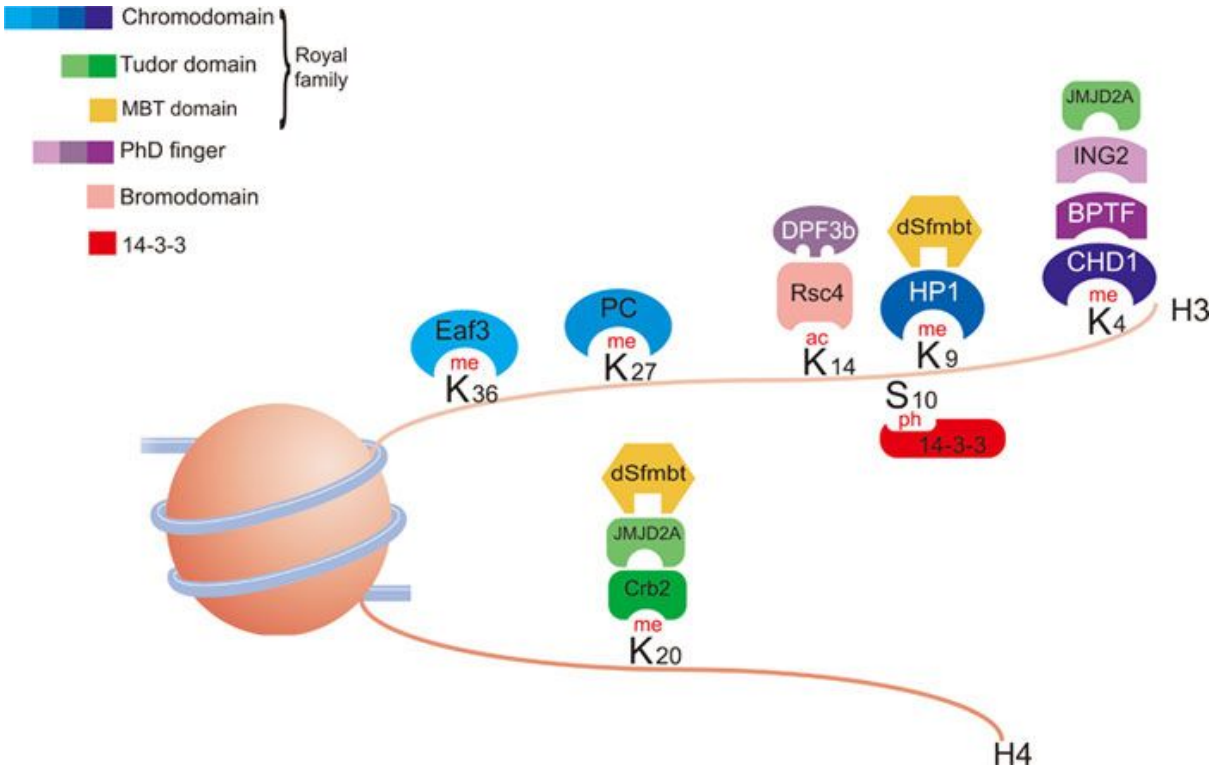


Figure 1.3. The histone code. This figure shows the complex landscape of modifications and interactions formed through the histone tails. This complex regulation is crucial for gene expression and chromatin packing. This example shows methylation, acetylation and phosphorylation with associated binders that recognise modified sequences. Figure reproduced from Bannister & Kouzarides, (2011) with permission from Springer Nature, Macmillan Publishers Ltd.

In all, protein disorder is at the heart of cell systems and essential for understanding biology at the systems level. Protein disorder allows proteins and cells to integrate many sources of signalling and regulation to make complex decisions about cell fate, development, and in responding to stressors.

1.4. Properties of short linear motifs

One of the central functional units in disordered segments is the short linear motif. It is the main contributor to functions involving interactions and chemical modifications as discussed in 1.3. Short linear motifs (SLiMs or simply "motifs") usually consist of a short linear sequence of amino acids as short as 2 and as long as 15, however the majority fall in the range 4-8 (Davey *et al.* 2012; Van Roey *et al.* 2014). Motifs are usually a combination of key amino acids that are recognised and bind to the pocket of an interacting domain, and several spacing amino acids that are less restricted but tend to have some limitations (Figure 1.4 and Figure 1.5). The amino acids that are key at defining interactions are enriched in hydrophobic amino acids and also charged and polar amino acids (Davey *et al.* 2012). The enrichment of hydrophobic amino acids is in contrast to the rest of the disordered region and they are important for binding to hydrophobic pockets in partner domains. Motifs tend to bind with a lower affinity than other modes of binding such as interaction surfaces between globular proteins, but in certain ways can increase the binding affinity through retaining higher entropy in some contexts, which has been described as fuzzy binding (Flock *et al.* 2014; Tompa & Fuxreiter 2008). Motifs can be divided into functional classes describing their range of functional interactions in the cell (based on classification in the Eukaryotic Linear Motif database (ELM) (Gouw *et al.* 2017)):

1. Localisation/Targeting: These motifs act as short barcodes being recognised by proteins involved in transport. Well studied examples for intracellular transport and localisation include the protein sorting machinery at the ER-Golgi interface which relies on motifs and transport proteins such as COPI and COPII to direct proteins to the right compartments (Gomez-Navarro & Miller 2016). Targeting motifs are also centrally involved in aspects of cytoskeletal transport, nuclear localisation and nuclear export to name a few (Honnappa *et al.* 2009; Kumar *et al.* 2017; Lange *et al.* 2007).

2. Cleavage motifs: These are recognition sites for proteases to cleave peptides and proteins, which is a key function in many signalling cascades and crucial for the activation of precursor proteins through cleavage. An example of this is in the activation of Thrombin receptors which are protease-activated receptors where the cleavage product of the terminus is also the receptor ligand (Gallwitz *et al.* 2012; Vu *et al.* 1991). Caspase 3 and 7 are also important examples of cellular proteases that are involved in regulation and execution of apoptosis through the recognition of cleavage motifs (Fischer *et al.* 2003).

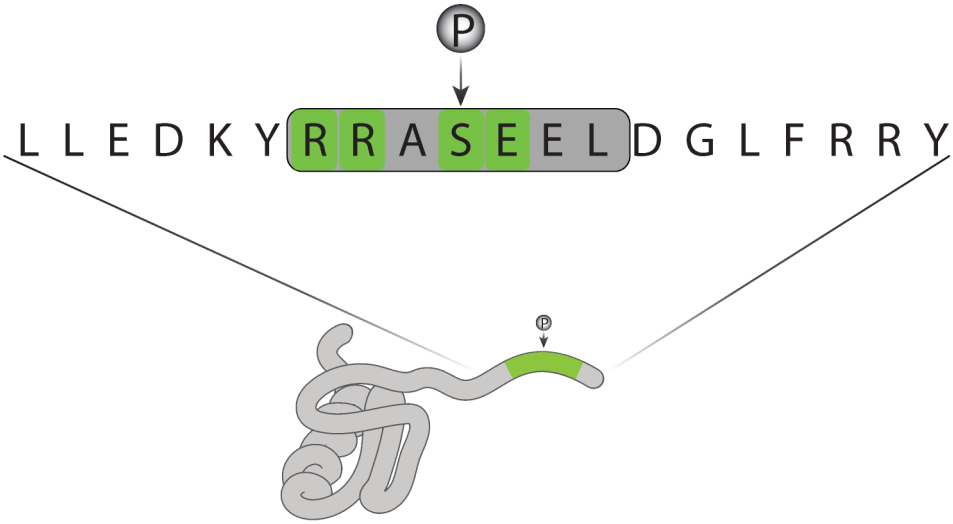


Figure 1.4. Typical representation of a motif in sequence. Motifs are predominantly found in accessible disordered parts of proteins where the linear sequence is recognised. In this example a consensus Protein Kinase A motif is shown, which is defined by [RK][RK]x[ST][^P]xx. The residues that define the binding are highlighted in green.

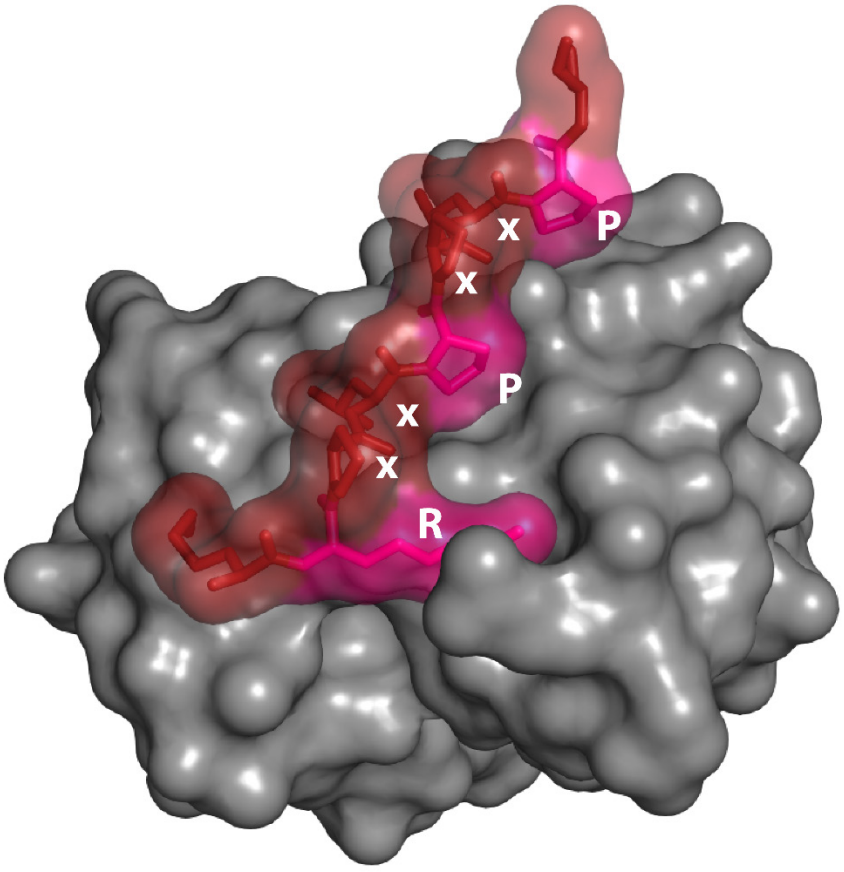


Figure 1.5. SH3 domain from PI3Kinase biding its sequence motif RxxPxxP in a small peptide.The prolines are aligned with pockets on the binding surface such that the 2-residue spacing positions them in the most optimal orientation to contact the SH3 domain with the arginine making an electrostatic interaction with glutamate and aspartate in its pocket. The interaction is further stabilised through backbone hydrogen bonds within the non-specific positions of the peptide. PDB: 3I5R

3. Degradation motifs/Degrons: Degrons are short sequences that are recognised by E3 ligases for degradations through the proteasomal pathway. Degradation motifs are key in regulating cell integrity and in timing many of the important cell cycle checkpoints through degradation via APC/C among others (Barford 2011). They play a role in protein quality control where it is hypothesised that buried or cryptic degrons made accessible through misfolding ensure degradation of these faulty proteins (Furth *et al.* 2011). Degrons also play a central role in signalling cascades and information processing through regulating the abundances of various signalling proteins, thereby allowing the cell to return to homeostasis (Grady *et al.* 1997). Well defined degrons include KEN- and D-box motifs.

4. Ligand interaction: Many scaffolding domains such as SH2, SH3 and PDZ recognise short motifs in their interaction partners. These domains are found in a vast range of proteins to form large multiprotein complexes and to regulate many signalling pathways (Harris & Lim 2001; Schlessinger 1994). These interactions form the backbone of many of the central signalling pathways and regulatory systems in every cell. Examples include key SH3 motifs that form interactions to mediate the ubiquitous receptor tyrosine kinase cascade and PDZ motifs in the C-terminal tails of GPCRs (Bockaert *et al.* 2003; Pawson *et al.* 1993).

5. Chemical Modification/PTM: A central aspect of cell regulation and complexity is reversible chemical modifications such as phosphorylation sites, glycosylation sites, methylation, acetylation, SUMOylation and ubiquitylation (Duan & Walther 2015; Gajadhar & White 2014). These are all defined by short recognition motifs that determine which residue gets modified by the active enzyme. These modifications act in tandem with all the other types of motif mediated interactions to regulate the accessibility and chemical properties of other parts of the protein.

All these different motif-mediated interactions take up very little space in the sequence compared to larger globular domains and other types of interactions. Their size and simple biochemical definition allows a whole range of different motifs to act in tandem in relatively short sequences and allow for a complex layering of regulatory and functional inputs. In the cellular context, disordered regions with motifs can be information dense and have the ability to process and regulate many different sources of information input as exemplified in the histone code discussed in section 1.3.

Given the small number of amino acids required to define a functional motif, many proteins (and any random peptide sequences) are likely to contain several motifs by chance. For example, an SH3 binding motif of PXXP is likely to occur once in every ~400 residues just by chance (ignoring any amino acid composition bias). To fully understand and categorise motif functionality in the cell it is therefore important to integrate several other factors required for *in vivo* motif function. A motif sequence pattern has to fulfil several key properties to be functional in the cell. It has to be accessible in its cellular

context, its interaction partner must be present when it is accessible, and the consequence of the interaction has to have an effect on the functional context of the proteins involved (Gibson *et al.* 2015; Via *et al.* 2009). This means that cellular localisation, expression timing, environmental properties and other present interaction partners are all important factors (Figure 1.6). Another critical consideration is the functional impact of the event; a phosphorylation that makes no functional difference can not necessarily be said to be a functional motif despite being physically phosphorylated by a kinase (Landry *et al.* 2013; Levy *et al.* 2009).

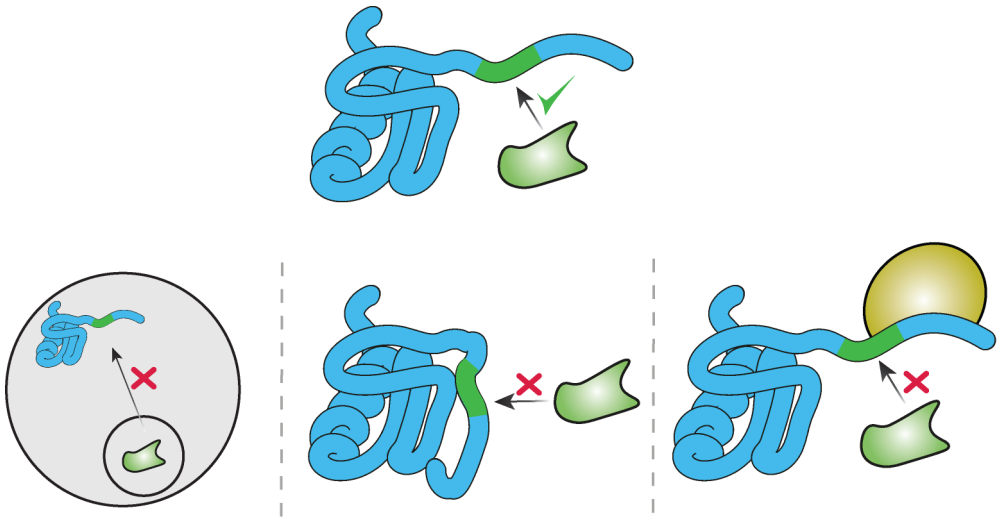


Figure 1.6. Functions of motifs are dependent on context. Having the motif consensus sequence is not the only requirement for a functional motif. The motif needs to be accessible within the structure, not blocked by other interactions and must be present at the same time and in the same region as the interaction partner.

This difficulty in defining and characterising functional motifs *in vivo* is one of the main limitations in our current ability to understand the full extent of motif mediated interactions in different parts of cell biology (Gibson *et al.* 2015). A combination of computational systems approaches and high throughput interaction studies alongside more painstaking biochemical characterisation of detailed pathways and regulatory mechanisms are all employed currently to try to elucidate the impact of motifs on these mechanisms. It has been estimated there are a million motifs in the human proteome, highlighting the importance of increasing our effort to study them (Tompa *et al.* 2014).

1.5. SLiMs in cellular information processing, decision making and regulation

Disordered regions with motifs are significantly associated with complexity through their central part in regulation and ability to respond to various cell stimuli and contextual cues (Van Roey *et al.* 2012, 2014). Disordered regions are also associated with interaction nodes and hubs and form some of the

key regulatory checkpoints in many networks (Haynes *et al.* 2006). Here I will use a simplified example to illustrate how motifs and protein disorder act together to form complex regulation that enables cellular decision making and information processing.

Considering a hypothetical cellular DNA damage monitoring system, we can envision three cellular decision making scenarios: No DNA damage, in which case the decision is to do nothing; minor DNA damage, in which case the decision is to activate repair; and extensive DNA damage, in which case the decision is to activate cell death (Figure 1.7). All of these are molecular decisions based on conditional information input states. To be able to make these decisions, molecules and cells thus have to process information about different conditions and contexts within the cell. In this way they are highly similar to other computational processes.

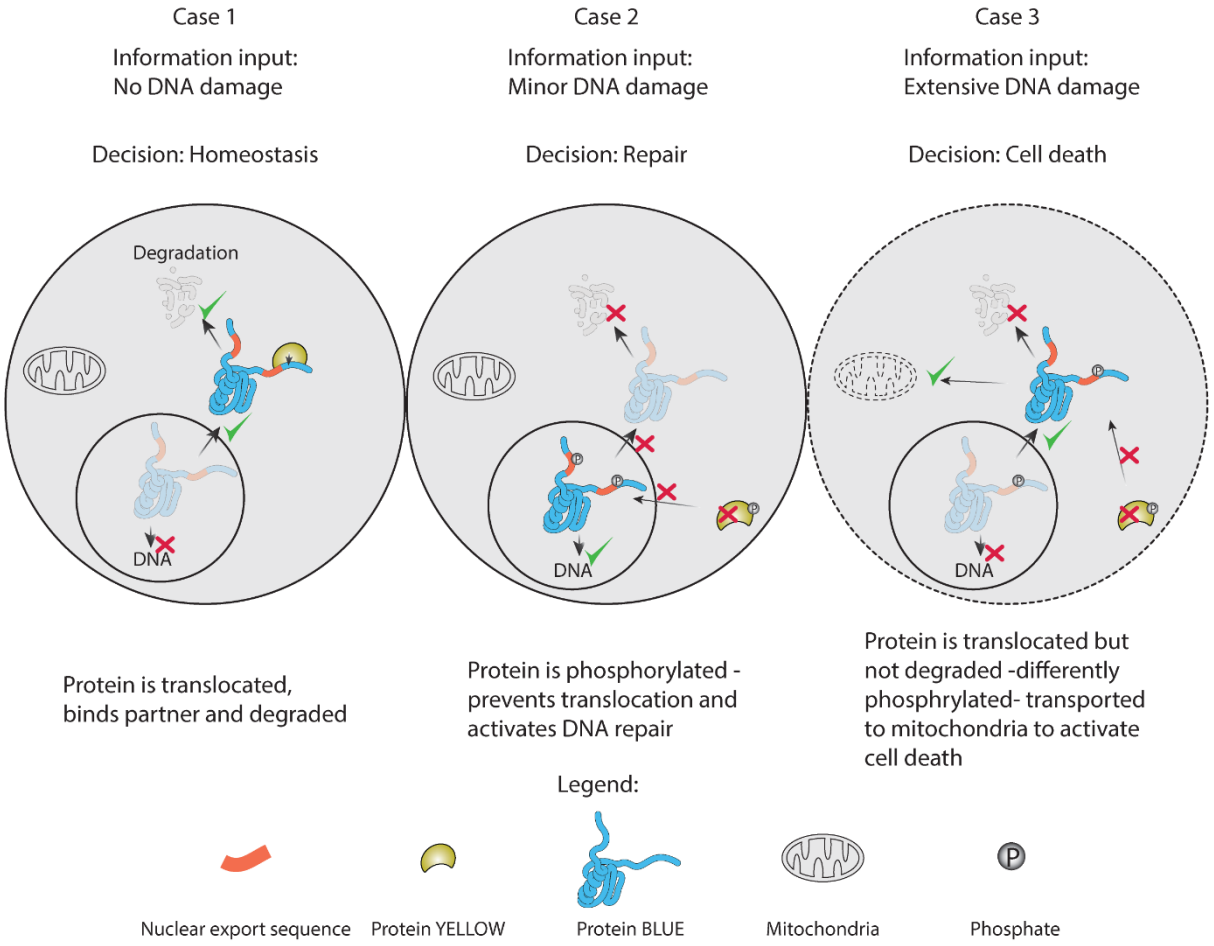


Figure 1.7. Cellular decision making through motifs. This example highlights three decision outcomes in a DNA damage repair system in a hypothetical cell. In case 1 protein BLUE is degraded and has no effect. In case 2 phosphorylation leads to nuclear accumulation of BLUE which results in DNA repair. In case 3 the damage is too extensive, and differential phosphorylation activates an alternative apoptosis pathway. This example is inspired by the p53-MDM2 DNA repair pathway.

In case 1 there is no DNA damage. Protein BLUE is not phosphorylated and the nuclear export sites are accessible causing the protein to localise to the cytoplasm. In the cytoplasm it binds protein YELLOW which is an E3 ligase that targets BLUE for degradation. YELLOW recognises a degron on BLUE that overlaps with a phosphorylation site and the nuclear export signal (Figure 1.7, case 1).

In case 2 there is minor DNA damage, which leads to specific kinases being activated. The kinases phosphorylate both YELLOW and BLUE, inactivating and sequestering YELLOW. BLUE is phosphorylated near both export sites preventing export and degradation. Increased concentration in the nucleus allows BLUE to activate genes involved in the DNA repair machinery.

In case 3 extensive DNA damage activates different kinases and phosphatases leading to a different phosphorylation pattern. One NES site is exposed which allows BLUE to be targeted to the cytoplasm. YELLOW is still phosphorylated and thus inactivated, and the binding site for YELLOW on BLUE is phosphorylated preventing the degron interaction. The abundance of BLUE increases in the cytoplasm, where it is targeted to the mitochondria. There it interacts with the apoptotic machinery through protease sites and motif based protein-protein interactions, which ultimately leads to cell death. In this simplified example, motifs are central to the decision making process through binding sites, degrons, cleavage sites, localisation motifs and PTMs.

This example is a simplified version of the decision making in DNA repair by p53 (BLUE) and MDM2 (YELLOW) in response to DNA damage. p53 is a protein whose fundamental job is to maintain genome integrity. It has been called the guardian of the genome and is crucial in DNA damage repair (Efeyan & Serrano 2007).

p53 activation in itself is a decision largely mediated through motif based interactions and modifications (Van Roey *et al.* 2014). Under normal cell conditions MDM2 continuously binds p53 through a motif-mediated interaction (degron) near the N-terminus of p53 (Brooks & Gu 2011). This binding has two effects, it blocks p53 transactivation functionality by physically blocking the transactivation domain and it also degrades p53 through ubiquitination. The “decision” to activate p53 in response to DNA damage is a result of p53 phosphorylation by DNA-dependent protein kinase (and stress related kinase, PIKK) which recognises and phosphorylates Ser15 (Canman *et al.* 1998). This phosphorylation triggers further nearby phosphorylation of Thr18 through Casein Kinase 1 which recognises pSer, and as a result the phosphorylated N-terminus is not recognised by MDM2. Ultimately, this prevents p53 degradation and causes an increase in P53 concentration and simultaneously enables p53 transactivation to occur (Sakaguchi *et al.* 2000; Shan *et al.* 2012).

p53 localisation is also influenced by motifs, in tandem with MDM2 binding and phosphorylation regulation. p53 contains 3 nuclear localisation motifs and 2 nuclear export sequences (Shaulsky *et al.* 1990; Stommel *et al.* 1999; Zhang & Xiong 2001). Nuclear localisation is important for its transcriptional activity but p53 has also been found to perform many cytoplasmic functions (Green & Kroemer

2009). The final localisation is decided by interactions between the localisation motifs and several PTMs. Upon sensing DNA damage, p53 is activated by phosphorylation which blocks both MDM2 binding and the nuclear export site preventing p53 nuclear export. MDM2 is also phosphorylated to reduce its ability to bind and ubiquitinate p53 (Cheng *et al.* 2011). In addition, subsequent multimerisation of p53 also obscures nuclear export motifs, further preventing export of activated p53 (Fischer *et al.* 2016). However, p53 translocation to the mitochondria is an important apoptosis inducing pathway that is partially independent of p53 transcriptional function. Translocation to the mitochondria has been shown to be mediated through monoubiquitination modifications (Vaseva & Moll 2009).

The decision making in p53 is highly complex and still not fully understood. What I have outlined here is a simplified description of some of the decisions made in this system, to highlight the role of motifs. In reality, a whole range of other motifs are also involved, including many more interaction sites and PTMs which regulate multimerisation, localisation, binding interactions and transcriptional activity, all of which together contribute to the final decision outcomes made within this pathway.

The many roles and sometimes even contradictory functions of p53 illustrate the importance of encoding multiple motifs in the same protein, which allows logical interactions with a range of other cellular proteins to assess the condition and execute the final outcome. p53 is one of the proteins we have studied in most detail, and we still have only a limited understanding of the levels of regulation and the motifs involved. Characterising the importance of motifs in cellular pathways is still in its very early stages, and many proteins are likely to be regulated by a range of different conditional motifs, similarly to p53.

This also highlights how important it is to understand the properties and dynamics of motif evolution. Through motifs, complex functionality can evolve quickly, and drastic regulatory changes can be achieved. Host systems can also quickly be hijacked by pathogens through these mechanisms. These motif aspects will be covered in the next sections.

1.6. Motif evolution

Motifs are key functional units in biology because of the interaction between a binding pocket on a domain and the general functionality that is required in many different instances, which results in a highly modular system. Modularity at this scale is important in allowing evolutionary innovation and restructuring of protein function and regulation (Tompa *et al.* 2014). The emergence of domains and motifs was essential for the evolution of modularity which is thought to have evolved as a means to generalise function and make the organisation of the proteome and networks more efficient, and most importantly as a way of enhancing evolvability and adaptability (Bhattacharyya *et al.* 2006; Hintze & Adami 2008; Jeff *et al.* 2013). Domains and motifs together act as basic functional building blocks

that can generate new function and allow adaptation to changing conditions through evolution. Domain-motif interactions are thus important in evolving modularity and evolving complex cell networks, but the extent and impact of motif evolution in these systems is still not well understood (Kim *et al.* 2014).

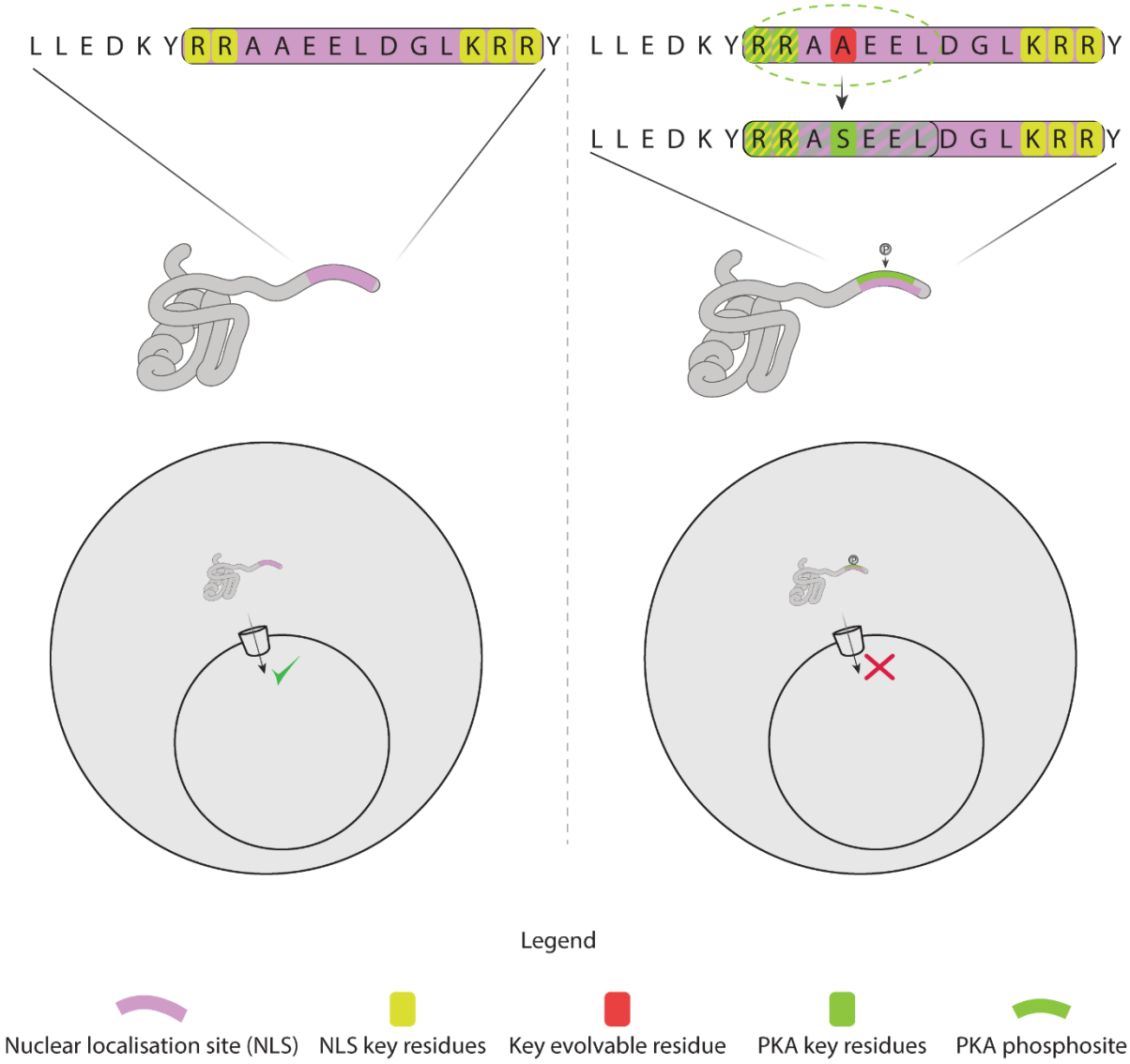


Figure 1.8. Evolution of a new phosphorylation motif. This is a general example of how a motif can evolve through a single amino-acid substitution (in this case A to S) which creates a new phosphorylation site. In this example the new site overlaps with an already existing nuclear localisation site, and the newly phosphorylated motif blocks nuclear localisation. This highlights the functional impact motif evolution can have in complex regulation.

The evolutionary properties of motifs are different from other interaction mediators. Their short length in linear sequence space, alongside their sequence degeneracy – allowing several amino acids with similar biophysical properties in defined positions – allow them to evolve quickly through a small number of mutations (Figure 1.8). This has been described as *ex nihilo* ("from nothing") motif birth

(Davey *et al.* 2015; Neduva & Russell 2005). These motif properties work in tandem with the properties of IDRs which have fewer evolutionary constraints on their sequences (Brown *et al.* 2010). This results in more rapid sampling of sequence space in disordered regions, and thereby sampling of many different motifs that selection can then act upon. Motif evolution has only been studied to a very limited extent and our understanding of the evolutionary properties and the biological significance of motif evolution is still in its infancy.

In a study on PDZ evolution in vertebrates the authors were able to track the emergence of *de novo* PDZ motifs in a range of proteins important for vertebrate neuronal development (Kim *et al.* 2012). They showed that vertebrate neuronal evolution has been drastically impacted by the evolution of these C-terminal motifs through point mutations, and that they had rewired many protein networks in neurons through acquiring new PDZ motifs. In their PDZ interaction network, around 1/3 of all PDZ motifs had evolved *ex nihilo* through random mutations. Motif evolution has been observed in important yeast pathways as well, where regulation of nuclear shuttling through the emergence of several Cdk1 phosphorylation sites evolved to regulate MCM3 in response to the cell cycle (*cf.* example in Figure 1.8)(Moses *et al.* 2007). In a study exploring the evolution of phosphorylation sites in the regulation of cell cycle progression it was similarly found that the high mutation rate in the disordered regions lead to the disappearance and emergence of phosphorylation motifs at a high frequency and that new phosphorylation sites evolve *ex nihilo* (Holt *et al.* 2009). They also found that conserved phosphorylation sites are not necessarily sequence conserved but may disappear and reappear such that they are conserved in the right IDR of the protein but move around on an evolutionary timescale.

Taken together, this suggests that motif evolution to a large extent happens through random mutations and *ex nihilo* emergence, giving motifs unique evolutionary properties in terms of the dynamic appearance and disappearance in proteins during evolution. This also has implications for how we interpret conserved functionality from a bioinformatics perspective since motif function can be conserved, despite motifs moving around in sequence space. The impact of motif evolution on functional organismal evolution and the importance of motif mediated functions in complex regulation in cells, together suggest that motif evolution has a high potential for functional innovation and rapid organismal adaptation. This implies that few sequence differences between closely related organisms can yield significant functional changes in information processing and regulation. This could easily be overlooked in analyses of closely related organisms as they would appear to function similarly based on the sequence and structural homology of the proteins involved. A detailed understanding of the impact of this is lacking, however a recent study explored the PTMs in two closely related (95% amino acid identity) bacterial species in two different ecological contexts (Li *et al.* 2014b). They identified a large shift in PTM patterns between the two species, hinting at the importance of this property of motifs in evolution. This has implications for the way we infer function through studies on related model organisms.

A better understanding of the evolutionary dynamics of motifs and their role in rewiring protein interaction networks and changing regulation will be essential to better understand what drives adaptation, speciation and other evolutionary events.

1.7. Pathogen hijacking of host systems through motif mimicry

The ease with which motifs can evolve through random mutations suggests they could be an important factor during infection and for pathogen function and evolution. Pathogens tend to have high mutation rates and need to respond on a relatively short evolutionary timeframe to changing conditions (Denamur & Matic 2006; Sanjuán *et al.* 2010). It has been established that a large number of pathogens that infect eukaryotes have significant amounts of protein disorder (Pancsa & Tompa 2012; Pushker *et al.* 2013). Alongside this, recent studies have begun identifying host-like motifs that are used by pathogens to interact with the host machinery for the benefit of the pathogen. These motifs have been found in wide-ranging pathogens including bacteria, eukaryotic pathogens and viruses (Chemes *et al.* 2015; Hagai *et al.* 2014; Van Roey *et al.* 2014; Via *et al.* 2015). These organisms rely on several mechanisms to subvert their host, but all use variations of similar methods which include promoting synthesis of pathogen proteins and genetic material, promoting an environment of growth, subverting host processes, redirecting cellular resources and subverting anti-pathogen host pathways (Alto & Orth 2012; Bruggeman 2007). To do this, they need a combination of first order and higher order means of attack which means they often encode both pathogen specific proteins that carry out a range of enzymatic activities, and also interfere with interaction networks and regulation to rewire host pathways (see Figure 1.9).

To illustrate common ways pathogens use motifs to interact with and rewire host systems I will use some well characterised examples from the literature:

1. Localisation motifs: Many pathogens have evolved localisation motifs to get proteins to the right compartments in cells. The most common motifs include nuclear localisation sequences, recognised by importins, utilised for example by several influenza A proteins and in SV40 which targets viral proteins to the nucleus (Fontes *et al.* 2003; Tarendeau *et al.* 2007). Importin binding motifs are also present in non-viral pathogens such as *Toxoplasma gondii* (Ahn *et al.* 2007). ER targeting and sorting motifs are also common in many coat proteins to get the proteins to the cell surface for viral particle assembly and export which we see in influenza and adenovirus (binding COP1 among other proteins) (Nilsson *et al.* 1989; Sun *et al.* 2013).
2. Interaction with ligand binding domains: PDZ and SH3 binding motifs are highly prevalent in many pathogens as both motifs are used by proteins involved in the host immune response

pathway. PDZ motifs have been evolved in several viral and non-viral pathogens including influenza A and HPV where it interacts with Scribble, which is involved in preventing apoptosis (Thomas *et al.* 2011; Zhang *et al.* 2007). A PDZ motif is also present in enteropathogenic *E. coli* where it is involved in binding and modulating scaffolding proteins associated with actin based intracellular transport (Martinez *et al.* 2010). SH3 binding motifs are similarly present in *Listeria*, *M. tuberculosis* and *Plasmodium falciparum* as well as viruses such as influenza A and HIV where the SH3 mediated interactions perform a range of roles (Arold *et al.* 1997; Rajabian *et al.* 2009; Ravi Chandra *et al.* 2004; Shin *et al.* 2007b). SH3 can mediate important interaction with PI3K (influenza A) which is central in the cellular immune response, FYN kinase (HIV) which alters T-cell signalling and c-Src kinases (*Plasmodium*) (Akhouri *et al.* 2008; Arold *et al.* 1997; Shin *et al.* 2007b).

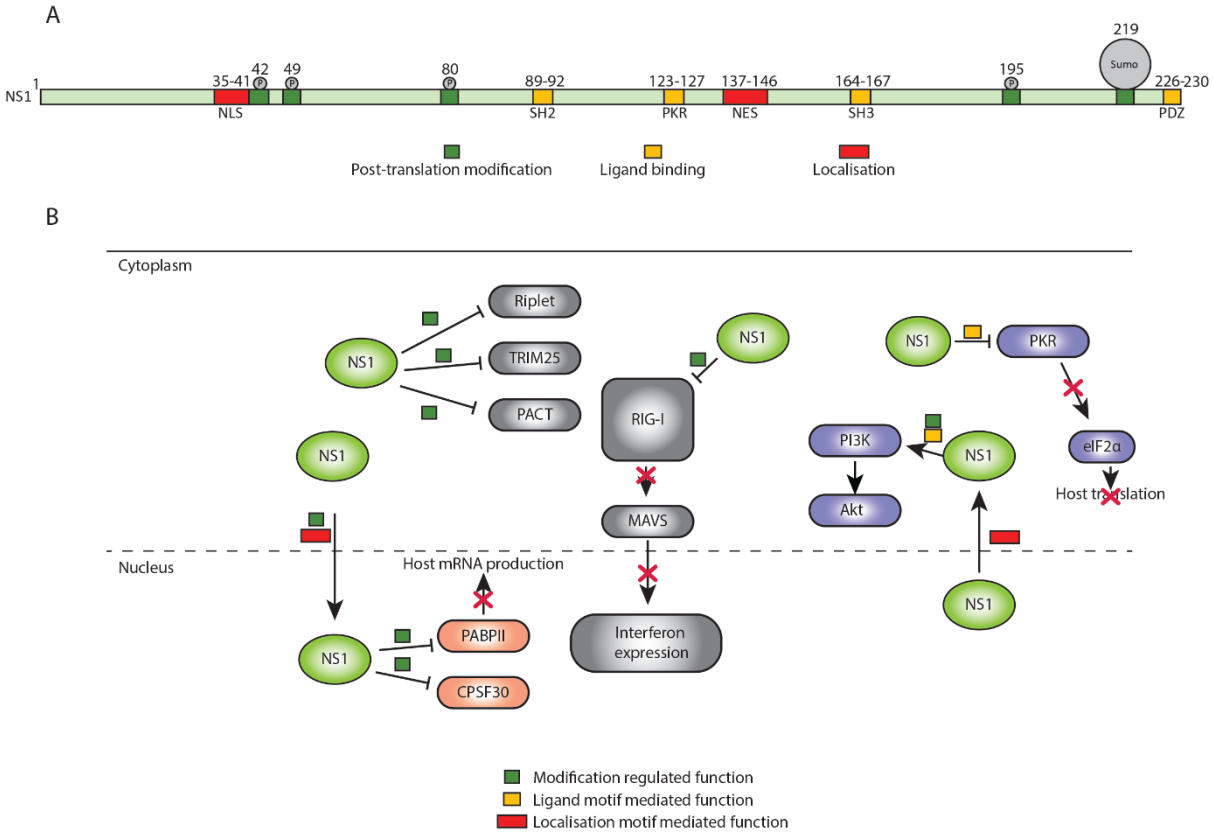


Figure 1.9. Viral subversion of host systems involve many motifs. A) Influenza A NS1 protein with several important motifs highlighted. B) NS1 subverts host cells through many different interactions and pathways. NS1 interacts with many proteins in the interferon signalling response to prevent interferon activation (grey proteins). These interactions are mediated by the dsRNA binding domain and regulated through several key phosphorylation and sumoylation sites. NS1 is also transported between the cytoplasm and nucleus. In the nucleus it reduces host mRNA expression through binding e.g. PABPII and CPSF30 (orange proteins). NS1 also plays an important role in activating PI3K/Akt signalling and in blocking translation through binding PKR through a motif. Figure inspired by (Klemm *et al.* 2018; Nogales *et al.* 2018).

3. Modification sites: MAPK recognition motifs (as well as docking sites) are present in pathogens such as *S. flexneri* and *T. gondii* and *Salmonella* where they modulate the innate immune response pathway and are key in the infection. Phoshorylation sites and glycosylation sites are plentiful in many coated viruses including influenza and HIV (Hutchinson *et al.* 2012;

Vigerust & Shepherd 2007). Glycosylation is important for coat function and recognition by the adaptive immune system and phosphorylation is used at all stages of infection to regulate transport, binding and other activities.

4. Degrons: Protein degradation is important in the lifecycle of many pathogens, but particularly for viruses. In hantaviruses, C-terminal degrons have been linked to their pathogenicity, being a defining difference between pathogenic and non-pathogenic variants (Sen *et al.* 2007). The cowpox is another example of a virus that relies on extensive protein degradation through the host proteasome machinery, predominantly for viral uncoating but also in other functional contexts (Grossegesse *et al.* 2018).

5. Cleavage sites: Host-like cleavage sites are common in many pathogen proteins. Influenza A relies on cleavage sites in the coat HA protein to enable cell entry (Bertram *et al.* 2010). It has also been suggested that late stage cleavage of influenza protein NP is crucial for pathogenicity (Zhirnov & Syrtzev 2009). A common strategy for viruses to subvert host apoptosis responses involve the use of viral proteins that are specifically targeted by apoptotic proteases (Richard & Tulasne 2012). This approach is prevalent in baculoviruses and HPV and allows them to prevent the cells defence response.

This pervasive use of host-like motifs in pathogens is a result of both the adaptability and evolvability mediated through motifs and the importance of motifs in regulating networks and the information flow within cells. This highlights the need to understand both the motifs themselves as well as their evolutionary properties better, to gain a better appreciation for the potential changes that can occur through simple motif gain and loss in pathogens.

1.8. Discussion and thesis objectives

In this introduction chapter I have outlined how motifs play a central role in cellular decision making and regulation, and have unique evolutionary properties compared to other functional protein features. Motifs fall in a really crucial interesting space and through studying their evolution we can start understanding how their evolutionary dynamics impact changes to cellular information processing.

The idea of the cell as an information processing machine is one that has been discussed previously and a model I find both fascinating and useful for thinking about the processes that occur within the cell. Motifs, more than any other aspect of cell biology, behave much like binary information states of chemical modifications and interactions that work together to create complex logic and decision making systems. Motifs make a compelling system for creating logic gates in cells through the use of layered and conditional motifs within proteins, as these are truly fundamental to the signalling and thus

decision making machinery (Fogelmark *et al.* 2016; Langan *et al.* 2019; Singh 2014; Van Roey *et al.* 2012). These fundamental properties are also what allows motifs to confer complexity to organisms.

My goals going into this PhD were to evaluate how evolution of motifs can act and has acted to rewire pathways in organisms. The idea that simple point mutations to motifs can drastically change the overall function of cells and organisms without large changes to the sequence is a very exciting and important one evolutionarily. I also think it is likely to have been much overlooked and underappreciated in biology as we have been lacking the ability to investigate these questions with conventional biochemical or bioinformatical approaches until recently. I further wanted to determine whether it is possible to better understand and even predict where motif evolution might happen and if there are markers at the sequence level that can help us better understand and characterise motif evolutionary dynamics.

Many of the viruses that we have the most difficulty treating and preventing such as influenza A and HIV have disproportionate amounts of protein disorder and motif mediated functionality. How the evolutionary dynamics of motifs contribute to these properties have not yet been investigated. To date very little is known about motif evolutionary dynamics in general. As has been discussed, motif emergence through random mutations is prevalent in nature but the extent to which motifs emerge, how frequently and in which contexts are unknown. It is likely there could be selective benefits in virus proteins that can sample motif space of relevant motifs, and also avoid losing important functional motifs. These costs and benefits to organisms of evolving new motifs and losing existing ones are of interest, but have not been addressed previously.

This thesis is my take on several of the questions regarding the nature and consequences of motif evolution. The overall outline of this thesis is as follows:

In Chapter 2 I first establish a framework for how we think about motif evolution, in particular in viruses and other rapidly evolving systems. The focus of this framework is the integration of mechanistic evolutionary models and the emergent properties on protein evolution resulting from the nucleotide and codon space. This framework thus integrates methods and ideas from several different fields of molecular and evolutionary biology. I go on to simulate instances of motif evolution in simple sequences and populations given certain mutation and evolutionary parameters to get a model for what to expect from motif evolution in general. Finally, I describe a computational pipeline that can be used to characterise motif evolution in RNA viruses given historical sequence data.

In Chapter 3 I apply that framework and methodology to evaluate motif evolution in influenza A. I evaluate whether the prior sequence, and in particular prior codon choice, impacts the evolutionary trajectory and evolutionary dynamics of motifs. I also investigate whether the evolutionary consequences of codon choice lead to codon bias in motifs and if it is possible to predict the functionality of putative motifs through inferred selection.

In Chapter 4 I investigate how new motifs get sampled and evolve in influenza A and how that relates to prior sequence and codon choices. I then explore the landscape of recently evolved motifs in influenza, important for the disease biology of the virus. Finally, I predict hotspots in the genomes of influenza strains where new motifs have the potential of evolving in the future and impact the virus biology and its interaction with humans.

In Chapter 5 I summarise and discuss the main findings and draw conclusions and delineate the most interesting questions in the field currently.

Chapter 2

A conceptual framework and computational resource for the analysis of motif evolution

2.1. Overview

Evolutionary forces can at the most basic level be divided into the generators of variation and the forces that determine which variations increase and decrease in prevalence. Genetic variation is generated in different ways in different organisms, and frequent sources of variation include single nucleotide mutations and genome rearrangements. For many viruses, in particular influenza A – which will be used as a model for evolution through mutations in this chapter – sequence variation is generated almost exclusively through point-mutations (Mehle *et al.* 2012; Shao *et al.* 2017). It has been observed that a large number of new motifs evolve through point mutations making this a good model for understanding motif innovation through mutations (Davey *et al.* 2015; Van Roey *et al.* 2012).

The second aspect of evolution is determining which variants become more prevalent and which are removed from the population. This is predominantly driven by the balance between selection and genetic drift which are closely linked to the population sizes and frequencies of different sequence variants (Bustamante 2005; Kimura & Ota 1974; Lande 1976). The fitness of any mutation-induced change alongside the relative frequency of that change within the viral population thus influences which mutational variants get fixed during an infection and further spread the virus (Moya *et al.* 2000).

To understand motif evolution, we thus first need to understand the fundamental interplay between the generation of variation and the factors that impact fixation including population frequencies, drift and selection. In this chapter I aim to define a way to think about motif evolution that captures the realities of the underlying driving forces that cause mutations and also the emergent properties through the organisation of the genetic code.

2.2. Background

This background acts as a short review covering relevant topics for the different aspects and properties of evolution of short linear motifs in polypeptides, to explore and define what is required to understand the details of motif evolution.

2.2.1. The mechanisms of mutation

Mutations are fundamental to evolution as a source for genetic variation that selection can act on. In general, mutations are caused either by external factors or replication errors (Ganai & Johansson 2016; Kodym & Afza 2003). External factors include: chemical compounds, which modify the bases changing the hydrogen bonding pattern during replication and introduce non-complementary bases in the newly replicated strand; radiation, which can break bonds in the nucleotides causing similar changes during replication; metals, which can both increase the presence of chemical compounds that alter genetic material, or interact with polymerases and change the accuracy of replication; and intercalators and cross-linking agents, which interfere with polymerase binding during replication. Polymerases also have natural replication error rates where a nucleotide is matched to a non-complementary nucleotide spontaneously, causing a mutation. This can be due to base-tautomerization which alters the pattern of hydrogen bonding (which is essential for the selectivity during replication) (Li *et al.* 2014a; Topal & Fresco 1976).

Since RNA viruses do not have any error correction mechanisms all these changes to new sequences will remain and propagate provided the sequence itself is viable (Boivin *et al.* 2010). These accumulating changes give rise to a pool of sequences with many variations during a viral infection that can both allow subpopulations within an organism or cell to perform slightly different functions, or reduce the overall number of functional sequences and thereby reduce the overall infectious fitness of the population (Domingo *et al.* 1985).

Importantly, these sources of mutations affect the nucleotides in different ways. Polymerases also have unequal rates of random mutations for the different bases. Overall this leads to large disparity in the mutation rates between the different nucleotides such that each “source nucleotide” to each different “target nucleotide” can have different rates (e.g. A-to-G compared with A-to-C and so on) (Pauly *et al.* 2017; Sanjuán *et al.* 2010).

A knowledge of the nucleotide specific mutation rates will be important to understand the likely frequencies of particular changes expected in viral evolution, which will be important when considering which point mutations cause motif formation at different rates.

2.2.2. Observed substitution frequencies: estimated mutation rates or "true" mutation rates?

To get to a point where we can accurately understand and characterise evolutionary dynamics we need to understand how frequently mutations are likely to happen. Most current methods use observed substitution rates to determine evolutionary dynamics over time (Nachman & Crowell 2000). Mutation rates describe the rate at which nucleotides change as a result of replication errors, while substitution rates describe the observed nucleotide changes in subsequent surviving linages over time. Mutation rates can be estimated from substitution data and knowing the time that separates sequences, however these estimates tend to drastically underestimate the actual mutation rate (Yang 1994; Yang & Nielsen 2000). To get the most accurate results we need to know the true mutation rate as closely as possible, as this will inform the actual rate at which different genotypes will emerge through mutation. The "true" mutation rate can be experimentally determined through measuring mutations by specific polymerases in active replication in what is known as a fluctuation test (Pauly *et al.* 2017; Pope *et al.* 2008).

Commonly both mutation and substitution rates are estimated from sequence alignments and phylogenetic trees and can thereby inform nucleotide, codon or amino acid substitution matrices where the probability of substituting a given element for another over a given time is defined (Henikoff & Henikoff 1992; Oscamou *et al.* 2008). Substitution matrices – in particular for codons and amino acids – are defined based on structural context and other protein features and can vary depending on a range of factors. For example, folded proteins have different matrices to disordered proteins since there are more restrictions on which amino acids in the core of the protein that can be tolerated (Goonesekere & Lee 2008). These matrices reflect that contextual information to some extent, as they are based on observed substitutions. Nucleotide substitutions are frequently more accurate reflections of the relative rates of the mutations, since silent mutations and mutations in non-coding regions act as mostly unbiased sources of variation. However, many factors can still skew those estimates since codon choice and codon bias, as well as RNA features and interactions with other molecules, all act to limit the tolerated mutations in many regions of the genome (Oscamou *et al.* 2008; Weatheritt & Babu 2013).

The most useful data for understanding evolutionary outcomes will therefore be experimentally determined mutation rates for all 12 nucleotide mutation types. This is generally rare to have for organisms and is also subject to many limitations, which is why in the past estimates based on substitutions have been used so predominantly. Many of the techniques used to determine mutation rates rely on sequencing which introduces biases from the errors during the PCR and sequencing steps (Foster 2006).

Recently, more accurate techniques have been developed that rely on direct readout of GFP fluorescence in molecules that require specific mutations to make functional GFP (Pauly *et al.* 2017). The

rate of these mutations can be measured as the rate at which GFP regains functionality through the target mutation, thereby circumventing the limits of PCR and sequencing. Having accurately measured nucleotide specific mutation rates is an important stepping stone that opens up many possibilities for elucidating features and consequences of viral evolution. Early evidence suggests that even closely related strains in influenza have significantly different mutation rates (Nobusawa & Sato 2006). This has consequences for their evolutionary trajectory and might be key information in better predicting evolutionary paths and in tackling the global health issues associated with influenza spread and outbreaks.

Thus, knowing the mutation rates is important for informing the frequencies of mutational events. These frequencies interplay closely with the population sizes and the likelihood of fixation of new genotype which will be discussed next.

2.2.3. The relationship between the drivers of sequence change (mutations) and the resulting evolutionary events

Mutations create the variation that selection can act on, but it is natural selection and genetic drift that determine which variants ultimately evolve. For new functional mutations to be fixed in a viral strain, and to potentially cause altered infectious properties for the virus, several factors play a role (Patwa & Wahl 2008). (1) The size of the population with the new phenotype compared to the effective population size of the strain during the infection. This is directly influenced by the mutation rate specifically, and also the replication rate of the new sequence (Wilke 2003). (2) Genetic drift which interacts with the population size to influence the outcome of new phenotypes through random events. The larger the population with the new phenotype is, both in absolute terms, and in relation to the rest of the population, the less susceptible to drift it will be. (3) The fitness of the new phenotype, which is the number of successful offspring left by an individual sequence in the case of viruses. A more successful sequence will generate more successful viruses that spread and grow rapidly. Fitness is impacted by drift and population size so that a less fit sequence that exists in high frequency might still be fixed in the population through drift, whereas a sequence of higher fitness that only exists in very small numbers potentially very rarely emerges and can easily be lost through drift (Olson-Manning *et al.* 2012).

2.2.4. Population variation and population fitness drives viral sequence heterogeneity: quasispecies

The relationship between mutation frequency and population sizes of different genotypes is particularly relevant in viral populations. Due to the extreme mutation rates in viruses, even within a single viral population within a single host or cell, the sequence diversity is very high (Holland *et al.* 1982; Russell *et al.* 2018). This diverse population of sequences in a viral infection is commonly referred to as a viral quasispecies (Domingo *et al.* 2012). The high viral mutation rate is counter-intuitive from a fitness perspective, as viral sequences are understood to be highly adapted and thus occupying fitness

optima in genotype space. Experiments suggest that the majority of substitutions have negative or neutral fitness effects (Sanjuán 2010). Despite this, viruses maintain high mutation rates. Experiments with altered polymerases with significantly reduced error rates in polio have shown that these strains actually have significantly lower fitness (Vignuzzi *et al.* 2006). It turns out that the high mutation rate increases fitness for the infection as a whole despite most mutations not having a positive fitness impact. Suggested explanations for this include molecular synergy through higher functional diversity in the many different cells infected, and higher rate of adaptation to adverse conditions. The fitness of the quasispecies population is thus higher than any individual homogeneous population would be. This is important for the evolution and sampling of new genotypes and protein features within the population. The mutation rate and the frequencies of different phenotypes emerging will impact the composition of the quasispecies and have a fitness outcome for the viral population and infection which will be important to understand (Moya *et al.* 2000).

2.2.5. The role of codons in sequence evolution: not all mutations are created equal

Mutations occur on the nucleotide level, however the phenotypes associated with those mutations are determined on the amino acid level in the case of motifs. This means that the combined effect of nucleotide mutations and their impact on amino acid changes are what ultimately determine the functional outcomes of mutations. Most amino acids are encoded by degenerate codons meaning several different codons encode the same amino acids. This causes mutations to have different effects depending on the codon that is mutated, which has implications for the composition and fitness of the quasispecies.

The codons used in viral sequences have in fact in many cases been shown to impact viral evolution and fitness. It is generally understood that viruses adjust their codon usage depending on hosts, often matching the dominant codon usage pattern of the new host as has been shown in influenza as viruses move from e.g. birds to humans (Belalov & Lukashev 2013). In part, codon usage is thought to shift to increase translation efficiency by using more optimal codons that are more readily translated by the ribosome due to the relative abundance of the host tRNAs (Sharp *et al.* 2010). In viruses this would lead to increased overall fitness by maximising the number of viral offspring per unit time. There are likely several other factors at play shaping the codon usage in viral populations. It has been shown that introducing hundreds of silent mutations into several viruses drastically attenuates the virus simply through a change in codon usage without necessarily a change in replication or translation efficiency (Fan *et al.* 2015; Le Nouën *et al.* 2019). A suggested explanation for this change is the shifted mutational landscape resulting from the new codons leading to a different composition within the quasispecies and possible lower mutational robustness during infection (Burch & Chao 2000).

Sequences can also be more prone to detrimental mutations depending on specific codon usage in some cases. In an experiment on Coxsackie B3 and influenza A viruses, leucine and serine codons

were changed with silent mutations to be encoded by codons that were a single point mutation away from one or several of the stop codons (Moratorio *et al.* 2017). This change resulted in a drastically attenuated virus as the frequency of non-sense mutations increased. This led to a virus less robust to mutations, where the specific mutation frequency and codon choice led to an increase in detrimental mutational outcomes. As a result, the viral fitness was reduced.

Overall, understanding the way mutational frequencies interact with codon and amino acid spaces within the viral quasispecies will clearly be important to elucidate the evolutionary dynamics shaping motif use and evolution in viruses. However, to date, these questions have not been addressed.

2.2.6. Robustness: sensitivity to mutations and the viral sequence population

As mentioned above, mutational robustness is hypothesised to play a role in viral fitness and mutational outcomes. Mutational robustness is a property of sequences that counteracts high mutation rates by reducing the impact of individual mutations (Kucharavy *et al.* 2018). For example, a protein with high mutational robustness requires several mutations to accumulate before protein function is compromised (Bershtein *et al.* 2006; Tóth-Petróczy & Tawfik 2014). Robustness tends to be a feature resulting from stability, functional/sequence redundancy or in the case of networks and pathways, signalling redundancies (Fares 2015; Félix & Wagner 2006). Mutational robustness has been argued to play a role in viral fitness and evolution, and in simulations flat fitness landscapes (more robust) were shown to lead to higher fitness in viral populations (Montville *et al.* 2005; Wilke *et al.* 2001). Conventionally, robustness has been studied as a systemic property in proteins when looking at evolutionary dynamics and simulations. This approach largely overlooks the fundamental sequence features that actually yield the mutational robustness in the first place. How robustness factors in during the evolution of specific protein properties and functional features such as motifs – if at all – is largely unknown. It appears likely that mutational robustness would impact the quasispecies dynamics of viral populations through changing the way in which functional properties in the sub-populations emerge, which can have major evolutionary consequences.

2.2.7. Bringing all these factors to the table to understand motif evolution – what to expect?

It will be important to consider how mutation rates, codon choices, robustness and the quasispecies population all influence viral and motif evolution. Since motifs are short, linear sequences with direct functional phenotypes, mutations and selection can quickly act to evolve new motifs in certain contexts. In this chapter, I establish a theoretical framework for motif evolution that takes mutation rate, population frequencies, codon choice and robustness into consideration when trying to understand the evolutionary dynamics of motifs in viruses and their potential for functional innovation. This is a

novel approach to looking at the evolutionary dynamics of specific protein properties in general, and for motifs in particular.

Current approaches to modelling viral evolution, with respect to robustness, population frequencies and fitness effects on the infection as a whole, rely on models that reduce sequence features to a few parameters. Where these approaches fall short is in understanding the molecular details and features that contribute to the properties that are being modelled. An approach considering sequence features on the molecular level would increase our understanding of the way population dynamics are affected by mutations, selection and drift. This chapter is an attempt to model viral population dynamics with motifs in mind, since this approach is feasible with the molecular knowledge we currently have. By being able to model the details of concrete molecular structures in viruses such as motifs, it gives us an insight into the added benefits of including these detailed attributes. This will hopefully allow us to find ways to expand this type of modelling to more complex molecular features and incorporate that into ever more comprehensive evolutionary models.

2.3. A theoretical model for motif evolution through random mutations

Here I want to define a framework for motif evolution based on the fundamentals of mutation rates, population dynamics and the organisation of the genetic code that shapes the mutational landscape through codons and the functional properties of amino acids within the motif context.

Historically motif evolution has been exclusively approached from the amino acid level, in large part due to the datasets available and the fact that motifs are functionally defined only on the protein level. However, mutations act on nucleotides. Therefore, a more accurate understanding of evolutionary characteristics would be gained by defining and studying motif evolution through the lens of nucleotide mutations.

In the previous section I detailed what is currently understood about the relative fitness of different codons in certain contexts. It is generally well established that the codon space available to any given codon through mutations varies drastically and impacts fitness (Lauring *et al.* 2012). It has even been proposed that some parts of the codon "landscape" have been shaped through selection to increase functional overlap in codon space such as for leucine, isoleucine and valine for example (Koonin & Novozhilov 2009). This is another important aspect that will impact motif evolution as well, as I will elaborate on in this chapter.

As motifs are short and degenerate in sequence, they are orders of magnitude more likely to emerge in sequences *de novo* than any other functional molecular feature, and are likely to emerge with relatively

high frequency. This also means that population dynamics of viruses and cells will be a more important factor in the evolution of new motifs.

In this section I will expand on these three principles and develop a comprehensive framework for understanding the fundamentals of motif evolution in various biological contexts.

2.3.1. Amino acid tolerances at key positions in motifs can be used to define a neutral substitution space

For a new motif to evolve, a prior sequence missing at least one key amino acid needs to acquire the right mutations to substitute in to an amino acid that is functional at that position. The fewer mutations required for that to happen, the more likely the event is to take place. Any given codon has 9 single nucleotide substitution outcomes and 27 double nucleotide substitution outcomes (Figure 2.1). For a non-motif sequence to gain a motif, the higher the proportion of the target codons that are within the single substitution space compared to the double or triple nucleotide substitutions space the more likely the motif will be to emerge.

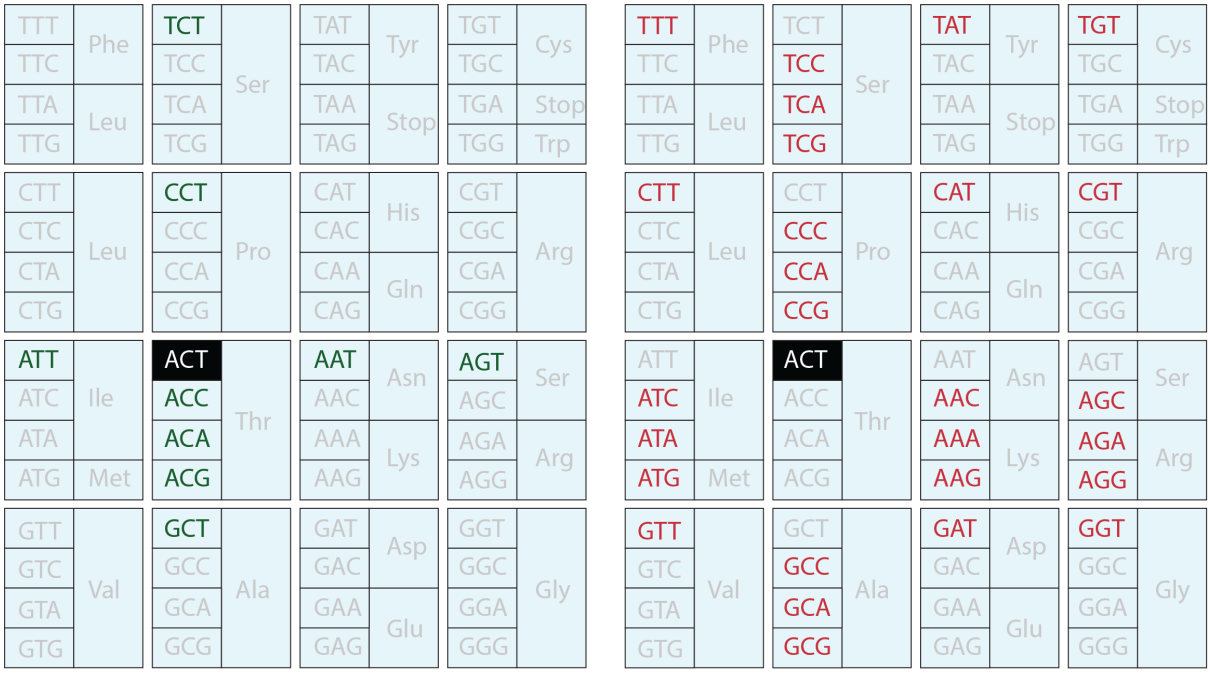


Figure 2.1. Codon mutation space for ACT. Left: ACT (Thr) has 9 possible single point mutation outcomes shown in green. Right: ACT has 27 additional double mutant outcomes shown here in red. The remaining codons can only be reached from ACT through triple nucleotide substitutions. ACT will therefore more frequently mutate into codons for e.g. Ser (single (2) or double (4)) than Val (double (1) or triple (3)) if all mutations are equally likely.

To elaborate on this point, the primary factor that impacts the frequency of specific amino acid substitutions at any given site, if not under selection, is the number of nucleotide mutations required, simply because the chance of a double nucleotide mutation happening within the same codon are several orders of magnitude lower than that of any single point mutation.

To determine if a sequence is likely or "primed" to evolve into a new motif, we have to define which amino acids are allowed in any given position to make a functional motif as this will determine the allowed codon space. As established in section 1.4, motifs are typically defined on the protein level through the right spacing and properties of amino acids in certain positions and usually described through a regular expression. In a typical Glycogen Synthase Kinase 3 (GSK3) recognition site the kinase recognises a serine or threonine followed by three spacer amino acids and then another serine or threonine. The motif is usually represented as a regular expression in the form: xxx[ST]xxx[ST], where x refers to any amino acid. From the allowed amino acids at each position we can define the full set of codons that correspond to the amino acids. In GSK3 that means 10 codons in each defined position (Ser/Thr), namely TCT, TCC, TCA, TCG, ACT, ACC, ACA, ACG, AGT and AGC (see Figure 2.1 for codon table). To determine if a given sequence is likely to evolve this motif, the question then becomes how likely each position is to evolve into one of the allowed codons. To illustrate this point, consider a nucleotide sequence that is -nnn nnn nnn CCC nnn nnn nnn AGA- encoding xxxPxxxR. This sequence is not near the motif on the amino acid level, having none of the required amino acids at the key positions initially. However, on the nucleotide level only a single nucleotide substitution per site is required to gain the motif. In addition to that, CCC (Pro) can either mutate to TCC (Ser) or ACC (Thr) giving 2/9 independent single substitutions that create the motif while for AGA (Arg) that number is 3/9. At the other end of the probability, a sequence like -nnn nnn nnn GAU nnn nnn nnn CGG- encoding xxxDxxxR is 4 nucleotide substitutions (2 per codon) away from forming a motif, making it an extremely unlikely occurrence (See Figure 2.2). In this way I can analyse any input sequence and determine whether it is primed to evolve a motif or if it has an extremely low probability of occurring.

The codon space will have a large impact on the prediction of motif evolution since a difference in allowed amino acids at a key site will alter which codons can easily substitute into the motif space. For example, the codons that are likely to substitute to Ser/Thr will be different from the ones for Ser/Thr/Cys, as the added cysteine will change the available codon space for many residues including glycine, arginine, tyrosine and phenylalanine (*cf.* codon table in Figure 2.2).

Motif definitions are to some extent related to the biophysical properties and similarities between amino acid residues, but there is significant variation here. Allowed amino acids at motif positions often differ from expected similarities since they are simply defined by the functional property of the motif interaction. In the Ser/Thr site for example, other similar amino acids such as asparagine are not allowed, because phosphorylation requires a phosphoacceptor. Thus, motifs are unique and highly suited to this approach of finding primed sequences because we are able to define the allowed residues in the functional space and because they are functional in linear sequence.

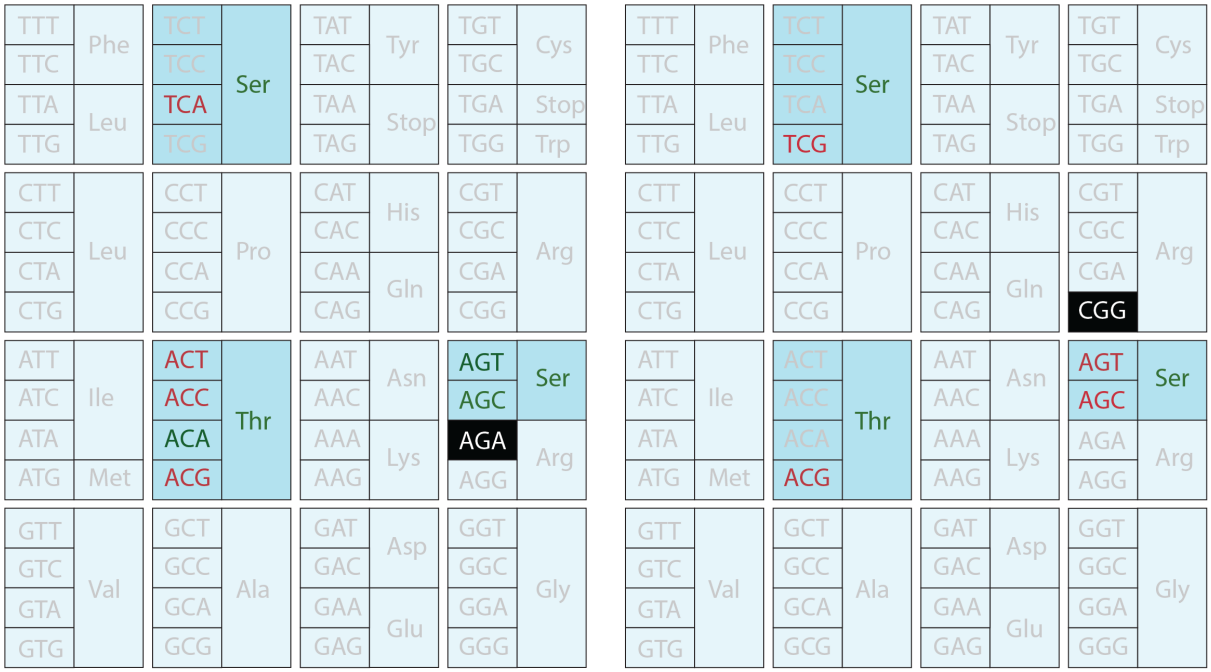


Figure 2.2. Different possible outcomes for two arginine codons. Given the codon space required to satisfy a Ser/Thr motif position, different initial codons will have a different likelihood of mutating into that codon space. On the left AGA is shown which can mutate into Ser/Thr through 3 different single substitutions and 4 double. On the right is CGG which has 0 possible single substitutions and 4 double. Given unbiased nucleotide mutation rates, AGA would be expected to mutate into Ser/Thr much more often.

2.3.2. Nucleotide specific mutation rates impact amino acid substitution probabilities in a codon centric model

Another variable that greatly influences the probability of a given codon to mutate into a codon within a motif-space is the nucleotide specific mutation rate. In this context, the nucleotide specific mutation rate is defined as the rate at which each nucleotide mutates into each of the other 3 possible nucleotides on average. In the example from the previous section (2.3.1) the rate of CCC (Pro) mutating to either TCC (Ser) or ACC (Thr) depends on the rate of C-to-T and C-to-A mutations specifically. There can be over a 100-fold difference between the mutation rates of certain nucleotide pairs. Having an accurate estimate of the different mutation rates would therefore greatly inform the evolutionary outcome of different initial sequences.

There are various ways to establish the nucleotide specific mutation rate. The relative rates can be estimated from a multiple sequence alignment and a phylogenetic tree. This approach to estimating mutation rates is susceptible to being influenced by selection pressures and other biases since many of the mutations included will be in coding regions. However, since the majority of mutations included will be either silent or in non-coding regions this still yields relatively accurate estimates of mutation rates

when a large enough dataset is used. Having these relative values can already improve the expectations for motif mutations, however even better would be to know the accurate mutation biases per replication in the absence of selection.

Fortunately, a recent experimental approach has been able to determine highly accurate mutation rates for viral sequences. In 2017, Pauly *et al*, developed a method that uses several sites in green fluorescent protein to accurately determine the mutation rate for all 12 possible mutation types in two influenza A strains. This method uses inactive GFP that will activate only if it acquires the mutation being measured, which allows direct readout of the mutation rate. This yielded a much higher mutation rate than had previously been reported. They found on average a rate of $1.8*10^{-4}$ mutations per nucleotide per replication, with as low as $5*10^{-6}$ for C-to-G and as high as $3*10^{-4}$ for A-to-G, highlighting the large range of mutation rates in influenza (see Table 2.5, in Methods 2.7).

These detailed mutation rates can be incorporated into the codon centric approach to motif evolution established in the previous section. The large variation in mutation rate will be a crucial component of the overall mutational outcomes in sequence space over evolutionary time, and over single genome replications. Consider the mutational outcome of CCC (Pro) to either TCC (Ser) or ACC (Thr); if the mutation rate C-to-T is $<10^{-6}$ and C-to-A is $>10^{-4}$ CCC would mutate into threonine >100 times for every single serine mutation that emerged. These differences will in large part change where we are likely to see motifs evolve as many sites will have big mutational differences alongside the available codon space (*cf.* outcomes in Figure 2.3).

To develop a model for motif evolvability in a given starting sequence I thus want to be able to quantify numerically the likelihood of all the possible mutations that would yield the right amino acids. With the understanding of the motif codon space from the previous section, along with the nucleotide specific mutation rate data, I can thus determine the likelihood of all the different outcomes that are within the target motif. With this I can get the expected probability, and thus the expected frequency of new motifs in a population which will be a key factor in the likelihood of the motif to evolve in an organism (Figure 2.3).

2.3.3. A probability model can assign probability values to motif evolution both for evolvability and evolutionary robustness

The goal is to be able to better understand where in a sequence a particular motif is likely to evolve, and to potentially be able to predict evolutionary outcomes on the motif level. Combining the motif codon space and the nucleotide specific mutation rate I can thus assign any initial sequence a probability for evolving a motif given that the motif consensus sequence is well understood.

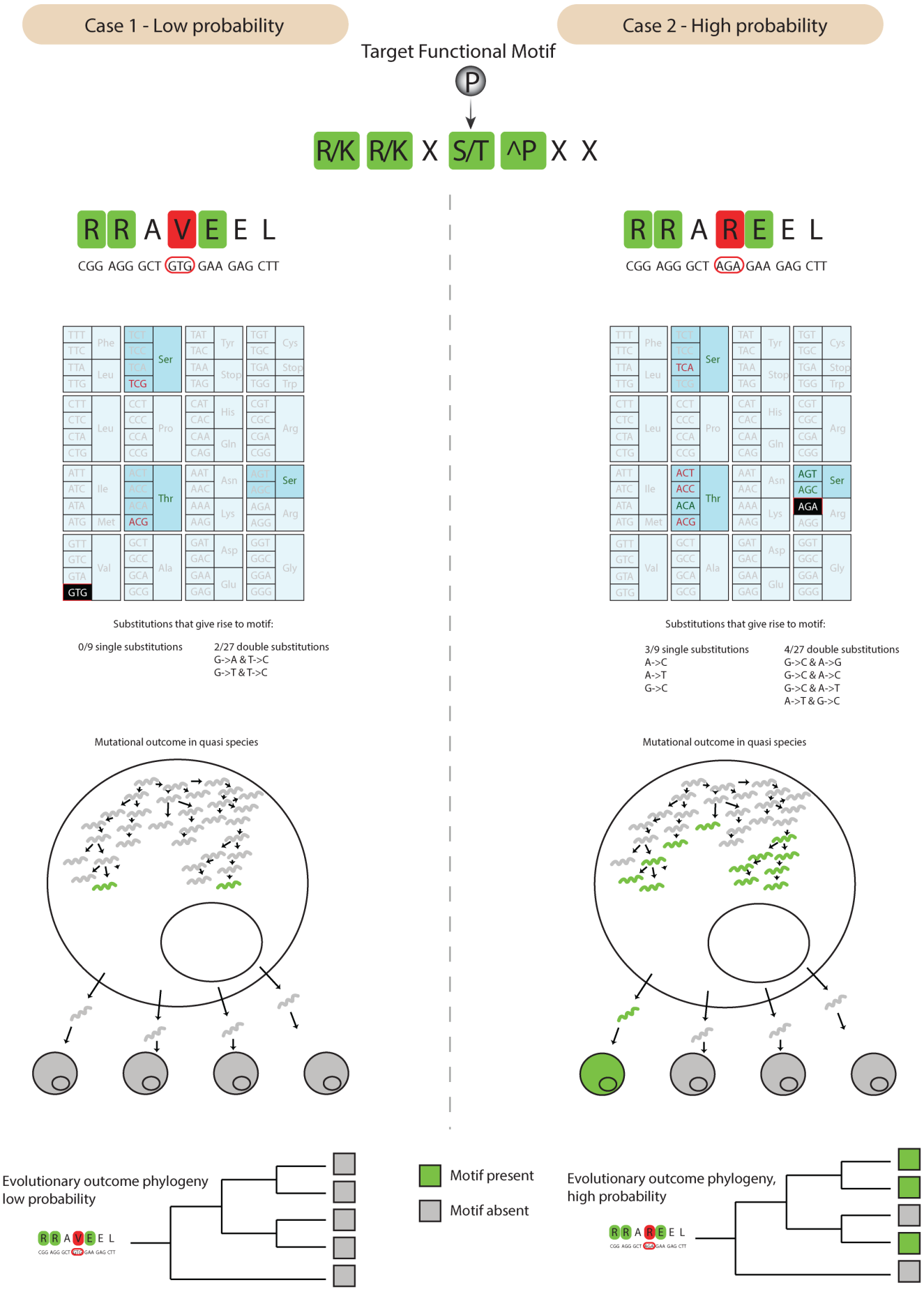


Figure 2.3. Diagram of the motif evolution model presented in this section. The left panel highlights the expected dynamics and evolutionary outcome of a low probability sequence and the right panel showcases the expected dynamics for a high probability sequence. The only difference between the sequences is the position for the Ser/Thr site which is GTG (Val) in the low probability scenario and AGA (Arg) in the high probability scenario.

The probability is calculated as follows: For a given initial codon (for example AAT) to substitute into serine or threonine, the probability for a codon substitution is first computed as the factorization of three single nucleotide substitutions, i.e.

$$P(AGC|AAT) = P(A|A) * P(G|A) * P(C|T)$$

All the substitution probabilities for the individual codons allowed at the position (i.e. encoding Ser/Thr) are then added together i.e.:

$$P(SerThr|AAT) = P(AGC|AAT) + P(AGT|AAT) + P(ACC|AAT) \ldots$$

All amino-acid substitution probabilities at each defined motif position are ultimately multiplied to get a whole motif probability, i.e.

$$P(Motif|Seq_{Init}) = P(Consensus_1|Codon_1) * P(Consensus_2|Codon_2) \ldots$$

where Seq_{init} denotes the starting sequence formed by $Codon_i$ at position i, and $Consensus_i$ the consensus amino-acid at position i defining the motif (Ser/Thr in the example above).

This results in high probabilities for sequences that have amino acids already matching the motif sequence at several positions as well as requiring few mutations (the fewer the higher probability) to gain any missing codon (and amino acid) required to form the motif. The probability is also increased if there are multiple mutational paths to reach an allowed codon space through single and double mutations with high nucleotide specific mutation rates. This calculation can be formalised as follows:

For a motif with i consensus positions we define the set of key positions as

$$M_i = \{i \in \{\text{MotifPos}_{\text{defined}}\}\}$$

For the probability of an initial sequence Seq_{init} with i codons, to gain the substitutions required to match a motif defined by i residues in $M_i$we can define a set of codons C, for each position i, such that

$$C^i = \{N_1^i, N_2^i, N_3^i \mid AA(N_1^i N_2^i N_3^i) \in \{AA_{motif}^i\}\}$$

where N_1, N_2, N_3 are nucleotides at codon positions 1, 2 and 3 respectively and AA_{motif}^i are the amino acids allowed at motif position i. The probability of the i^{th} codon in Seq_{init} to substitute into C^i can thus be expressed as

$$P(C^i|S^i) = \sum_{N_1^i, N_{2,}^i N_3^i \in C^i} P(n_1 = N_1^i | n_1 = S_1^i) * P(n_2 = N_2^i | n_2 = S_2^i) * P(n_3 = N_3^i | n_3 = S_3^i)$$

where $S^i = S_1^i S_2^i S_3^i \in Seq_{init}^i$. For the whole motif, i.e. all positions i, the probability of substituting any initial sequence, Seq_{init}, to a motif can finally be expressed by the single-site factorization over the motif:

$$P(motif|Seq_{init}) = \prod_{i \in M_i} P(C^i|S^i)$$

This tool can be used to scan through coding sequences to establish where motifs can easily evolve within the sequence and which mutations are required for this to happen. It describes the probability of the required mutations occurring over a single round of replication. This information can then be used either to inform analysis of past evolutionary events or to predict likely evolutionary events in the future. Different nucleotide substitution or mutation rates can be used to calculate the codon substitution probabilities, for example the experimentally determined per-replication mutation rates or substitution rates inferred from a phylogenetic tree.

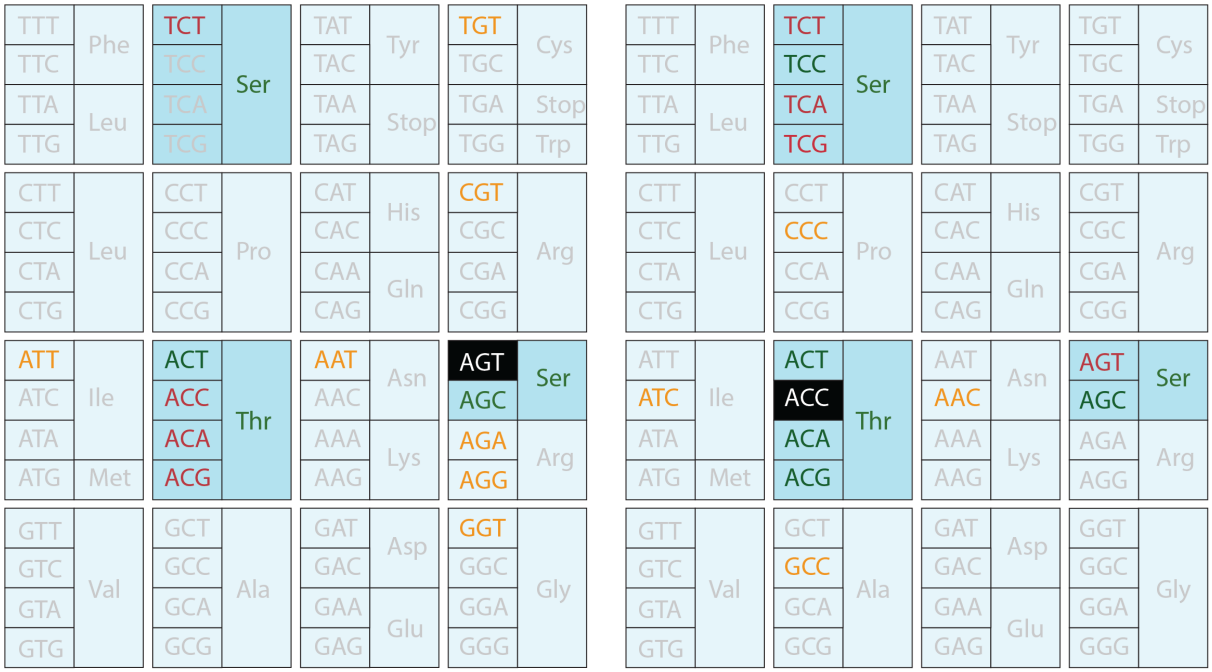


Figure 2.4. The difference in codon space for codons within Ser/Thr. AGT on the left has 2/9 of its single substitution space within Ser/Thr (green) and the rest result in non-motif substitutions (orange). AGT also has 4 available double substitutions and the rest of Ser/Thr space requires 3 substitutions. Conversely ACC, shown on the right, has 5/9 single mutations in Ser/Thr space (green) with only 4 outside of it (orange). It also has the remaining Ser/Thr space in reach within double substitutions, and none require 3. Assuming unbiased mutation rates ACC should be more robust to mutations for positions requiring Ser/Thr as a higher proportion of mutation outcomes are within Ser/Thr space.

The probability provides the expected outcome under neutral conditions and does not account for any fitness benefits or costs of the mutations. In general, establishing the evolutionary potential in the absence of selection will be informative as a means of predicting and analysing origins of evolutionary innovations and features on which selection can then act. Often, evolved features will not have an immediate benefit but will emerge neutrally, until a change of conditions provides a fitness benefit to the new feature (Wagner 2012). This was famously demonstrated in *E. coli* populations where one lineage evolved the ability to metabolise citrate which was facilitated by a prior neutral mutation (Blount *et al.* 2008).

The probability metric is not only a tool to evaluate properties of new evolutionary innovations in motifs, but can also be used where motifs are already present and functional. Given the amino acids defined as functional within given motif positions, different codon choices can theoretically result in different probabilities that random mutations are silent in the motif context or result in losing a motif (Figure 2.4). I will refer to this as motif robustness or codon robustness in the motif space. This idea will be discussed in later parts of this thesis.

With this method I have provided a novel conceptual framework to examine how likely random mutations are to turn a sequence lacking a given motif into a sequence containing a motif. Given that this probability calculation is estimating what is happening to sequences on a per replication basis, before any selection takes place, it will be important to further try to understand the effect this will have on an evolving population of mutating sequences such as a virus. Importantly, because the probabilities describe the expected rate of emergence, a model for the population distribution of different genotypes would be an important first step in understanding how likely new "motif genotypes" are to be fixed in a population. Within viruses we therefore want to look at the quasispecies model of infection.

2.3.4. Quasispecies model assumptions suggest a plausible evolutionary mechanism for motif evolutionary properties

The suggested importance of quasispecies dynamics (introduced in 2.2.4) during viral infections also has major implications for evolutionary innovations in viruses, in particular for motifs. Mutation frequencies have been implicated in fitness and evolvability through experimental work in viruses. Of particular relevance to motifs are studies looking at codon choice and fitness within the viral quasispecies. As mentioned in the introduction to this chapter (2.2), it has been shown that leucines and serines using codons that are 1 mutation away from a stop codon, rather than 2 mutations away, have drastically lower fitness. This suggests that, like in the model developed here for motifs, the rate of unfavourable nonsense mutations increases enough to impact the overall infectivity, showing that these codon-based frequency-of-outcome effects play a real role in viral evolution. In another set of experiments, codon choice and fitness have been shown to be linked: replacing influenza codons with less optimal codons reduces viral fitness and infectivity despite there being no difference in viral replication rate (Lauring *et al.* 2012).

Overall, these experiments indicate that different codon choices lead to different mutational codon outcomes or in other words different frequencies of mutating into more or less favourable codons and amino acids. This ultimately affects fitness through the impact on the diversity within the quasispecies. This has a direct relevance for motif evolution since motifs are short linear sequences within the proteins. Given the probability model I have established in this chapter we would expect motifs that have a higher probability of evolving to emerge with a higher frequency during replication in viruses. This

would lead to a larger proportion of the viral quasispecies having a new motif phenotype with potentially new interactions and functionality in a way that may be neutral or beneficial for the virus. In these instances, the expectation would be that high probability motifs would emerge earlier on average during replication cycles and also more frequently independently and thereby make up a larger population (see model in Figure 2.3).

It has been suggested that the quasispecies nature of viral populations improves overall infectious fitness. This may be due to an interaction between viral subpopulations with mutations that give them a synergistic relationship which enables them to interact differently with the host, thereby favouring the infection as a whole (Andino & Domingo 2015). It may also be due to the increase in diversity, allowing the viral population to adapt more easily to shifting environmental and immune-response related pressures. In each of these potential functional aspects of the quasispecies, motifs could plausibly play a significant role by allowing viral proteins to explore a range of new interactive and regulatory functionalities through sub-populations that acquire new motifs from high-probability sites.

I also expect through similar mechanisms of quasispecies mutation frequencies that functionally essential motifs in viruses would tend towards reducing the probability of mutations that lead to a loss of that motif. Given the motif probability metric I have established that can be used to determine the loss probability for functional motifs as well (i.e. motif robustness), it would be expected that certain codons are more robust to mutations – have lower loss probability – with respect to the motif phenotype. Within the quasispecies fitness paradigm I would expect that motifs that are essential but frequently lost would reduce overall fitness in the same way that leucines that frequently mutate into stop codons reduce fitness. Conversely, motifs that are encoded by mutationally robust codons would tend to be lost later in the infectious cycle on average, and be lost fewer times independently. Therefore, I hypothesise that there might be a motif-centric codon bias that allows viruses to optimise the evolutionary robustness of essential motifs within key proteins. These hypotheses will be explored in detail in chapter 3 in this thesis.

2.4. Simulations of motif evolutionary dynamics

With the model for motif evolution developed in the previous section, I wanted to determine how different combinations of codons and mutation rates are likely to impact the gain and loss of a range of motifs in evolving sequences. The model implies that initial codon choice, through motif codon space and the nucleotide specific mutation rate, will determine the rate at which motifs appear or disappear in populations. This suggests that certain evolutionary outcomes can be predictable, and that there might be detectable codon patterns for putative, previously uncharacterised motifs that indicate selection for motif function.

The expected impact of either motif codon space and mutation rate on the motif evolutionary outcomes are not immediately obvious. I will thus first outline some base expectation given the assumptions of the model. In influenza, mutation rates vary from ~10^{-4} mutations/nucleotide/replication to ~10^{-6} mutations/nucleotide/replication depending on the nucleotide (Pauly *et al.* 2017). The impact from codon space and codon choice will depend on the motif in question and whether motif loss or motif gain probability is considered. For motif gain, a position requiring a single mutation will be expected to emerge with a frequency that is on a similar order to the mutation rate or slightly higher (10^{-4}~10^{-3} in influenza) depending on the number of independent mutations that will give the motif outcome. To highlight this, proline codon CCC can mutate to Ser/Thr either through a C-to-T or a C-to-A mutation and through a range of double mutations. If those mutations are on the order of $3*10^{-4}$, the added probability for all paths to Ser/Thr will be ~10^{-3}. Each genome segment during influenza infection will be replicated between 10 000 and 15 000 times on average per cell, and in addition encode ~80 000 mRNA strands (Frensing *et al.* 2016; Heldt *et al.* 2012). For a motif with a gain probability of ~10^{-3}, this would result in motif emergence on average 10-15 independent times per cell, and additionally affect the mRNA pool and proteins translated. In contrast, at positions requiring more than a single point mutation, the combined mutation rates compound fast yielding very unlikely outcomes. Two required mutations are on the order of 10^{-6}-10^{-12} and three mutations <~10^{-12}. For motif loss the interaction between codon choice, mutation rate and probability is difficult to predict intuitively but will depend significantly on the degeneracy of the allowed amino acids and the number of favourable outcomes that any given codon yields as well as the mutation rate. This provides a basic intuition for the general effects of mutation rate and codon space at key positions within motifs.

In this section I have performed a range of simulations exploring the effect of different codons and mutation rates on the frequencies of mutational outcomes. I have explored the frequency of evolutionary events in a simulated phylogenetic tree for both the gain of new motifs and the loss of existing motifs. I have also performed simulations exploring the population growth dynamics of certain motif phenotypes in simulated viral sequence populations to approximate the conditions of the viral quasispeices under varying starting and mutational conditions.

2.4.1. Simulating motif evolution in viral strains in a phylogenetic tree

To explore the evolution of a range of sequences of viral strains over a phylogenetic tree I have used Pyvolve (Spielman & Wilke 2015). Pyvolve is a python package that allows users to evolve a set of sequences over a phylogenetic tree of defined topology using custom mutation probabilities. It is thus a useful tool for looking at expected neutral evolutionary outcomes given known mutation parameters. I have evolved a range of sequences over the tree topology using different mutation rates and parameters. I have specifically explored the dynamics of motif evolution for a range of sequences using different codons to explore the potential impact of prior sequence on the emergence of new motifs in the

absence of selection (Figure 2.5). This provides insights into the generation of diversity on which selection can act and also the rate at which certain motifs can be expected to emerge assuming they are not detrimental to viral fitness.

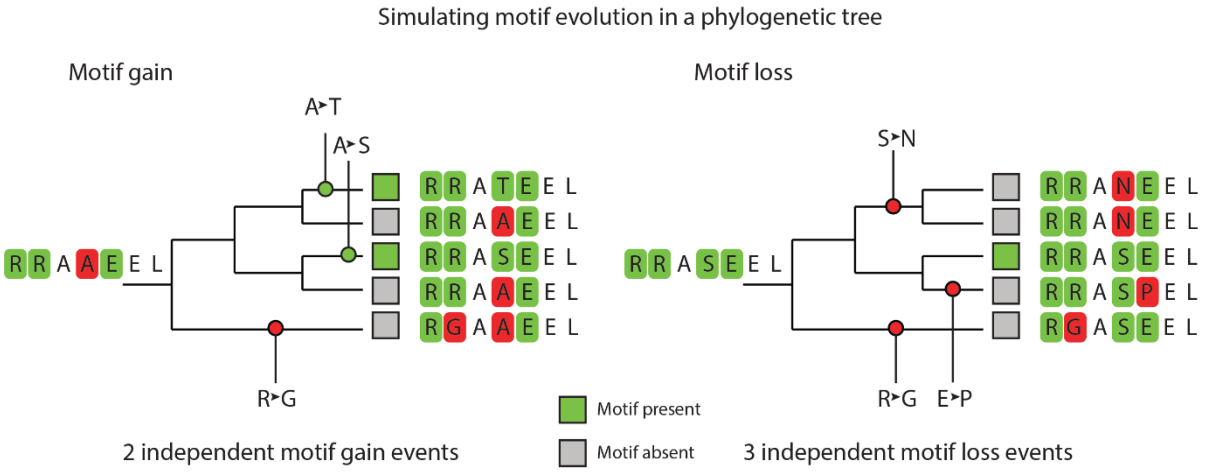


Figure 2.5. Diagram of the simulation and motif scoring approach for the gain or loss of the PKA consensus motif [RK][RK]x[ST][^P]xx. Independent motif gain and loss events were scored across simulated trees for a range of initial sequences. Two example sequences are shown to highlight how changes were determined across the tree.

2.4.1.1. Simulating motif emergence in evolution

To simulate motif gain I used the Protein Kinase A (PKA) recognition sequence. Phosphorylation sites are functionally important in the majority of influenza proteins and can have drastic effects on protein function and regulation when lost or gained (Hutchinson *et al.* 2012). The PKA consensus considered here is [RK][RK]x[ST][^P]xx, where x is any amino acid (MOD_PKA_1 from ELM (Gouw *et al.* 2017)). I have simulated the evolution of five different short sequences all with slightly different initial codon choices in key positions to explore how and when new PKA motifs could evolve through point mutations over a phylogenetic tree. The tree was constructed from real influenza NS1 sequences from H1N1 strains taken from previously published work by Worobey *et al*, (2014) (See Methods, section 2.7.1). The query sequences were then evolved over all the different branches, all using the same relative mutation rates, determined experimentally by Pauly *et al*, (2017). This allowed me to investigate the accumulation of mutations and the evolution of features under the given mutational conditions but without any selection acting on the system. For this analysis a total of 4569 strains were in the phylogenetic tree, and it spans a period equivalent to about 100 years of H1N1 strain evolution.

To explore the impact of codon choice in sequences that are within 1 mutation of the motif consensus, I determined the codons most and least likely to evolve into a serine or threonine given the nucleotide specific mutation rate. The codon most likely to mutate is ATC (Ile), which requires a single point mutation to mutate into ACC (Thr) and can also mutate into AGC (Ser).

Table 2.1. Results from running simulations with different starting sequences over a phylogenetic tree. Results from 1000 replicate simulation runs. Different starting sequences encoding protein sequences that require one or two amino acid changes to form a new motif evolve new motifs at very different frequencies depending on the codons used.

ProtSeq	NucSeq	Independent Evolution	Trees with Motif (%)	Avg # strains with motif per tree
RRGINGA	**AGG** AGG GGT **ATC** AAC GGT GCA	1715	72.3	131.9
RRGINGA	**AGG** AGG GGT **ATA** AAC GGT GCA	673	43.8	102.3
RRGRNGA	**AGG** AGG GGT **CGC** AAC GGT GCA	554	36.4	86.2
RRGVNGA	**AGG** AGG GGT **GTC** AAC GGT GCA	677	36.8	66.2
RRGQNGA	**AGG** AGG GGT **CAA** AAC GGT GCA	40	3.4	5.2
WRGINGA	**TGG** AGG GGT **ATC** AAC GGT GCA	178	11.2	31.8

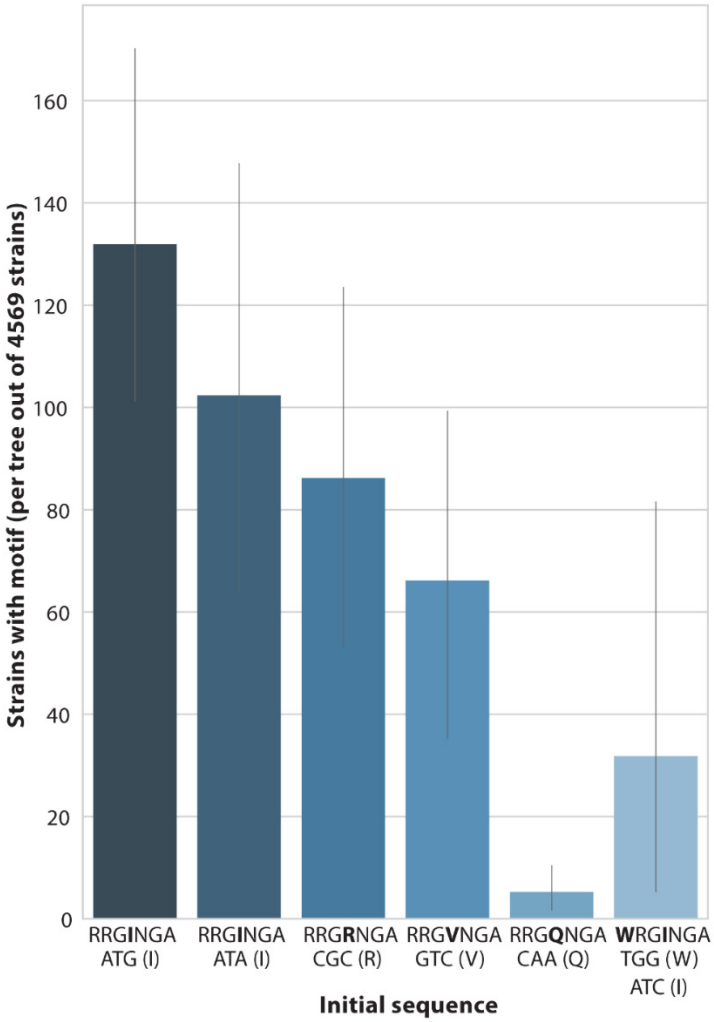


Figure 2.6. Strains with a new evolved motif in simulated data. Motifs require the consensus sequence [RK][RK]x[ST][^P]xx, meaning the input sequences are either missing Ser/Thr or Arg/Lys. Depending on the codon in the starting sequence (different codons used highlighted under sequences in x-axis) motifs have different frequencies in the strains across the tree, highlighting the big impact prior sequence can have on functional motif innovations in viruses.

The codon within 1 point mutation that is the least likely to acquire the right mutation is CGC (Arg) which can only mutate to S/T through a C-to-A mutation which happens at a very low rate. I simulated otherwise identical sequences using either ATC (Ile) or CGC (Arg) to compare the impact on motif evolution that might result from a single codon difference. For the high probability input sequence I thus used AGG AGG GGT **ATC** AAC GGT GCA encoding RRx**I**Nxx and for the low probability sequence I used AGG AGG GGT **CGC** AAC GGT GCA encoding RRx**R**Nxx. I also analysed sequences that were two mutations away from the PKA consensus, either two mutations in a single codon or a single mutation in two different codons. This allowed me to investigate the importance of the difference in a codon based approach to sequence evolution compared to amino acid based approaches by comparing sequence similarity on the amino acid level with the different outcomes of the simulation. Over 1000 independent simulation runs I scored the number of trees that saw the motif evolve, the number of independent evolution events and the total number of strains with the motif (Table 2.1).

Interestingly, the initial codon choice drastically determined the number of times a new motif evolved, both if it at all emerged and how many independent times it evolved. As a result, the number of strains with the motif for each independent tree simulation also closely correlated with the initial sequence (Figure 2.6). The high probability, single point mutation sequence with ATC (Ile) evolved in ~72% of all independent simulations and 1715 times independently. By contrast, the low probability sequence with CGC (Arg) only evolved in 36% of simulations and 554 times independently. The distribution of the number of independent events in the high probability sequence also show a marked difference, with several more independent events per tree on average (Figure 2.7). This difference is a result from only the difference from ATC (Ile) or CGC (Arg) with the rest of the sequence being identical. On the amino acid level this is the difference between an Ile-to-Ser/Thr substitution compared to and Arg-to-Ser/Thr substitution. Interestingly, there was a large difference even between the two isoleucine codons, ATC and ATA, with ATA only evolving in 44% of simulations compared to 72% for ATC. From a biological perspective it is clear that – under the assumption that this is a functional new phosphorylation site with a small positive fitness effect – a prior sequence with ATC (Ile) drastically increases the chance of this new feature evolving, even comparing to another isoleucine codon.

Conventionally, motif evolution has been studied on the amino acid level, and finding instances where only a few amino acid changes are required can seem like a good indicator of where motifs may evolve. However, these simulations clearly show that codon evolutionary probability matters more than amino acid sequence distance, as different codons within a single residue can have different probabilities (*cf.* Table 2.6 for all [S,T] probabilities). In addition, an amino acid sequence that appears closer in sequence space can evolve into a new motif less frequently than another sequence that appears more distant based on the amino acid level. This is exemplified in these results by comparing the rate at which WRxIxxx and RRxQxxx evolves into the motif. RRxQxxx appears closer based on

amino acid sequence alone, however, both require two point mutations to evolve the PKA motif. Despite this, WRxIxxx evolves in 11% of all simulated trees compared to RRxQxxx evolving in less than 3.5% of trees and the number of independent gain events is more than 4-fold lower in RRxQxxx. This effect is a motif-specific interaction between the codon space for each position in the motif and the nucleotide specific mutation rate in the organism and needs to be calculated in a context-dependent manner. This is counter intuitive from a traditional perspective but makes sense within this new framework of motif evolution.

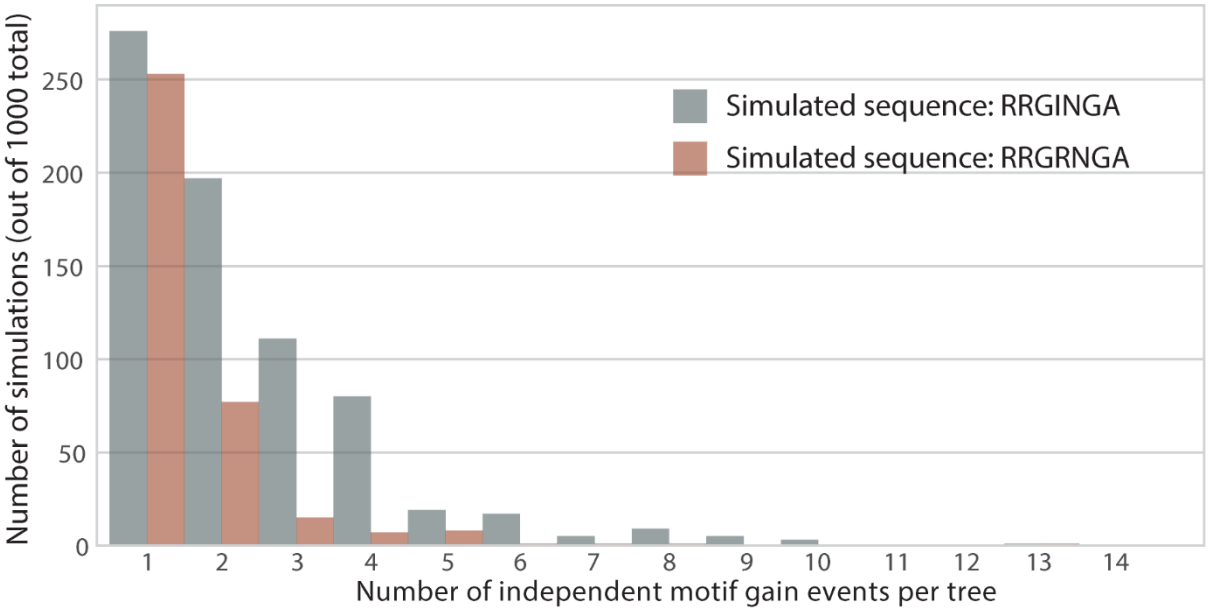


Figure 2.7. Number of independent evolution events in a tree (x axis) and how many trees saw a given number of independent events (y-axis). In grey is the sequence RRGINGA and in red the sequence RRGRNGA, both a single point mutation away from evolving a motif. These simulations show a stark difference between how many times strains in a phylogenetic tree of this topology are likely to evolve a motif *de novo* depending on a single codon difference.

2.4.1.2. Simulating motif loss

In the same way prior codon choice impacts gain of new motifs over evolution, I hypothesised in the previous section in chapter 2 that the loss of already existing motifs can also be influenced by the initial codon choice. The different potential shape of the codon space, alongside nucleotide specific mutation rates, will make certain codons more or less likely to mutate. In addition, when they do mutate, they will be more or less likely to mutate into another codon that is still within the motif consensus (see previous section, Figure 2.4). To investigate the evolutionary dynamics of motif loss given different codon choices I have simulated different sequences over the same tree as for motif gain. I again ran 1000 independent simulations for each sequence, scoring the number of strains where the motif is lost, total independent loss events and the number of simulated trees with very high (>95% of strains) loss of motifs. Comparing the same phosphorylation motif as for motif gain I ran simulations for two

sequences: one that is predicted to be more robust to mutations that disrupt the motif and one that is predicted to be less robust to such mutations. The more robust sequence used was **CGA CGA** GGT **ACC AAC** GGT GCA encoding **RR**G**TN**GA and the less robust sequence used was **AGG AGG** GGT **AGT CTT** GGT GCA encoding **RR**G**SL**GA. I determined which codons were high and low robustness within this motif consensus using the calculations described previously (see section 2.7.3 for details). Given these specific mutation rates, CGA emerged as the most robust Arg/Lys codon, ACC as the most robust Ser/Thr codon and AAC (Asn) as the codon least likely to mutate into proline (satisfying the [^P] position). In contrast AGG was the least robust Arg/Lys codon, likely due to low A-to-C mutation rates and low G-to-A mutation rates as well as a lack of four-fold degeneracy in the third position leading to fewer silent mutational outcomes. Similarly, AGT is the least robust Ser/Thr codon due to a lack of four-fold degeneracy and high mutation rates for mutating into glycine and asparagine among others. Finally, CTT (Leu) is one of the codons most likely to evolve into proline.

I found again that the evolutionary pattern for loss correlated directly with the predicted robustness of the codons chosen (Table 2.2). The less robust set of codons mutated so that the motif was lost earlier in the tree on average, leading to a larger part of the tree and a larger total number of strains without the motif than for the more robust codons. The loss frequency of the motif was determined by the number of independent times the motif was lost normalised by the number of strains with the motif as strains that inherit a genotype without the motif are unable to lose it. The loss frequency was over 50% higher in trees with the non-robust motif compared with the robust motif. The average number of strains with the motif in trees using the robust sequence was ~40% whereas in trees with the less robust sequence that number was only ~29%, which is a ~28% reduction in motif presence in strains. On average the number of strains without the motif was ~20% higher in the trees using the less robust motif. The number of simulated trees that had an almost total loss of the motif (>95% of strains) was also much higher in the non-robust motif at ~16% and only ~10% when using the more robust motif. Taken together, this clearly shows that the expected mutational dynamics in viral strains can theoretically be impacted by codon choice in functional motifs, and that motifs encoded by codons with different robustness will have potentially different fitness and thus evolutionary outcomes over time.

Table 2.2. Results from running simulations with different starting sequences over a phylogenetic tree. Results from 1000 replicate simulation runs. Different starting sequences encoding protein sequences that are more or less robust to mutations that disrupt the motif. Sequences that have high robustness are also lost less over a phylogenetic tree, suggesting there can be a fitness gain from sequences that maintain more robust codons in motifs.

ProtSeq	NucSeq	Independent Loss Freq (per 1000 strains)	Strains with motif (per tree) %	% trees with >95% motif loss
RRGTNGA	**CGA CGA** GGT **ACC AAC** GGT GCA	11	40.3%	10%
RRGSLGA	**AGG AGG** GGT **AGT CTT** GGT GCA	17	28.8%	16.2%

2.4.2. Simulating motif emergence frequencies in a growing viral population (quasispecies)

In contrast to simulations over a phylogenetic tree, in this section I wanted to simulate motif emergence and evolution in a context more similar to that during viral infection and rapid sequence expansion. During infections influenza forms a quasispecies, which is a heterogeneous population with many different genotypes that emerge through mutations, as discussed in the background to this chapter. To simulate the dynamics during viral replication in a single cell or host I wanted to capture this replication behaviour of influenza (for details on influenza infectious cycle, see section 3.2.1). During replication, negative sense RNA strands form the viral RNA (vRNA) pool that will make the future generation of viruses and that also act as the mRNA template (Samji 2009). Positive strand RNA – copied from vRNA – forms the pool of complementary RNA (cRNA), which act as templates for making more vRNA. To capture this in a way that closely reflects our best current understanding of influenza replication, I evolved sequences in two pools representing vRNA and cRNA (Figure 2.8). A single founder sequence made the first pool of vRNA, which was replicated using the experimentally determined mutation rates, to generate the first pool of cRNA. This new pool of cRNA was then replicated creating new vRNA sequences, which were added to the existing pool of vRNAs. This back and forth was performed until the cRNA pool contained 300-400 sequences, reflecting expected cRNA amounts (Frensing *et al.* 2016). After this point, vRNA-to-cRNA replication was stopped and cRNA-to-vRNA replication generated the remaining pool of sequences up to a final population of ~10000 vRNA sequences. This captures what we currently know about the average sizes of molecular populations and replication dynamics during influenza infection in individual cells (Frensing *et al.* 2016; Heldt *et al.* 2012; Kawakami *et al.* 2011). To date, no studies characterising the mutational heterogeneity in a single cell during infection have been performed. However, single-cell transcriptomics analyses have shown a remarkable heterogeneity in expression levels and the immune response generated in individual cells (Russell *et al.* 2018). Together with the experimentally measured mutation rates, this suggests that the level of mutational heterogeneity in single cells is likely to be significant. All newly formed sequences are replicated with a set of mutation rates that have been established through direct experimental measurements of influenza polymerase error rates (Pauly *et al.* 2017). After several generations of replications, the quasispecies was analysed to determine where motifs had evolved or been lost and what proportion of the population had new motif phenotypes.

These simulations are likely to reflect some of the same correlations and patterns found in the phylogenetic tree simulation. However, the specific dynamics of a population of sequences forming a quasispecies is likely to better reflect the true neutral rate of motif emergence and loss in the population since mutations occur during each replication. In addition, the phylogenetic tree spans a much larger timeframe and the topology of it is shaped not only by mutation frequencies and sequence space, but

by selection, human socio-political factors, geography and other factors which can skew the evolutionary dynamics I am trying to investigate. In this way, the population simulations here should be more informative of overall evolutionary behaviour as the temporal resolution is higher.

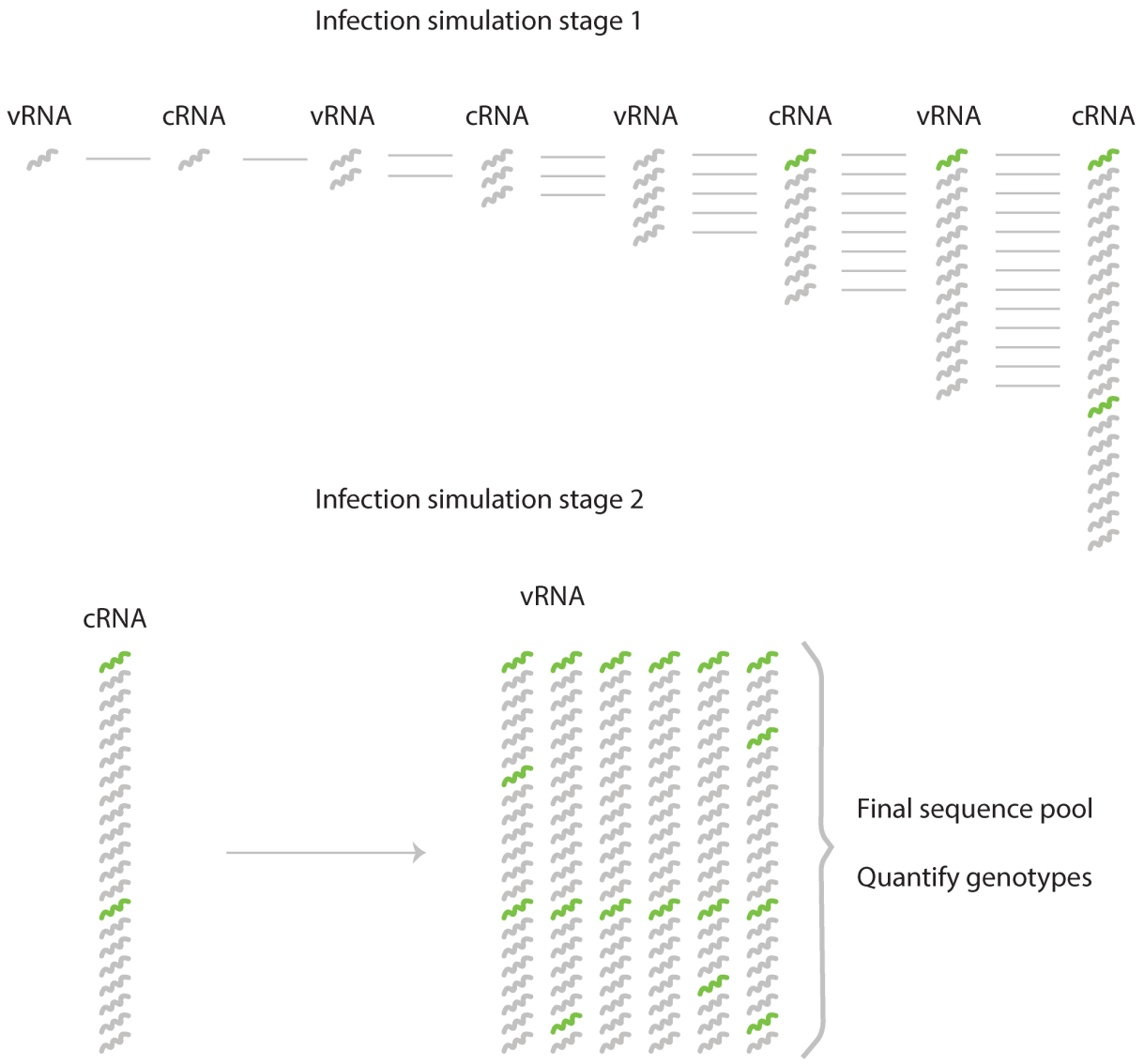


Figure 2.8. Diagram visualising the approach to quasispecies sequence simulations. Simulations were performed in a two stage process. The initial stage evolves the sequence between two populations (cRNA and vRNA) under exponential sequencec increase. In the second stage a fixed cRNA pool replicates to create the final full vRNA pool.

2.4.2.3. Simulating quasispecies motif gain

To determine how the gain of new motifs depend on the prior sequence, and how that impacts the proportion of the population with a novel functional motif, I ran the simulations using the same sequences as I used for the phylogenetic tree evolution. Specifically, I have compared the differences between

the two sequences that are a single point mutation away from the PKA motif. The first sequence encodes Ile using ATC (high probability of emerging) and the second encodes Arg through CGC (low probability of emerging). The proportion of the population with the new motif was then determined after the replications had been performed. 3000 independent simulation runs were performed and analysed, with each run being equivalent to approximately the population expected in a single cell. For the sequence using ATC (Ile), the motif emerged in ~47% of all simulations. By contrast, the motif only evolved in 5% of simulations using the CGC (Arg) sequence (Table 2.3).

Table 2.3. Results from motif gain analysis of two sequences in a quasispecies population. Even on a short scale within a simulated single cell the difference in motif evolution in two sequences with a single codon difference is striking. Higher gain rate through the high probability sequence is likely to impact fitness and evolutionary dynamics.

ProtSeq	NucSeq	Average Gain (percent)	Freq. of motif gain in pop.	Freq. of >1% motif gain	Freq. of >10% motif gain
RRGINGA	AGG AGG GGT **ATC** AAC GGT GCA	40.8 (0.4%)	47%	0.8%	0.3%
RRGRNGA	AGG AGG GGT **CGC** AAC GGT GCA	9.8 (0.09%)	5%	0.1%	0%

For the sequence using ATC (Ile), the average size of the subpopulation with the new motif was ~0.4% of the final population size which was around 40 sequences in a population of 10000. The proportion of the population with the motif was highly variable since it is strongly affected by when the mutation happens, as it is subsequently transferred through replication (the so called "jackpot" effect, (Luria & Delbrück 1943)). The average timing of the mutation is, unsurprisingly, closely related to the probability of the motif gain. In the ATC sequence the mutation rate for the motif change is on the order of $2.7*10^{-4}$ or approximately 1/3500. Conversely for the CGC (Arg) sequence the probability for this specific mutation is around $7.7*10^{-6}$ or 1/130000 which explains why it very rarely happens and the motif gain rate is so low in a viral population. These rates correlate closely with the number of replications before the mutation occurred on average, which explains the population sizes observed.

On average in ATC (Ile), 0.1-1% of the population end up with the new motif phenotype which is 10-100 viral sequences given this population size. This indicates that it mostly emerges around generation 10-11 when the total cumulative number of replications in the population is around 4000. In ~1% of simulations, >1% of sequences in the final population have the motif and in ~0.3% of simulations, >10% of the sequences have the motif (a potential "jackpot" event). In contrast, in the simulations with the CGC (Arg) sequence, when the motif did emerge it only did so very late in the replication cycle leading to a very small fraction of the population with the motif phenotype. On average, the few simulations in which the motif emerged only had <10 copies of the motif-bearing sequence or less

than 0.1% of the population, suggesting it only emerges in the last few replication rounds (Table 2.3). No simulation using this sequence had any "jackpot" events (>10% of the population with the motif).

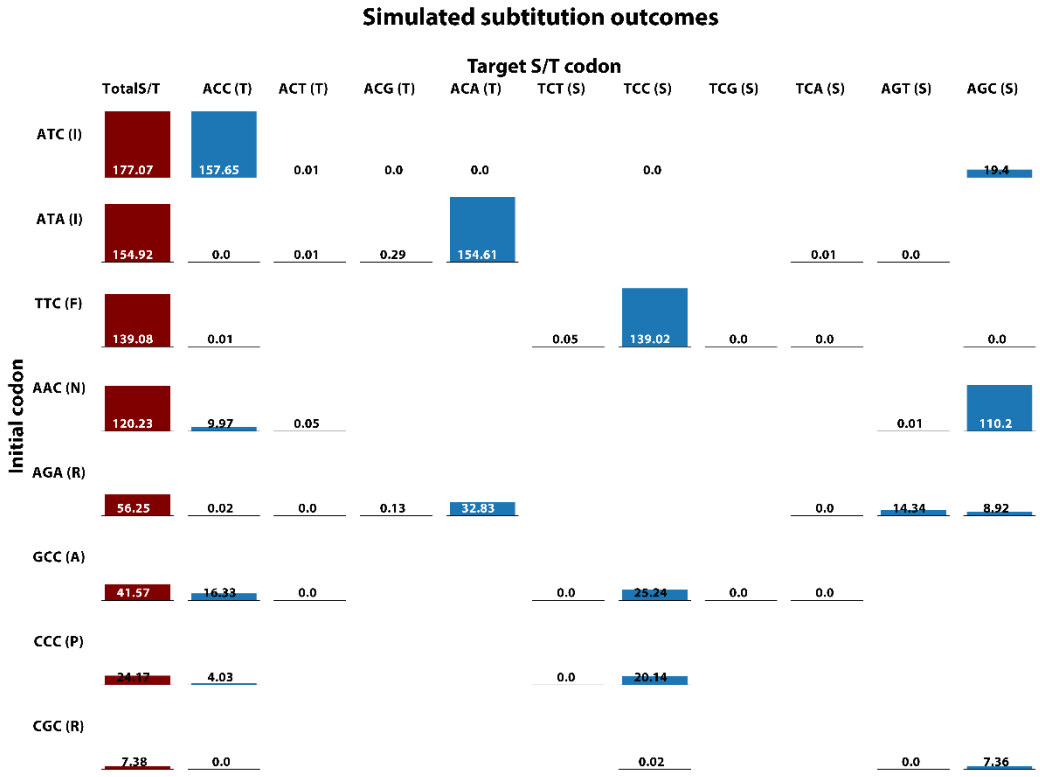


Figure 2.9. Simulation results of populations of different sequences and codons mutating into Ser/Thr codon space. All codons on the left are a single point mutation away from mutating into Ser/Thr but end up with drastically different mutational and evolutionary dynamics in population simulations. ATC (Ile) mutates into ACC (Thr) with approximately 25 fold higher frequency than CGC. Knowing the nucleotide specific mutation rate and the prior codon can greatly inform mutational sequence space and outcomes.

Finally, to get a more general idea of the mutational dynamics for the codons that are a single point mutation away from the Ser/Thr codon space I simulated seven different codons in isolation. I determined which mutations happened and summed all of the ones that were within Ser/Thr codon space (Figure 2.9). Here we can clearly see that isoleucine ATC is the most likely to mutate into Ser/Thr space because of a very high rate of ATC (Ile) to ACC (Thr) mutations. In addition, there is also a difference between isoleucine codons ATC and ATA from the added possibility of ATC substituting into AGC. Analysing all the mutational outcomes in the population in this manner clearly highlights the

interplay between codon space and nucleotide specific mutation rate for the evolutionary outcomes of codons.

These results from simulating sequence populations in single cells corroborate the conclusions from the earlier simulations and also strongly indicate that sequence regions with higher probability of evolving new motifs can be important for predicting the evolution of these new features.

2.4.2.4. Simulating quasispecies motif loss

To look at motif loss in sequences with different robustness I used the same sequences as in the phylogenetic motif loss analysis. Populations were replicated to the same size and through the same process as described for the motif gain simulations in a quasispecies. Again, the purpose was to determine the frequencies of motif-genotypes in the final population as a result of mutations and in the absence of selection. For the PKA motif there is a drastic difference in the frequency of the population without a motif in the sequence using more robust codons compared to less robust codons (Table 2.4). Maximising robustness lead to motif loss in less than 0.4% of the population on average. The most frequent non-motif genotype resulted from an arginine to proline mutation through a CGA-to-CCA mutation creating either PRGTNGA or RPGTNGA. In contrast, for the non-robust motif, motif loss occurs on average in 1-2% of the population leading to a substantially larger pool of sequences without the motif which is effectively a direct reduction in the fitness of the virus (fewer successful offspring). The number of simulations where motif loss was observed in >1% of final sequences was 4.4% and 16.8% for the robust and non-robust respectively. Motifs lost in >5% of final sequences (what could be considered an "anti-jackpot" effect) was seen in 1% and 5% of simulation outcomes for the robust motif and non-robust respectively. One additional implication of significant motif loss of an essential motif is likely to be the impact on viral fitness through the number of functional proteins within a cell as well as the successful number of sequence offspring resulting from each infected cell. This results from the fact that vRNA sequences encode both the proteins crucial for the ongoing infection and make up the genetic material for subsequent virions.

Table 2.4. Results from mutation loss analysis of two sequences in a quasispecies population. There is on average a 4-fold difference in the frequency of motif loss between the more robust sequence and the less robust one. In addition, very large population motif loss also happens more frequently in the less robust sequence.

ProtSeq	NucSeq	Average Loss (percent)	Freq. of >1% motif loss	Freq. of >5% motif loss	Freq. of <0.1% motif loss
RRGTNGA	**CGA CGA** GGT **ACC AAC** GGT GCA	45 (0.4%)	4.4%	1%	69%
RRGSLGA	**AGG AGG** GGT **AGT CTT** GGT GCA	140.7 (1.3%)	16.8%	5%	5.6%

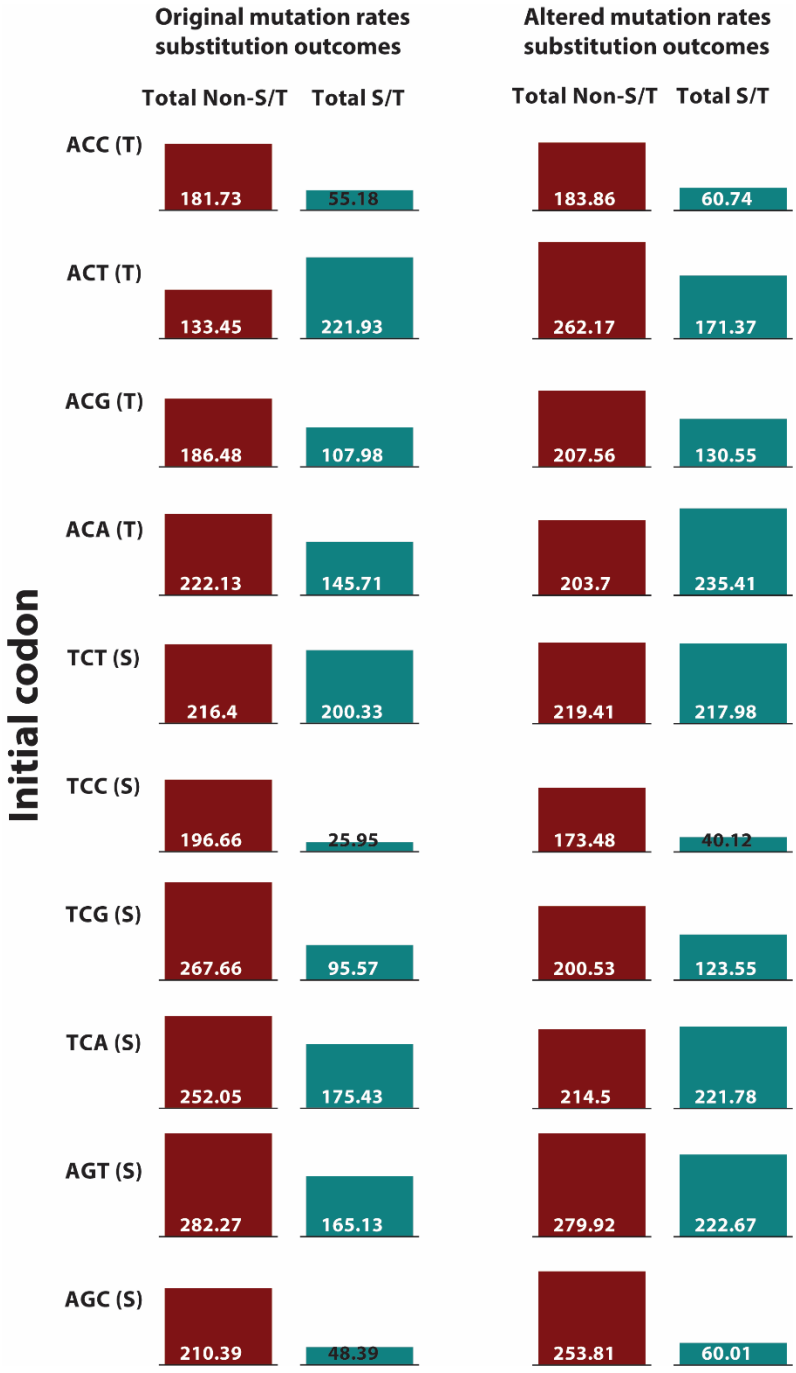


Figure 2.10. Comparing Ser/Thr codon substitution outcomes. The different Ser/Thr codons substitute away from Ser/Thr codons and silently within Ser/Thr codons at significantly different rates in these simulations for both subtitution rates (Original rate: χ^2 (9, N = 30000) = 30.0, $p < .001$, Altered rate: χ^2 (9, N = 30000) = 28.1, $p < .001$). In addition, when altering the mutation rates within the observed margin, the codons substitution rates change markedly. There is a significant difference in mutation rates between the codons ACC, ACT, ACG, ACA, TCA, AGT and AGC between the two different rates (Mann-Whitney test, $p<0.005$) This illustrates the importance of the mutation rate as well as the codon space in mutational outcomes. The original mutation rates are from A/Puerto Rico/8/1934 (H1N1), and the alternative rate is from A/Hong Kong/4801/2014 (H3N2) (see methods section 2.7.1).

Finally, to determine the general loss dynamics for the different Ser/Thr codons I simulated all 10 codons in isolation and determined the frequency of each mutation in a similar way as for gain in the previous section (Figure 2.10, left panel). There is a large difference in the number of substitutions to

non-Ser/Thr codons. The difference between ACT (Thr), which is one of the least mutated, and AGT (Ser) which is highly mutated, is more than 2-fold. There is also a striking difference between the number of within-Ser/Thr "silent" mutations between the different codons. ACC has a very low number of silent mutations due to C in the 3rd position having the lowest overall mutation rate, whereas ACT has a very high silent mutation rate with T in the 3rd position.

I also wanted to compare the impact from different nucleotide mutation rates on these outcomes. I used a slightly different set of mutation rates, that were observed in a different influenza strain from Pauly *et al*, (2017) and ran the same simulations again (Figure 2.10). This had a big impact on the outcome of some of the codons mutation rates, in particular between the codons ACC, ACT, ACG, ACA, TCA, AGT and AGC which all were significantly different between the two rates. Serine AGT and AGC were still among the codons with highest mutations in both data sets. This highlights that the nucleotide-specific mutation rates, unsurprisingly, have a very important impact on the mutational outcomes in the codon space, which needs to be considered and measured accurately for high quality predictions.

2.5. A resource for evolutionary characterisation of motifs in RNA viruses

In order to have a generalised tool that can be used to study motif evolution in RNA viruses and other short, fast evolving sequences (with good sequencing coverage) I have developed a pipeline that allows a user to study these systems. In summary, the pipeline uses virus nucleotide sequences to generate a multiple sequence alignment and a phylogenetic tree. The viral sequences are translated into proteins computationally, and each sequence is scanned through to identify the presence of any motif-matching sequence-patterns. Sequences, phylogenetic trees and motifs are then analysed together to track mutations and motif changes over time. This allows the user to get a detailed understanding of the events and evolutionary pressures that act on motifs within viral sequences and how various factors interact with motif evolution over time. The pipeline is summarised here (Figure 2.11).

2.5.1. Data collection and quality filters

The pipeline has been designed to work with data from the NCBI sequence databases (NCBI Resource Coordinators, 2017). It has been designed around analysis of virus sequences but can work for any nucleotide sequence that is well behaved from a feature perspective, meaning it is straightforward to translate the protein from the nucleotide sequence. Input data can take a number of different forms depending on user preference, input data can be specified as an NCBI search term

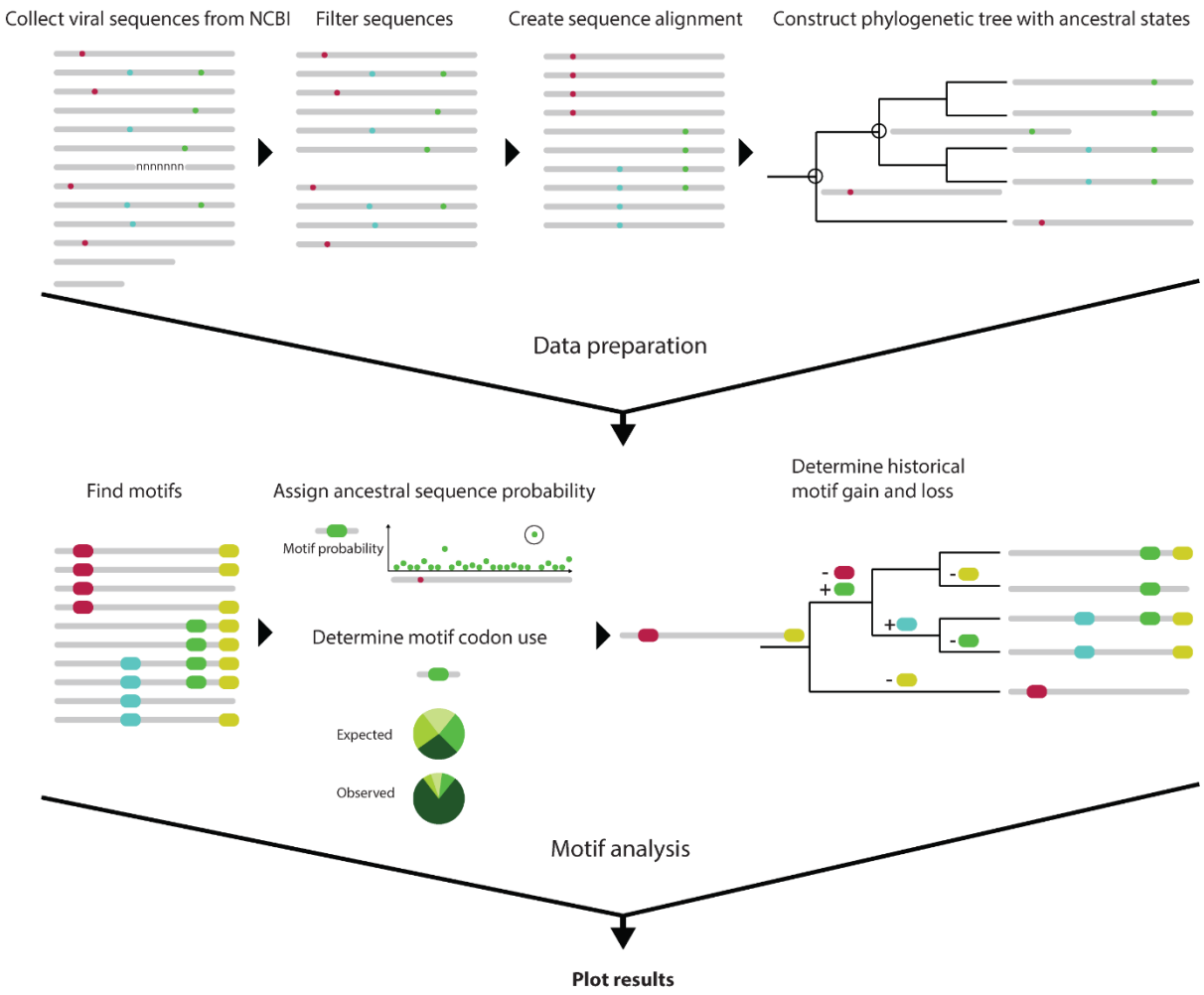


Figure 2.11. Summary of the pipeline for analysis of motifs in viral sequence records. Data preparation steps include collecting and filtering viral sequences and creating a sequence alignment and a phylogenetic tree which can be used for comparative motif analyses. Motif analysis steps illustrated here are regular expression identification of motifs, determination of codon usage and sequences probability for motifs and comparison of motif gain and loss over time.

(influenza A [strain] [gene id]) according to the biopython NCBI search format (Cock *et al.* 2009). The program will automatically fetch all the top matching NCBI sequence entries up to a user provided limit, save them locally and generate a master list of the NCBI entry IDs. Alternative data inputs can be a text file with NCBI IDs or a pre-processed fasta file containing coding-sequences of interest. Given user filtering parameters the entries can then be filtered according to a number of quality checks such as completeness of sequence, number of unknown nucleotides and sequence similarity between entries. The program also trims the sequences (in the case of NCBI data files) to include only the relevant coding sequence by finding the specified reading frame in the NCBI file. All sequences are then checked for translation compatibility and files containing sequences with missing start codons or files that cannot be translated properly due to sequencing gaps or other file errors are flagged and removed. If the reading frame is intact and sequencing gaps only generate unknown nucleotides these files can be kept and handled by the pipeline. The missing nucleotides are treated as unknown on the amino

acid level and that region does not interfere with the subsequent motif analysis as the unknown status is kept throughout for that part of the sequence. These filtering steps then generate the working list of IDs and sequences for the later steps of the pipeline.

2.5.2. Multiple sequence alignment and phylogenetic tree construction with ancestral sequences

For the full analysis of motif probabilities and motif evolution over time in the selected sequences, a multiple sequence alignment and a phylogenetic tree are required. The user can use external files of common file formats as inputs and this would be recommended for higher quality alignments and trees. However, for smaller datasets and highly similar short sequences such as in the case of many viral strains, I have implemented local multiple sequence alignments and phylogeny software within the pipeline that can run automatically on the collected data. The pipeline acts as a wrapper for ClustalW for multiple sequence alignments and RAxML and/or TreeTime for phylogenetic trees and ancestral sequence reconstruction (Larkin *et al.* 2007; Sagulenko *et al.* 2018; Stamatakis 2014). The wrapper for ClustalW checks for any current full or partial alignments in the working directory compared to the NCBI entries in the working list of entries from the previous step. If there is a partial alignment from before, any new files are added to this alignment which saves a significant amount of running time when new entries are added to an existing alignment. If no alignment is present a new alignment is created. Alignments are checked against any pre-existing trees to determine if a new tree is needed and if it is, this alignment is used to create a phylogenetic tree using either RAxML or TreeTime. To generate tree files with complete ancestral reconstruction of sequences, a text file is generated from the NCBI list with all the node names (usually NCBI IDs or similar) alongside the year extracted from the NCBI file. If some NCBI files are missing the year information, these entries will be flagged and the user will have to add them in the text file manually or remove them from the file and also specify they be pruned from the tree and removed from the alignment. The generated tree, alignment file (with matching ID names) and date file are then used by TreeTime to generate a new tree with full sequence reconstruction at all ancestral nodes.

2.5.3. Short linear motif identification, filtering and probability scoring

To prepare sequences for linear motif analysis a protein sequence alignment file is required (in contrast to the nucleotide alignment used to create the phylogeny). Using the nucleotide alignment file as a source I have created a custom translation algorithm that is faster and more suitable for RNA-virus data and motif analysis than the standard biopython translation function. There are two possible approaches to generate a protein alignment within the pipeline. One is a direct nucleotide alignment to protein alignment conversion, which works well for many RNA virus sequences as they are very closely related and contain few gaps. If all gaps are consistent with the reading frame being main-

tained i.e. in multiples of three or simply missing nucleotides rather than gaps, an alignment conversion is also possible. The second alternative is to generate a new protein sequence alignment using the sequence alignment wrapper after making a translated file. Alternatively, for a high quality alignment an external alignment file can be read. For motifs the Eukaryotic Linear Motif (ELM) database is used (Gouw *et al.* 2017). The user is required to download the ELM database of motifs in a CSV format to the working directory and the program can read in the relevant data, importantly including Motif ID, regular expression and optional gene ontology information. Using the protein sequences and ELM files a custom algorithm scans through each protein sequence looking for regular expression matches for each ELM, based on motif consensus sequence definitions. Matches are saved including their position and matching sequence. Given a list of motifs for each sequence the user can then choose to apply appropriate filters to exclude motifs in specific regions of the sequence. Motifs can be excluded based on predicted sequence features such as falling within predicted transmembrane helices or not being within disordered parts of a protein using the programs TMHMM or IUPred respectively (Dosztányi *et al.* 2005; Krogh *et al.* 2001). Where protein structures exist, a user created list of accessible surface area for residues can be used to exclude motifs in buried regions. After filtering, motifs are correlated with their nucleotide sequence such that each amino acid in the motif is assigned with the codon used to encode it to allow for direct mutational analysis of motifs through the sequences. After codons have been assigned, the nucleotide sequence can be used to calculate a predicted probability or robustness score for the motif at that position. If the motif is missing in that same region in some of the sequences these can also be assigned with a probability of regaining the motif in question at that particular location.

2.5.4. Probability analysis/scoring

The probability scoring is performed through a custom algorithm developed in accordance with the details outlined in 2.3. The user is required to provide a list of the nucleotide specific mutation rates for the organism in question. The algorithm uses these values to then calculate the combined likelihood of all possible mutational events for all the different codons at all the motif positions. The final probability is thus the probability of the input sequence being a motif after 1 replication. An alternative calculation can yield the probability of the input sequence being a motif after a specified branch length in the phylogenetic tree. This probability calculation uses the substitution rates estimated from the tree and also considers the equilibrium frequencies of the nucleotides. Probability scoring can either be performed in tandem with motif identification on the codons that encode the motif instance, which yields the robustness of that motif (and thus the probability of the motif being lost through mutations). Alternatively, probability scoring can be performed for a list of motif types over an entire sequence of interest thereby giving data on the probability of any motif evolving at any given position within the sequence. The algorithm performs a sliding window probability analysis over the input sequence (if longer than the regular expression motif definition) and determines a probability for each

position in the sequence and saves it in a list. The algorithm uses a nucleotide sequence as input alongside a motif regular expression and an optional time window which selects the branch length. Without a branch length time window, the per replication probability is the default. The mutation rate variables can also be adjusted in a text file as appropriate.

2.5.5. Analysis of motif evolution dynamics

The functional analysis of the evolutionary properties of motifs relies on having a tree with known ancestral sequence states at the nodes, a sequence of filtered motifs of interest associated with each sequence at each node and a protein alignment file. The user can create the list of motifs of interest from within the pipeline. To establish motif evolutionary trends, the presence or absence of each selected motif is first established for each node in the tree, accounting for alignment positions to accurately assess motif position and identity. To achieve this, motifs are compared with each other to create a set of unique motifs that appear one or more times and are associated with specific sequence positions in the alignment. Subsequently at each node in the tree each unique motif is assigned a set of properties such as presence, sequence, probability and whether it has recently evolved (i.e. was not present in its ancestor). Data can then be analysed quickly, providing detailed information about loss of conserved motifs, gain of new motifs, average number of descendants retaining the altered state (motif presence or absence) and the associated probabilities with these events. In addition, since the codon-level information is saved for each position of each motif in each strain, summary codon usage data and the specific mutations that cause gain or loss can be assessed as well. Motif details and sequences are quality controlled to ensure that false assignments of presence or absence do not occur due to missing sequences, gaps or ambiguously assigned nucleotides.

2.5.6. Motif evolution visualisation

The pipeline can automatically generate a range of figures and diagrams to facilitate analysis of the results and to further enable exploration of preliminary data. For exploration of motifs in different virus strains, figures of the phylogenetic tree can be generated with each sequence and associated sequence features illustrated adjacent to relevant nodes. This easily showcases presence and absence of motifs in strains of interest and is valuable to get an overview of branches and strains that have evolved or lost specific motif features. To analyse the probability landscapes of sequences in detail, a range of plots can be output that can showcase both sequence features, predicted disorder and transmembrane domains as well as existing motifs, and can layer any number of user chosen motif types and their probability landscape over the sequence. These figures allow the user to quickly find hotspots and regions where sequences are more or less likely to gain or lose motifs. To quantify gain and loss of motifs over evolutionary time, scatterplots can be output correlating expected frequencies of events given sequence probabilities with the observed frequencies of evolutionary events in the different strains over the tree. Observations of loss of known functional motifs can be compared over

time and over strains and correlated with the sequence probability of those motifs. For codon use and mutations, figures summarizing mutational outcomes are generated. These plots are useful to show evolutionary trends and any divergence from expected frequencies of events indicating selection and other forces playing a role in shaping their evolution.

2.5.7. Special considerations

Given the range of genome organisation in various viruses, many of them cannot simply be treated as a single coding sequence to a single protein. The pipeline allows for some special consideration for certain groups of viruses. Flaviviruses are encoded as a single polyprotein and NCBI sequencing files contain the whole polyprotein. The polyprotein matures into 12 individual proteins during infection through protease cleavage at known sites, also encoded in the NCBI files or available in literature. For the purposes of motifs, analysing proteins individually with defined N and C termini is more appropriate. Many motifs are only functional at termini, or conversely are only functional internally. Thus, for flavivirus analysis, the pipeline requires the user to input the start and end amino acid position for each protein in the polyprotein and they are then analysed as individual proteins automatically. For proteins in viruses that have multiple overlapping reading frames, these can be processed individually through a dedicated pipeline/function to generate two alignments both in correct reading frames. These can then be analysed separately as well.

2.5.8. Simulations

The pipeline also handles simulations and analysis of simulation data as data generated through the phylogenetic simulations take the same shape as actual biological data. Phylogenetic tree simulations are analysed in the same way generating the same output and figures as biological data. The pipeline acts as a wrapper to the program Pyvolve which performs the actual simulations and resulting data is fed through the pipeline and can be integrated with biological data and compared (Spielman & Wilke 2015). The simulations of motif evolution can be plotted alongside actual motif evolution provided evolution has been simulated over a tree of the same topology as the biological data (the expected input). These comparisons can provide insight into evolutionary pressures and events. Populations simulation data is handled differently as the parent-child relationship between each simulated virus sequence is tracked. Data from these simulations can be analysed through a separate pipeline where mutation frequencies and motif evolutionary patterns can be established. For population simulations a single sequence is the founder sequence, simulating infection by a single virus. This sequence is replicated for a number of generations simulating early stages of a viral infection (fast growth). Motif evolution patterns given different starting conditions can then be evaluated.

2.6. Discussion

In this chapter I have established a novel framework that can be used to explore the mutational and evolutionary dynamics of motifs. This framework incorporates three key principles for the first time in motif evolution research. Firstly, the use of detailed nucleotide-specific mutation rates to elucidate motif evolution has not, to my best knowledge, been addressed before. Instead, an amino acid centric approach has been used in the past due to limitations in available data. Secondly, I have incorporated the structure of the codon space for motifs to improve our understanding and expectations of common substitutions and predicting evolutionary outcomes. Thirdly, I have taken into account aspects of population genetics, which help delineate how different rates of mutational events for motifs can impact the fitness of sequences in the population, and how this impacts the outcomes of different motif-containing sequences.

This is a novel approach to studying the evolution of specific protein features, that is supported by several observations from previously published work in the fields of population genetics, evolution, virology and motif research among others. Some of these observations include: experimentally observed fitness effects from the mutational landscape of viral sequences (Burch & Chao 2000; Lauring *et al.* 2012); direct fitness effects from codon choice that has a higher probability of mutating into stop codons (Moratorio *et al.* 2017); sequence robustness benefitting survival in simulated populations of large sequences (Wilke *et al.* 2001); observations that motifs frequently evolve from single amino acid substitutions resulting in functional innovations (Davey *et al.* 2015); observations of the sensitivity to mutation of different codons in some influenza functional contexts (Plotkin & Dushoff 2003). Taken together, applying these principles to studying the evolutionary dynamics and propensities of motifs is a logical and intuitive continuation of these observations.

Using this approach, I have through simulations shown that different initial sequences greatly influence mutational outcomes of motifs. Gaining new motifs as a mode of functional innovation is important for viruses to regulate protein behaviour (Davey *et al.* 2011; Hagai *et al.* 2014). As shown here, the potential space available for new motifs is highly dependent on the mutation rate and the prior nucleotide sequence through codon choices. The simulations also indicated that already existing motifs are likely to have different fitness depending on codon choices, through the different mutational dynamics exhibited by the different codons that are allowed within a certain motif. Together, this suggests that exploring virus sequence evolution from a motif perspective using this framework could elucidate new aspect of viral function and evolution. Primarily it could allow us to predict the fitness of current motifs through codon use and thereby to infer selection pressures. It could also allow us to predict sequence hotspots that are primed to evolve new interaction sites and PTM sites thus changing the virus regulation within the host.

Given the importance of the nucleotide specific mutation rate in motif evolution outcomes in these simulations, the accuracy of any further analyses will rely on having accurate mutation rates. This is not currently available for most circulating viruses to the level required, however they have recently been published for influenza A which is part of the reason for how this analysis is now possible. However, it is also well established that the mutation rates vary even between influenza A strains. How much they vary is currently not known, but several known mutations influence the polymerase mutation rate, and measured rates in different strains from Pauly *et al*, (2017) show that some of the nucleotide mutation rates are significantly different. This further highlights how important this type of data is and it will add a lot of value to evolutionary analyses when it is expanded to include most known viruses.

Another important consideration to make when thinking about evolutionary dynamics of motifs in viruses is that selection will play a significant role in evolutionary outcomes. The different frequencies I see in these simulations are under neutral evolution, and while these frequencies will shape the sequence landscape that selection can act on, the fitness effects of motif gain and loss will likely play a larger role in the final strains that emerge. Having an understanding of the underlying frequencies is essential for accurate predictions of evolutionary trajectories, so together with models for selection of given features, accurate evolutionary predictions could be achieved.

A key discussion point is the accuracy of motif definitions. The predictions rely on the defined codon space which is based on the motif consensus. For many motifs the consensus is well established and experimentally validated, and there is a good understanding of the contribution of the different positions and amino acids. It will be of great value to establish this kind of understanding for more motifs than we have today, as it will drastically improve our ability to predict evolutionary outcomes. High throughput approaches such as phage display and IDR-screen are allowing us to define these motif properties (Davey *et al.* 2017; Ravarani *et al.* 2018). There are also some fundamental limitations for how selection can act on and shape the fitness of particular motifs. Depending on the motif consensus, different outcomes will be expected. For a consensus such as that for the SH3-binding motif, which has a defined core pattern of PxxP, there will be no fitness differences between the different codons from a motif perspective. This results from the fact that P is 4-fold degenerate and therefore all codons share the same mutational outcomes (Position 1, and 2 are the same and mutations in position 3 are silent). In addition, positions defined by a large number of residues such as [^P] (not P) or [LVIF] will share the majority of outcomes for the majority of codons and any fitness differences are likely to be negligible, except for a small number of codons.

In conclusion, the codon centric framework for motif evolution developed in this chapter has the potential for elucidating interesting evolutionary properties of motifs, and the possibility of predicting motif evolutionary trajectories. Motif evolution probability informs how organisms can increase the

frequency and thus fixation rate of functional motifs. There are some caveats to this, centred around the current limit of our understanding of motif definitions and accurate mutation rates, which will impact predictions. However, many motifs have well defined consensus sequences, and organisms including influenza have accurately established mutation rates for the different nucleotides. With this in mind and based on the outcomes of the simulations run in this chapter, I use this approach to investigate the evolutionary dynamics of motifs in circulating influenza A strains in the next chapter.

2.7. Materials and Methods

2.7.1. Datasets

High quality sequences were downloaded from the curated dataset used in an analysis of influenza evolution by Worobey *et al.* (2014). Their high quality sequence alignment of the influenza protein NS1 was the basis for building the phylogenetic tree used in the simulations. Their NS1 alignment consisted of 10982 sequences from all influenza A strains. It was constructed using the open reading frame for both NS1 and NS2 as they overlap. They aligned their sequence in MUSCLE v3.6 (Edgar 2004) and refined it manually. They also removed identical sequence.

To reduce the complexity and improve the speed of the tree for the simulation analysis I filtered the above alignment to only include NS1 from H1N1 strains. I also removed any sequences with continuous gaps or ambiguous positions where a definitive nucleotide had not been assigned. The final dataset consisted of 4569 sequences. The oldest sequence in the dataset was the 1918 strain and the most recent was from 2012. The motif used as the basis for the simulations ([RK][RK]x[ST][^P]xx) was defined in the ELM database (MOD_PKA_1) (Gouw *et al.* 2017). The consensus sequence defined there was used unaltered, and no alternative motif definitions were considered for the PKA motif for the purposes of these simulations.

Nucleotide mutation rates established by (Pauly *et al.* 2017) were used to calculate probabilities for the simulation experiments. The mutation rates used were the ones from strain A/Puerto Rico/8/1934 H1N1 (Table 2.5).

2.7.2. Building a phylogenetic tree

The sequence alignment outlined above was used to build a phylogenetic tree. To construct the tree the software RAxML was used (Stamatakis 2014). The command line argument used to build the tree was

[raxmlHPC-PTHREADS-AVX -T 8 -m GTRGAMMA -s ../Input-Alignment.fasta -f a -p 12345 -x 12345 -# 100 -n Outputfile].

It was run multithreaded using 8 threads on a local cluster. –m defines the substitution model GTR-Gamma, which is standard for nucleotide sequences. –f defines the algorithm RAxML uses, with “a”

being “a rapid Bootstrap analysis and search for the best-scoring ML tree in one single program run” (RAxML documentation).

Table 2.5. Experimentally determined mutation rates reported by (Pauly *et al.* 2017).

	Mutation rate (mutations per nt per strand replicated)[a]	
Mutation class	A/Puerto Rico/8/1934 H1N1	A/Hong Kong/4801/2014 H3N2
A to C	$1.5 \times 10^{-5} \pm 0.9 \times 10^{-5}$	$3.4 \times 10^{-5} \pm 1.0 \times 10^{-5}$
A to G	$2.0 \times 10^{-4} \pm 1.1 \times 10^{-4}$	$3.0 \times 10^{-4} \pm 1.5 \times 10^{-4}$
A to U	$1.8 \times 10^{-5} \pm 1.8 \times 10^{-5}$	$1.3 \times 10^{-5} \pm 0.3 \times 10^{-5}$
C to A	$7.7 \times 10^{-6} \pm 4.4 \times 10^{-6}$	$1.7 \times 10^{-5} \pm 0.9 \times 10^{-5}$
C to G	$5.1 \times 10^{-6} \pm 2.3 \times 10^{-6}$	$9.7 \times 10^{-6} \pm 7.8 \times 10^{-6}$
C to U	$2.7 \times 10^{-5} \pm 0.7 \times 10^{-5}$	$4.6 \times 10^{-5} \pm 1.6 \times 10^{-5}$
G to A	$3.1 \times 10^{-5} \pm 0.2 \times 10^{-5}$	$7.2 \times 10^{-5} \pm 1.1 \times 10^{-5}$
G to C	$5.4 \times 10^{-5} \pm 2.4 \times 10^{-5}$	$2.8 \times 10^{-5} \pm 0.7 \times 10^{-5}$
G to U	$3.5 \times 10^{-5} \pm 0.9 \times 10^{-5}$	$6.0 \times 10^{-5} \pm 1.6 \times 10^{-5}$
U to A	$1.4 \times 10^{-5} \pm 0.7 \times 10^{-5}$	$4.5 \times 10^{-6} \pm 1.8 \times 10^{-6}$
U to C	$2.3 \times 10^{-4} \pm 0.5 \times 10^{-4}$	$3.1 \times 10^{-4} \pm 1.2 \times 10^{-4}$
U to G	$3.5 \times 10^{-5} \pm 2.3 \times 10^{-5}$	$3.6 \times 10^{-5} \pm 2.3 \times 10^{-5}$
Overall[b]	1.8×10^{-4}	2.5×10^{-4}

[a] Arithmetic mean plus or minus the standard deviation calculated from at least three replicates.

2.7.3. Calculating motif probabilities and defining sequences

To determine a set of sequences to use for the simulations, I performed the probability calculations to identify the codons most and least likely to **evolve into** defined sets of codons and also codons most and least likely to **remain** within a given set of codons. Using the mutation rates as defined above, the probabilities were calculated by the following formula:

For the probability of an initial codon to evolve into the sequence space for e.g. Ser/Thr or Arg/Lys we can define a set of codons C, for e.g. Ser/Thr such that

$$C^{ST} = \{N_1, N_2, N_3 \mid AA(N_1N_2N_3) \in \{S, T\}\}$$

where N_1, N_2, N_3 are nucleotides at codon positions 1, 2 and 3. The probability of codon ATC to substitute into C^{ST} can thus be expressed as

$$P(C^{ST}|ATC) = \sum_{N_1,N_2,N_3 \in C^{ST}} P(n_1 = N_1|n_1 = A) * P(n_2 = N_2|n_2 = T) * P(n_3 = N_3|n_3 = C)$$

This produced a ranked list for codons giving their probability of evolving per replication. The codon probabilities for the different Ser/Thr outcomes are listed in Table 2.6.

Table 2.6. Calculated probabilities for all codons to gain Ser/Thr through mutation after a single round of replication given the above mutation rates.

CODON	PROBABILITY TO SER/THR
ATC(I)	0.000265
ATT(I)	0.000265
ATG(M)	0.00023
ATA(I)	0.00023
TTC(F)	0.00023
TTT(F)	0.00023
TTG(L)	0.00023
TTA(L)	0.00023
AAC(N)	0.000215
AAT(N)	0.000215
AGG(R)	0.000143
AGA(R)	8.70E-05
TGC(C)	6.80E-05
TGT(C)	6.80E-05
GCC(A)	6.60E-05
GCT(A)	6.60E-05
GCG(A)	6.60E-05
GCA(A)	6.60E-05
UGG(W)	5.40E-05
UGA(*)	5.40E-05
CCC(P)	3.47E-05
CCU(P)	3.47E-05
CCG(P)	3.47E-05
CCA(P)	3.47E-05
GGC(G)	3.10E-05
GGU(G)	3.10E-05
AAG(K)	1.50E-05
AAA(K)	1.50E-05
UAC(Y)	1.50E-05
UAU(Y)	1.50E-05
UAG(*)	1.50E-05
UAA(*)	1.50E-05
CGC(R)	7.70E-06
CGU(R)	7.70E-06
GUC(V)	1.63E-08
GUU(V)	1.63E-08
GUG(V)	1.52E-08
GUA(V)	1.52E-08
CUC(L)	8.25E-09
CUU(L)	8.25E-09
CUG(L)	7.98E-09
CUA(L)	7.98E-09
GAC(D)	7.19E-09
GAU(D)	7.19E-09
GGG(G)	6.32E-09
GGA(G)	4.59E-09
CGG(R)	2.56E-09
CGA(R)	2.13E-09
CAC(H)	2.06E-09
CAU(H)	2.06E-09
GAG(E)	9.91E-10
GAA(E)	9.90E-10
CAG(Q)	5.21E-10
CAA(Q)	5.21E-10

I did the same calculation for the Arg/Lys position and [^P] position and picked sequences in a range of predicted probabilities. The sequences used are shown in Table 2.7.

Table 2.7. The sequences used for motif gain and loss analysis.

Protein Sequence Gain	Nucleotide Sequence Gain
RRGINGA	**AGG** AGG GGT **ATC** AAC GGT GCA
RRGRNGA	**AGG** AGG GGT **CGC** AAC GGT GCA
RRGVNGA	**AGG** AGG GGT **GTC** AAC GGT GCA
RRGQNGA	**AGG** AGG GGT **CAA** AAC GGT GCA
WRGINGA	**TGG** AGG GGT **ATC** AAC GGT GCA
Protein Sequence Loss	**Nucleotide Sequence Loss**
RRGTNGA	**CGA CGA** GGT **ACC AAC** GGT GCA
RRGSLGA	**AGG AGG** GGT **AGT CTT** GGT GCA

2.7.4. Simulating evolution using pyvolve

To simulate sequence evolution over the tree determined above, the software Pyvolve was used (Spielman & Wilke 2015). The RAxML constructed tree was used as input alongside a founder sequence. Each sequence constructed above was used as the founder in subsequent pyvolve simulations. And each sequence was simulated 1000 independent times. The program requires a user defined custom_mu parameter which determines the relative substitution frequencies for the different nucleotide combinations. The custom_mu was created by determining the ratio of mutational frequencies from the experimentally determined mutation rates, with G-to-A arbitrarily defined as 1.0 and the other rates defined in relation to that (Table 2.8).

Table 2.8. Relative mutation rates for all twelve mutation classes relative to G to A.

A to T	0.58
A to C	0.48
A to G	6.45
C to A	0.25
C to T	0.87
C to G	0.16
T to A	0.45
T to C	7.42
T to G	1.13
G to A	1.00
G to T	1.13
G to C	1.74

The frequencies of each nucleotide were assumed to be fixed in the population and were also required as input. The NS1 nucleotide frequencies based on the RAxML phylogeny were used (A=0.32, T=0.20, C=0.24, G=0.24). The simulation produced output data in the form of alignments in fasta for-

mat. The genotypes of each evolved sequence were determined, translated and analysed for motif consensus. Summary data including frequency of motif evolution, trees where the motif evolved and average number of strains with the motif per tree was compiled.

2.7.5. Analysis of phylogenetic simulation data

To determine when motifs evolved or were lost I used DendroPy v.4.4.0 in python to traverse the tree and correlate each node with the simulation alignment (Sukumaran & Holder 2010). At each node I translated the sequence and determined if it was a motif or not based on the PKA consensus. For each tree I determined the number of nodes with the motif, how many trees the motif emerged in at least once, and how many independent evolutionary events took place. To assign independent evolutionary events I determined the number of times an ancestral sequence lacking the motif had a descendant with a mutation that lead to a motif. Since the theoretical number of independent evolutionary events possible is the total number of sequences without the motif, it is important to consider the number of strains that inherit the motif and exclude them from the independent count. This means that if a lineage gains a motif early on and all descendants have the motif, there are no opportunities in that lineage to gain a motif again, which in turn can affect the conclusions if not considered. For motif gain this has very little impact as motifs are gained between 0.004 times to 2 times per tree on average, and between 5-131 strains out of 4569 have the motif on average. For motif loss, which is much more prevalent, this impacts the independent loss estimate, and it is important to normalise the number of independent loss events with the number of possible loss events given that only a subset of nodes in the tree have the motif. For this reason, in the independent loss calculation, I divide the number of independent losses observed with the number of strains with the motif in the tree. If an early strain loses the motif 3 independent times such that only 1000 out of 4569 strains ultimately had the motif, and another strain loses the motif 10 times but still 4000 out of 4569 strain had the motif in total I would compare 3/1000 with 10/4000 as there were more possible opportunities for loss in the second strain. This provides a rough normalisation that should be representative when averaged out over 1000 trees, however it does not consider branch length which could skew the outcomes somewhat.

2.7.6. Simulating quasispecies evolutionary dynamics

Quasispecies simulations were run using a custom simulation algorithm. A founder sequence was replicated with nucleotide specific mutation rates as defined above. Sequences were replicated between two populations, simulating actual influenza replication behaviour (Samji 2009). The sequences were replicated using the probabilities as before (Table 2.5). Sequences were replicated between viral RNA (vRNA) and complementary RNA (cRNA), doubling each replication until the cRNA population passed 300 sequences, reflecting the amount of cRNA usually seen for each gene segment in cells (Frensing *et al.* 2016; Heldt *et al.* 2012). After the cRNA population cap, new vRNA sequences were generated from the cRNA pool until the vRNA population passed 10000 sequences, again reflecting

the number of vRNAs average cells produce per gene segment during viral infections (Frensing *et al.* 2016). The final sequence analysis was performed on the 10000-12000 vRNA sequences. The sequences used for this analysis were the same ones as were used in the pyvolve simulations, which had been calculated to show different evolutionary properties as described previously. As before, all the genotypes in the population were counted and translated to determine if they matched the motif consensus. The proportions of the different motif genotypes within the whole population were then determined and quantified. To determine the mutational outcomes of individual codons, quasispecies simulations were run in the same way. Mutation outcomes were quantified independently for each codon and plotted.

Chapter 3

The evolution of motifs in influenza A is shaped by motif centric codon bias

3.1. Overview

Influenza is a negative strand RNA virus with a very high mutation rate. It causes yearly influenza outbreaks and has historically been responsible for many of the most lethal pandemics ever known (Potter 2001). Influenza A is a virus that has been studied extensively due to its impact on human history. We therefore have a long record of strain and sequence data and an in depth understanding of its biology.

Influenza A uses many host-like short linear motifs throughout its infectious cycle. It utilises the host machinery for post-translational modifications, binding interactions and transport to infect and interact with the host, all through convergent evolution of host-like motifs (Zhao *et al.* 2017). Some of these motif-based functions are known to evolve and change between strains but our knowledge of the evolution of these motifs is currently limited. An in depth understanding of the evolutionary dynamics of motifs would be an important step towards a more complete model for understanding viral mechanisms, spread, pandemics and infectious effects when new strains emerge. In addition to pointing to new possible therapeutic targets, it can also provide a deeper understanding of their potential for rewiring virus-host interactions, thereby highlighting mechanistic differences in infection between strains. Insights into motif evolutionary dynamics in viruses can also act as a general model for how motif evolution behaves in other contexts, in particular where mutation rates are high and populations are large and heterogeneous. Implications for our understanding of motif evolution in viruses could thus be applied to other pathogens and potentially to cancer cells as well. In addition, studying motifs in a good model system can allow us to identify fundamental properties about motif evolution that would be difficult to observe in more slowly evolving systems.

In this chapter I have used the methods and framework developed and described in chapter 2 as a basis for studying the evolution and sequence properties of motifs in influenza proteins. I compare the observations from the simulations in chapter 2 to those of evolution of known motifs in influenza. I find that prior sequence impacts motif evolution in the viral strains and that there is a distinct bias towards

codons that are more robust to mutations in key conserved motifs, most notably phosphorylation and glycosylation sites. These results point to the fact that motif-based codon bias can act as sequence signals that can be used to better predict if motif patterns observed in a protein sequence are likely to be functional. They can also provide insights into how key functional motifs have fluctuated (i.e. become gained or lost) during the evolutionary history of influenza.

3.2. Introduction

The biology of influenza makes it an interesting model system to use to study motifs and motif evolution. Firstly, influenza interacts with several key cell pathways using a range of motifs. Secondly, the availability of strains from many different hosts and backgrounds, including many host shifts, allows for interesting comparisons of the biology and independent evolution of sequence properties of the virus. Thirdly, it is one of the most well studied and well documented viruses; samples have been collected and frozen since the early 1900s and have since been sequenced. That means it is possible to create an accurate sequence history from historical data and track changes and mutations over time. Fourthly, it is a virus with wide ranging health implications both historically and presently for humanity. Through world-wide pandemics it has killed several hundred million people in the 20^{th} century alone. Through the seasonal influenza, it infects millions of people every year, representing a significant burden to the economy and society, and kills hundreds of thousands of people throughout the world. Finally, influenza has a mutation rate on the order of $2*10^{-4}$ and mutations per nucleotide per replication. It is replicated between 10000-15000 times in each cell leading to an enormous amount of mutational sampling, on average 2-3 mutations per nucleotide. Taken together, influenza is a highly useful model that can allow us to elucidate properties of motif function and evolution in a biologically relevant system where high mutation rates and variability meet strong selection pressures.

3.2.1. The biology of influenza A

By firstly delineating the lifecycle and functional context of influenza A, and the sequences and proteins that make up the virus, we can get a better understanding for the role of motifs in the infection. Influenza A infects the respiratory tract of humans, birds, pigs and several other mammals. It is commonly transmitted between different species, and all known pandemics have been the result of animal to human transmissions. Examples of this include bird to human in 1918 (H1N1), bird to human in 2005 (H5N1) and swine to human in 2009 (H1N1) (Taubenberger & Morens 2010). There are also seasonal variants that circulate yearly within the human population specifically, and evolve rapidly, requiring new vaccinations each year (Petrova & Russell 2017). Influenza A encodes 11 proteins from 8 separate gene segments of a total size of around 13500 nucleotides depending on strain (Bouvier &

Palese 2008). The proteins Haemagglutinin (HA), Neuraminidase (NA), nucleoprotein (NP), polymerase basic protein 2 (PB2) and polymerase acidic protein (PA) are encoded by unique gene segments (Figure 3.1). The six remaining proteins are encoded by three gene segments from overlapping reading frames (Figure 3.1). Matrix protein 1 (M1) and matrix protein 2 (M2) share overlapping reading frames, and M2 is additionally spliced by the host splicing machinery. Non-structural protein 1 (NS1) and non-structural protein 2 (NS2) share a similar gene architecture to M1/M2, where NS2 is spliced and the reading frames overlap. Finally the gene segment encoding polymerase basic protein 1 (PB1) also encodes a short peptide, PB1-F2, that plays a role in apoptosis (Chen *et al.* 2001).

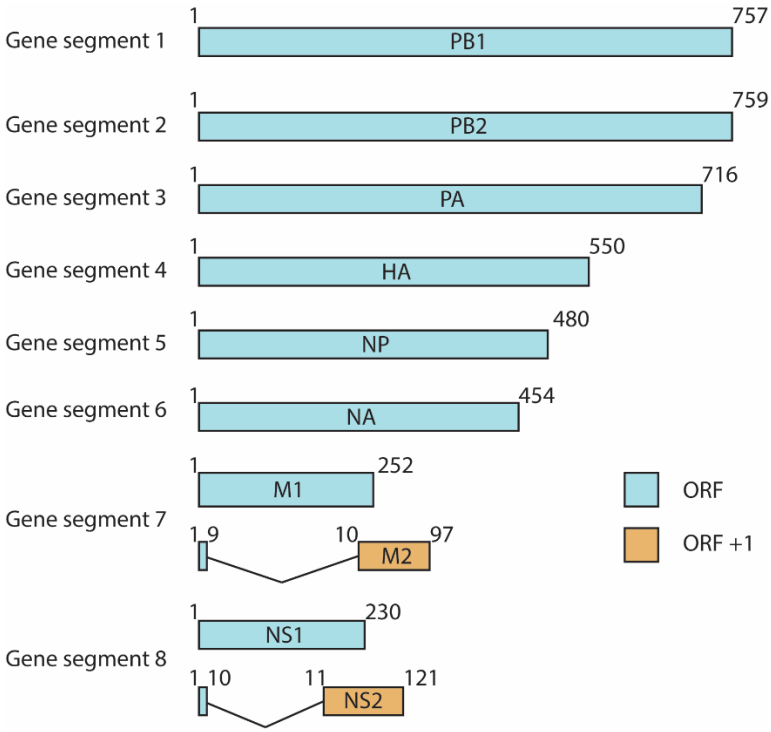


Figure 3.1. The gene architecture of influenza A. From strain A/Puerto Rico/8/34. The 8 gene segments encode the above 10 proteins and an additional short peptide (not shown). Segments 7 and 8 encode two proteins each, M1+M2 and NS1+NS2 respectively. M2 and NS2 are spliced and encoded partially by an alternative overlapping ORF.

The virion itself consists of a protein coat made up of Haemagglutinin (HA) and Neuraminidase (NA) and structural and functional variants of these proteins are used to classify viral strains such as H1N1, H5N1 and H3N2 (Figure 3.2) (Bouvier & Palese 2008). HA is more numerous in the coat, with around 500 copies per virion, whereas NA has only about 100 copies, however there is a large diversity in viral coat assembly (McAuley *et al.* 2019; Vahey & Fletcher 2019; Varghese *et al.* 1983). The protein M2 also forms part of the coat, but does not interface with the host during infection. It forms proton channels and exists in very low copy numbers of about 12 copies per virion (Duff & Ashley 1992). In addition, M1 forms an important part of the virion coat as a whole through interacting with both coat protein and vRNPs at the membrane interface on the inside of the virion (Figure 3.2) (Shilova *et al.*

2017). It is essential to the structural integrity of the coat and exists in high copy numbers of about 3000 copies per virion (Hutchinson *et al.* 2014).

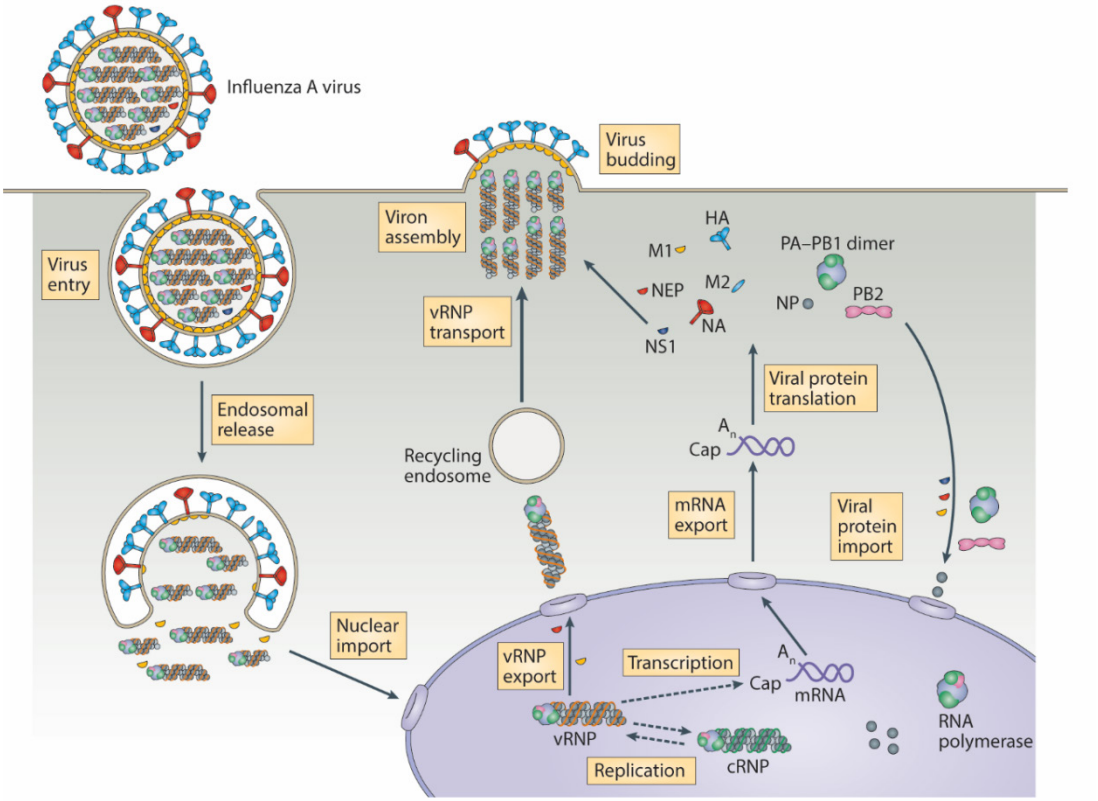


Figure 3.2. The viral infection cycle. Shows the general means of viral entry, nuclear import, vRNP replication, virion assembly and budding. At all stages, the viral particles interact with host proteins and rely on host pathways and functions. Figure reproduced from te Velthuis & Fodor, (2016) with permission from Springer Nature, Macmillan Publishers Ltd.

Upon infection the virus binds cellular receptors and gets incorporated into the cell. The vRNPs are released and subsequently transported to the nucleus (Wu *et al.* 2007). The vRNPs are made up of the 3 subunits of the viral RNA polymerase PB1, PB2 and PA as well as Nucleoprotein (NP) and contains one of the eight negative strand RNAs that make up the viral genome (Baudin *et al.* 1994; Noda *et al.* 2006). Upon entering the nucleus, the RNA is transcribed to generate the viral products, and replication of the viral RNA also concurrently begins (Samji 2009). The viral polymerase achieves both transcription and replication (Eisfeld *et al.* 2014; Fan *et al.* 2019). Replication of the negative strand vRNA results in a positive strand RNA product (cRNA) which gets replicated to create more vRNA (negative strand). The pool of negative strand vRNA also acts as templates for mRNA synthesis by the same polymerase. These replication cycles create a pool of vRNAs and a pool of cRNAs, that when bound by the nucleoprotein and polymerase are referred to as vRNPs and cRNPs. Synthesised HA, NA and M2 assembles on the cell surface after moving through the ER and Golgi (Figure 3.2). M1 assembles from the cytosol and stabilises the coat proteins. NS1 acts to rewire cell behaviour but does not directly take part in viral assembly. NP, NS2, PB1, PB2 and PA assemble around the new vRNAs

and get exported from the nucleus and transported to the cell membrane where the 8 vRNPs, containing one each of the vRNAs, assemble with the coat proteins and bud off as a new virion (Rossman & Lamb 2011). Taken together, the infection cycle of the virus involves repeated interactions with the cell machinery at many different levels, opening the question of how influenza motifs and motif evolution impact the different stages of the infection processes.

3.2.2. Motif use and relevance in influenza

Influenza relies on several classes of motifs throughout the infection cycle and several such motifs have been well studied and characterised. Within the cell, influenza proteins rely on phosphorylation, SUMOylation, transport and binding motifs to perform their range of functions (Han *et al.* 2014; Li *et al.* 2015; Sierra *et al.* 1998; Wang *et al.* 2013; Zheng *et al.* 2015). Influenza replication occurs in the cell nucleus, so initially the vRNPs are localised to the nucleus using nuclear localisation sequences in NP (Wu *et al.* 2007). In addition, newly translated proteins involved in replication and in regulating host processes such as transcription in the nucleus all have to be transported there successfully. NP, PB1, PB2, PA, NS1 and M1 all contain nuclear localisation signals as well as nuclear export signals (Li *et al.* 2015). The coat proteins rely on glycosylation for both cell fusion and evasion of the immune system (Tate *et al.* 2014; Tsuchiya *et al.* 2002). In addition, HA requires protease cleavage to function which it achieves through a host-like protease site that is recognised by a host protease (Bertram *et al.* 2010).

NS1 is one of the key proteins regulating host cell functionality through interactions with the host machinery. It contains several binding motifs to achieve this regulation including PDZ, SH2 and SH3 binding motifs that are important for interacting with some of the host anti-viral machinery (Hale *et al.* 2008). Targeting of host innate immunity systems, including IRF3 and RIG-I associated proteins, is essential for viral evasion of the cell defence machinery (Bottermann & James 2018; Krug 2015; Lin *et al.* 2007).

Many of these motifs, in particular PTM sites, have been characterised in only a small number of strains but show considerable variation across strains. Understanding the evolutionary dynamics of motifs in the context of influenza will not only improve our understanding of influenza infections, mutational dynamics and selection pressures, but will also act to improve our general understanding of how motifs evolve and what role motif evolution can play in the complex changes to molecular interactions that happen during infection.

3.2.3. Role of glycosylation motifs in influenza

Glycosylation sites are some of the most important motifs for the external coat proteins and they are consequently some of the most evolutionary dynamic. Changes to HA – and to a lesser extent NA – are relevant for viral evasion of host antibodies (Tate *et al.* 2014). It has been shown that different

strains have different sites glycosylated, and that many of these sites are poorly conserved (Altman *et al.* 2019; Alymova *et al.* 2016). In the past, a study that looked at where new glycosylation sites emerge found that the majority evolve through single point substitutions (Igarashi *et al.* 2008). This suggests that new site prediction could be improved using the methods developed in chapter 2. Improving our understanding of the evolutionary dynamics of glycosylation sites would be of great importance in dealing with influenza infections, and could play a role in predicting evolutionary events associated with changes to influenza glycosylation patterns, including adaptation during vaccination efforts.

3.2.4. Role of phosphorylation sites in influenza

Phosphorylation sites are another prevalent, and functionally important, motif type in influenza (Hutchinson *et al.* 2012). Phosphorylation is used to regulate transport and binding of viral proteins, both between the different viral proteins and between viral-host pathways. Phosphorylation regulation has been particularly characterised in the protein NP. Here phosphorylation sites near the N-terminus are important for regulating the interaction with host importins, thereby controlling nuclear shuttling (Zheng *et al.* 2015). Phosphorylation sites at the NP oligomerisation surface on the other hand, regulate NP assembling around the vRNP (Mondal *et al.* 2017). Overall, differential phosphorylation is hypothesised to be important for timing the stages of viral infection (Mondal *et al.* 2015). It has also been found that phosphorylation sites evolve and change on a rapid timescale and are different between many strains (Petri *et al.* 1982). However, details of their different functionality have not yet been studied, and in general the molecular functionality and specific timing of the majority of phosphorylated sites even in the best characterised strains are mostly unknown.

3.2.5. Mutation rate and quasispecies dynamics in influenza

In chapter 2 I introduced the concept of the viral quasispecies and highlighted its role in viral fitness. To reiterate, influenza has some of the highest known mutations rates in the viral world, which results in a highly diverse population of sequences during infection. The mutation rate between each of the 12 nucleotide pairs have been determined accurately through experiments for 2 strains of influenza A (Pauly *et al.* 2017). On average, during each full genome replication of influenza A, at least 2-3 new mutations are expected and 10000-15000 replication events occur in a single infected cell (Frensing *et al.* 2016). The resulting quasispecies thus has a very high sequence diversity in each cell, and in each host. This sequence diversity is important for the fitness of influenza, and impacts the sequence space and mutational outcomes in each viral population as has been noted in a number of experiments. As established in chapter 2, this is predicted to have fitness implications for motifs both on the single cell scale and the larger population scale. For key functional phosphorylation sites and binding sites, motif loss would be likely to result in reduced fitness and thus should be minimised in evolution. In addition, the high mutation rate would allow sampling of some motifs more frequently suggesting that motifs

that provide a gain in fitness would be likely to emerge and be fixed if sampled with enough frequency. More frequent mutations can thus have significant fitness impact in the population given the functional outcome of the mutation.

3.2.6. Codon use and fitness in influenza

Given the results from the chapter 2 simulations, the hypothesis is that influenza codon space will impact its motif evolution dynamics. General codon choice and codon bias has been studied previously in influenza. A previously reported observation is that codon usage in influenza reflects the general codon usage in the host (Wong *et al.* 2010). This is thought to allow influenza to have more efficient translation as a consequence of differential tRNA expression as well as potentially different rates of translation for different codons.

A 2003 study by Plotkin *et al.,* looked specifically at the codon choice in HA antigenic sites in influenza A (Plotkin & Dushoff 2003). They found that codons in antigenic regions were chosen to increase missense mutations. Less degenerate codons were used more frequently as they can more easily mutate into different amino acids. They suggest the high substitution frequency at these sites are maximised for immune evasion. As discussed previously, studies have also shown the importance of codon choices in influenza through experimentally mutating multiple codons within the genome (Le Nouën *et al.* 2019; Moratorio *et al.* 2017). This generally led to attenuation of the virus and fitness decline in the majority of cases.

To date, the influence of codon choices on the evolution of specific functional units, such as motifs, have not been investigated. Despite extensively studying relative codon usage and codon bias in influenza and other viruses, we still lack a full understanding of the underlying reasons for codon bias in many contexts (Kumar *et al.* 2016; LaBella *et al.* 2019). The observations thus far suggest however that codon choice and the potential variable sequence outcomes given different initial codon conditions are crucial for influenza infectious fitness (Lauring *et al.* 2012). Teasing apart the more detailed molecular causes and consequences of these fitness effects would be a very important step towards understanding how evolution of fast evolving RNA viruses changes the functions and interactions of the virus over time.

3.2.7. Chapter summary

In this chapter I establish what is currently known about motifs in influenza and then explore the evolutionary dynamics of these known motifs. I have determined the detailed codon choices and mutations that occur over time through the use of high-resolution sequence data and phylogenies. I find that codon choices impact the rate of evolution of specific motifs. In particular, for phosphorylation and glycosylation sites, the codons used are optimised to maximise fitness of the virus through the robustness of the motif. To counteract the loss of key motifs influenza strains tend to favour robust codons at

these locations to reduce the likelihood of mutations causing motif loss. I also find different codon preferences for glycosylation motifs between HA and NA proteins which suggests that motif robustness can be modulated in response to specific context requirements.

3.3. Results

3.3.1. Mutation rates and codon choice influence codon mutational outcomes in influenza sequences

In order to determine if the evolutionary outcomes seen in chapter 2 are reflected in nature I have looked at the sequence history of influenza A. I wanted to assess to what extent the per-replication mutation rates inform the codon-level substitutions for different codons in influenza, as interesting differences were observed for codons in the simulations in chapter 2. If actual influenza sequence substitutions show similar biases and trends for certain codons to mutate into motif-related codon spaces more frequently, it would be a strong indication that the model described in chapter 2 is relevant. In other words, it would suggest that the mutational biases impact codon substitutions, and they are not solely shaped by selection and fitness. This would mean that the probability of events based on mutation rates could prove useful in improving our understanding of motif evolutionary dynamics in viruses. Two opposing scenarios can be generally envisioned. Under a hypothesis where the fitness difference between genotypes is very large, the mutation rate and codon space would have little bearing on the final outcome, as the most fit genotype will dominate in the population as it arises. On the other end of the spectrum, where small fitness differences separate many genotypes, the biggest driver for genotype fixation would be the population size of different genotypes (Sironen *et al.* 2008; Wilke 2003). By extension this would lead more frequent mutational outcomes to dominate the evolution of sequence features.

To explore on which side of the spectrum influenza sequence evolution falls, I created gene specific phylogenies for the influenza gene segments. The python software TreeTime was used to create time-informed phylogenies with ancestrally reconstructed sequences at intermediate nodes (Sagulenko *et al.* 2018). TreeTime was developed specifically for viral phylogenies where sequence data exists for strains that were active over the past decades. A more accurate phylogeny can be determined through accounting for the year at which the strain was circulating. This approach allowed me to reconstruct an accurate strain timeline and ancestry and thereby to infer reliable directional mutation events and track changes to sequences over time. For this analysis I specifically used the genes encoding NP, NS1 and M1. The number of strains in the alignment for each were 9592, 8663 and 9921 respectively. After sequence reconstruction at each node the NP alignment contained 17564 aligned sequences, the NS1

alignment 15834 and the M1 alignment 18441. TreeTime uses a maximum likelihood approach to infer divergence times of the strains and uses a general time reversible substitution model estimated from the alignment for sequence reconstruction (Sagulenko *et al.* 2018). Overall, because of the short distances between influenza sequences and large amount of sampled sequences across time, this is expected to result in accurate sequence reconstruction comparable to the best available, more computationally expensive methods. Using this data, I thus compared all amino acid and codon positions over time to observe the frequency at which different substitutions have taken place. I traversed the phylogenetic trees and identified all instances of codon substitutions between ancestral and descendant nodes. Since the different codons have different prevalence across the sequences, I normalised the number of observed codon substitutions by the frequency of the codon in the record. This allowed me to determine how frequently the different codons substitute into the codons encoding serine or threonine across influenza NP, NS1 and M1, which could in a motif context lead to the emergence of a new phosphorylation site (Figure 3.3).

There is a striking overlap between the codons that were highly predicted to evolve into serine or threonine codon space in chapter 2, and the codons observed in this dataset to frequently evolve into Ser/Thr (Figure 3.3 and Figure 3.4). The expected frequencies of different mutational events based on the experimentally determined mutation rates are clearly reflected in the actual evolutionary substitutions that occur. For example, AAC (N) more frequently substitutes to AGC (S) than ACC (T) because the A-to-G mutation rate is higher than the A-to-C rate. Thus the combined effect of accessible codon space from a particular starting codon, along with the mutation rates of the relevant nucleotides appears to determine the frequency of the different outcomes, which in large part shapes the overall observed rate of substitutions in the record. This results in large variations in how frequently different codons evolve into serine or threonine with implications for sequence trajectory and predictability of different outcomes.

Comparing the codon evolution predictions based on the probabilities as outlined in chapter 2, with the outcome of this analysis suggests a balance between the impact of mutation rate and selection through preservation of the physicochemical properties of the amino acids in determining the final substitutions. The amino acids (and codons) with high predicted probabilities of evolving into Ser/Thr include isoleucine (ATC, ATT, ATA), phenylalanine (TTC, TTT), asparagine (AAC, AAT), alanine (GCC, GCT, GCG, GCA), proline (CCC, CCT, CCG, CCA), glycine (GGT, GGC) and arginine (AGG, AGA). Comparing these to the results here of the top codons that frequently get substituted into Ser/Thr in the evolutionary record, proline, isoleucine, alanine, glycine and asparagine are by far the most frequent. Isoleucine is seen as being less frequently mutated to Ser/Thr than predicted, which is likely caused by the change from hydrophobic to more hydrophilic/neutral. This is probably also the reason why arginine and phenylalanine, despite being predicted to frequently mutate into Ser/Thr, very rarely do.

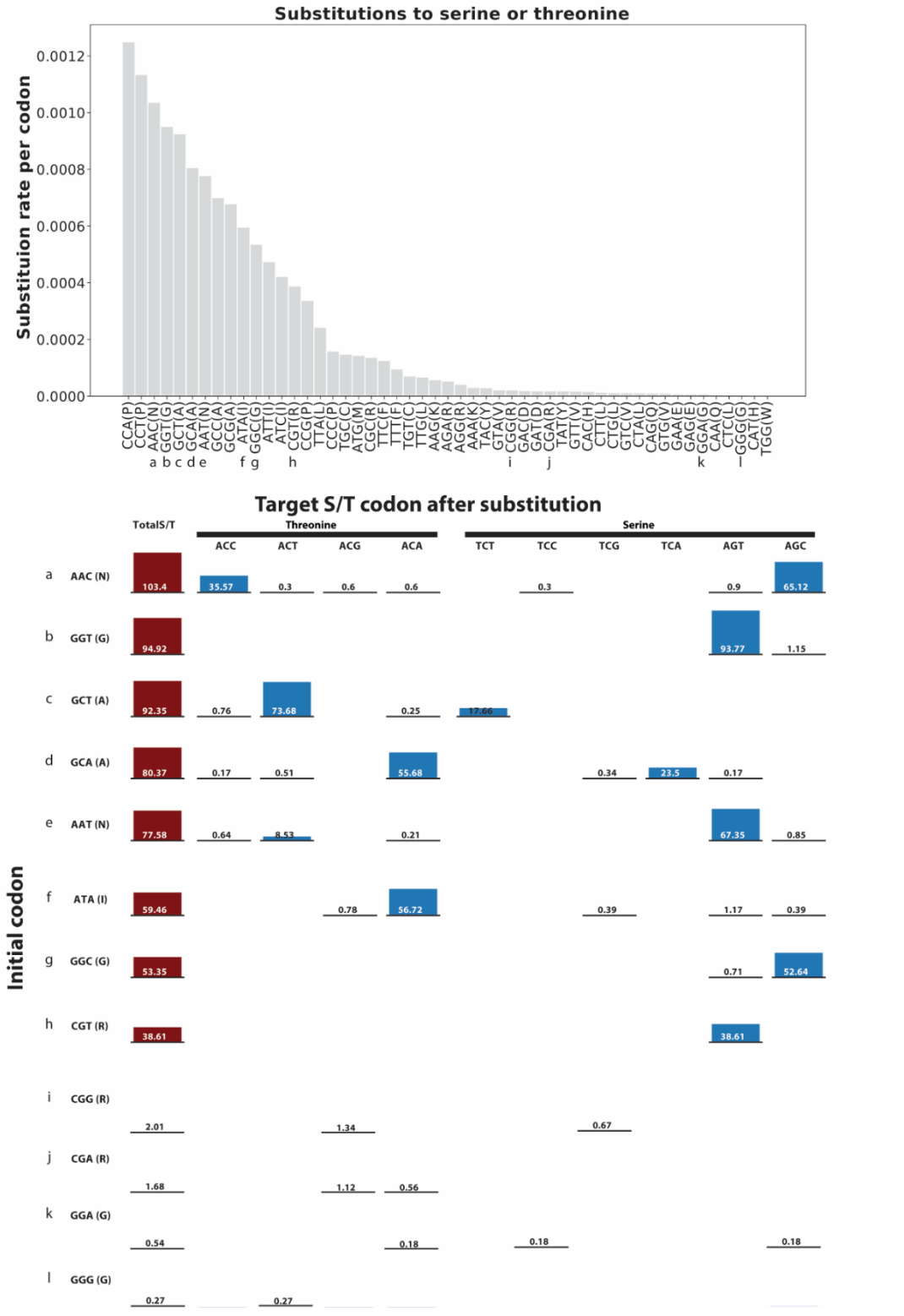


Figure 3.3. The substitution landscape of influenza codons to serine or threonine. The numbers denote the number of substitutions per 100000 codons. The heights of the red bars are internally normalised, and the height of the blue bars are internally normalised. Codons with more mutational paths and where mutation rates are high evolve Ser/Thr more often in the record. The rate of serine or threonine is also largely determined by mutational probability. Some amino acids such as arginine and glycine have both codons with high substitutions and with very low substitutions, illustrating the difference the probability makes and the enhanced detail from codon level analysis.

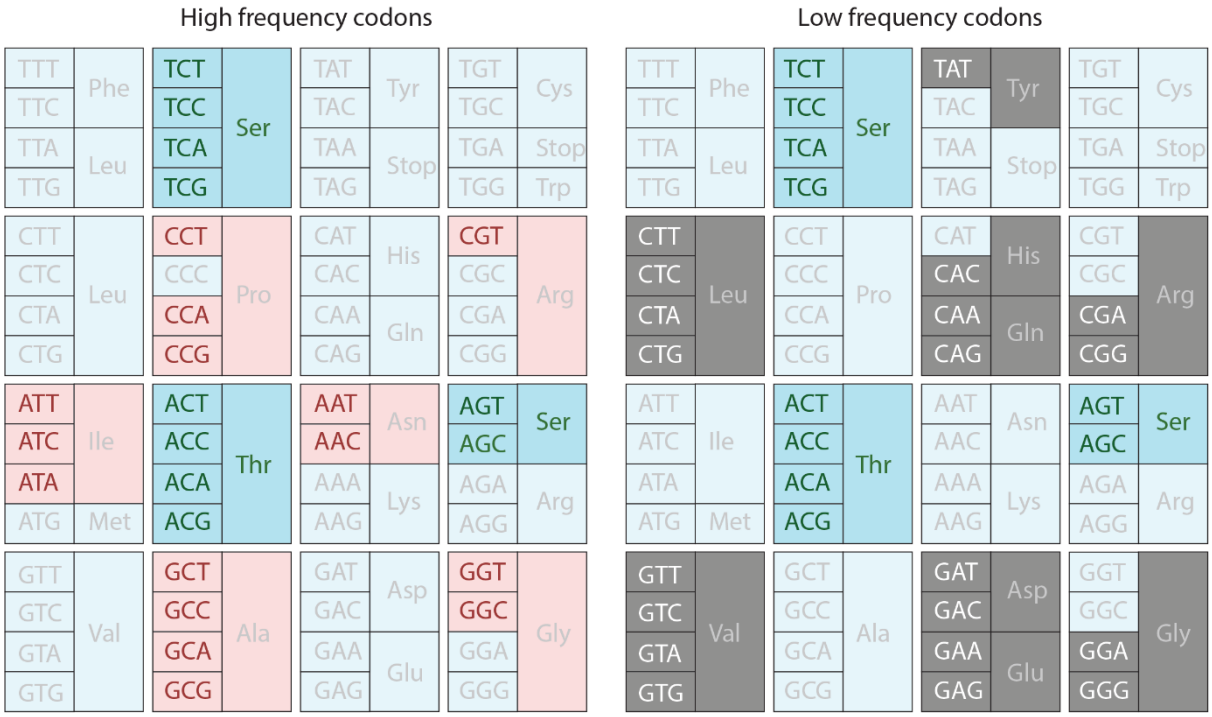


Figure 3.4. High and low substitution codon spaces to Ser/Thr. Most common codons to substitute to serine or threonine (left) and least common codons to substitute (right). This figure shows how the codon space largely determines the mutational outcomes. The most frequently substituted residues to Ser/Thr in influenza record are highlighted on the left and the least substituted on the right.

I also compared the codons that are predicted to very rarely mutate into Ser/Thr with the data from the evolutionary record. Here the results are even more striking. Predicted residues and observations here, agree very closely by almost never being substituted. These residues include arginine (CGG, CGA), glycine (GGG, GGA), histidine (CAC, CAT), aspartate (GAG, GAA) glutamate (GAC, GAT), valine (GTC, GTT, GTG, GTA) leucine (CTC, CTT, CTG, CTA) and glutamine (CAG, CAA). Interestingly, because of the shape of the codon space for Ser/Thr, both glycine and arginine have some codons that are among the most likely and frequent to mutate and some codons that are among the least likely to (Figure 3.3). This observation further illustrates the importance of considering mutation rate and codon space when analysing evolutionary dynamics and substitutions. When considering only amino acid space the difference between codons in arginine would be overlooked despite the arginine to serine/threonine mutational outcome clearly depending on which codon is used.

Taken together, these observations suggest that incorporating specific mutational rates and factoring in the relevant codon space can help explain patterns of evolution and that some codons are more evolvable than others from the perspective of motif outcomes. An AAC (N) codon will more frequently evolve a new Ser/Thr phosphorylation site than a GGC (G) codon will, all else being equal, meaning sites with asparagine that are encoded by AAC are more evolvable for new Ser/Thr motif sites.

I also looked at the robustness of different codons in the record, i.e. which codons in a given set most frequently substitute away from that set and which are less frequently substituted. Here the results

from influenza history also agree with those that were predicted by the model (Figure 3.5). Looking at Ser/Thr again in this context it becomes clear that TCT is among the least substituted codon in general within this space, with threonine codons ACA and ACT and serine codons AGT and AGC being the most substituted. This reflects the predictions from the chapter 2 simulations. The different codons here also have very different substitution outcomes with AGC (S) and AGT (S) most frequently substituting into asparagine whereas the remaining serine codons predominantly substitute into alanine and proline. By contrast, ACC (T) and ACT (T) to asparagine substitutions are much rarer, despite also being one point mutation away. Instead, threonine codons predominantly substitute into alanine and to some extent isoleucine.

This analysis clearly shows that the mutation rate difference between different codons and amino acids significantly shapes the mutational outcomes and directions of influenza sequences on a relatively short timescale. Taken together with the motif-centric codon space, this illustrates how much the sequence space affects the evolutionary trajectory and how much of a potential impact it can have on the evolutionary dynamics of PTMs and other motifs. To further delve into the dynamics of motif evolution in influenza in the following sections I will look at the evolutionary characteristics of functional motifs and their amino acids.

3.3.2. Known functional motifs in influenza A and their evolutionary conservation

The observation made in the previous section suggest that motifs in influenza will be subject to differences in loss and gain as a consequence of codon choice. In order to explore these evolutionary dynamics of motifs in influenza, I first set out to identify known and previously characterised functional motifs in the various influenza proteins. To this end I performed a literature review and collected high confidence functional motifs. These motifs form the basis for the majority of analyses in the subsequent sections of this chapter and include functional phosphorylation sites by a range of kinases, glycosylation sites, nuclear localisation and export motifs, interaction motifs and SUMOylation sites, covering the majority of motif classes and a number of eukaryotic motifs. In the majority of instances these motifs have been characterised both by sequence analysis and experiments and their functional importance for influenza established. Based on the availability of motif data, and for time considerations I have limited the analyses in this chapter to the proteins NS1, NS2, NP, M1, M2 and HA (specifically strains H1 and H3) and NA (strains N1 and N2) as these are the most important seasonal strains and also have the largest datasets. I have excluded the three polymerase proteins as they have very few well-studied motifs and are very expensive time-wise to analyse due to their sizes.

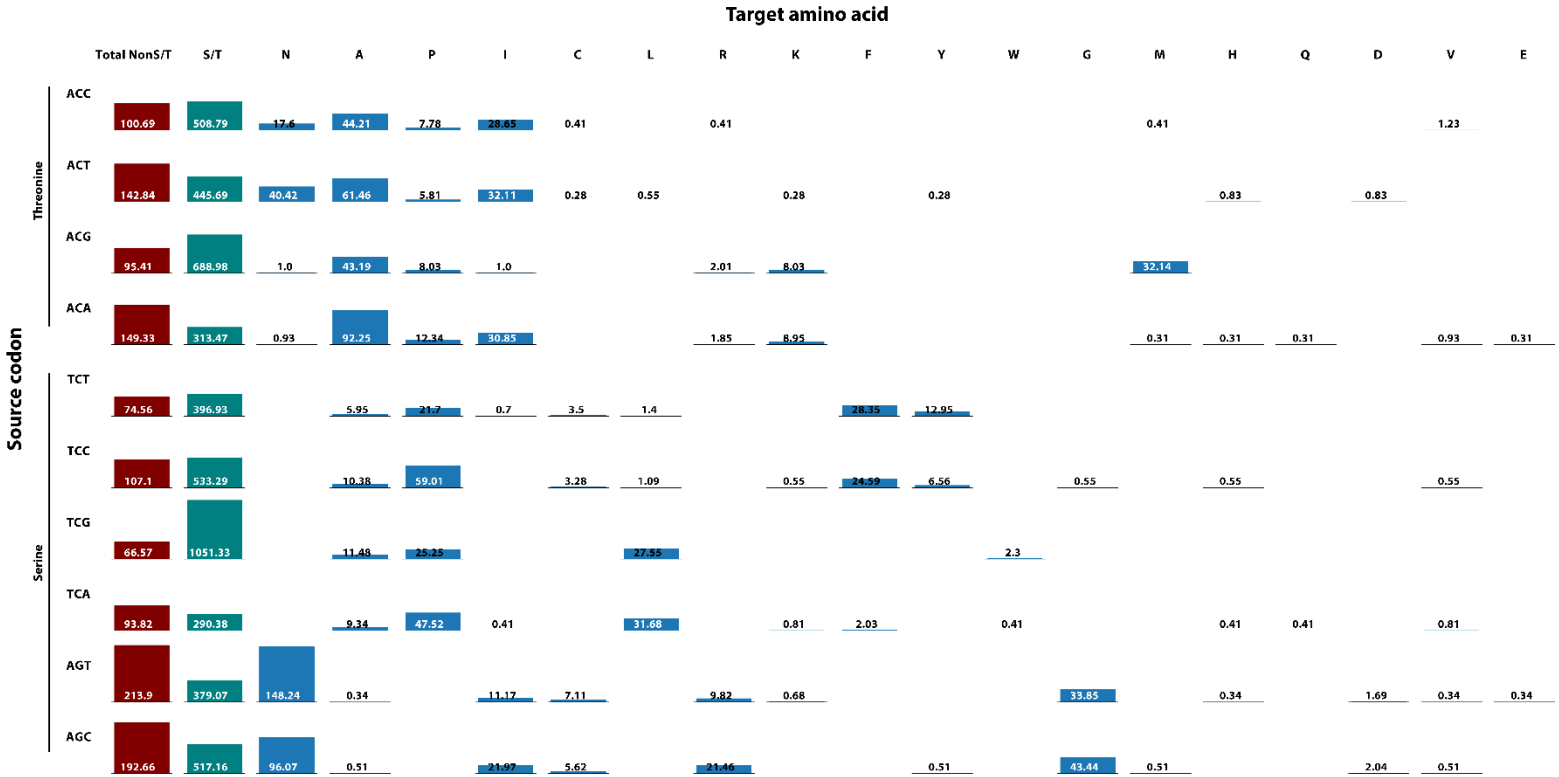


Figure 3.5. Substitutions from serine or threonine codons to the other amino acids. There are large differences in the substitution rates of the different Ser/Thr codons. AGC and AGT are substituted to e.g. N and G at higher rates than e.g. TCT or ACC are substituted to non-Ser/Thr space. The height of the bars of each colour is normalised to the range within that colour. The numbers denote substitutions per 100 000 codons (for each source codon)

The motifs in this section are defined by observed function in studies, and often conform to expected consensus sequences, but sometimes diverge and use non-consensus sites despite having been identified as a functional motif. Phosphorylation sites and SUMOylation sites are defined based on observed modifications and sometimes also have defined consensus recognition sequences. Evidence levels were assigned manually by evaluating the quality and number of studies that have identified the motif and its properties. Mixed indicates conflicting results where the motif status is unsure. Low generally indicates that the motif has been shown not to impact replication of the virus in experimental conditions, medium indicates well studied motifs with a known but minor effect on fitness and replicative success and high denotes essential motifs that have been shown to be important during replication. Conservation levels are manually assigned to reflect how widely conserved the motif is across strains. High denotes conserved in all strains, medium denotes conservation in certain lineages but absent in some lineages. Low denotes generally not conserved and mixed denotes sequence deletions in some strains causing the loss of motifs but general conservation across strains of a particular sequence length.

3.3.2.1. NS1

Influenza gene segment 8 encodes the NS1 protein which most commonly is 230 amino acids in length (Lin *et al.* 2007). NS1 contains a large number of known functional motifs. It does not form part of the new virion or take part in viral assembly, instead its main function is in subverting the host antiviral response (Lin *et al.* 2007). NS1 is localised both to the nucleus and cytoplasm and interacts with several host proteins (Li *et al.* 2015). It is also regulated through host post translational modifications. The known motifs are summarised in Table 3.1. PDZ, SH3 and PKR binding motifs all interact in with different partners involved in the interferon pathway and in PI3K activation (Shin *et al.* 2007b; Thomas *et al.* 2011). These interactions have been shown to be important for viral infectivity and overall fitness. Nuclear and cytoplasmic shuttling is important for NS1 to mediate functions in both compartments, and this shuttling is mediated through conserved NLS and NES motifs (Li *et al.* 2015). These interactions, movements and other functions are regulated through a complex set of modifications, importantly phosphorylation sites and SUMOylation sites. S42, T49, T80 and S195 are the most well conserved phosphorylation sites and the ones that have well established functional impacts on fitness and regulation, importantly through regulating dimerization, dsRNA binding and interactions with RIG-I (Kathum *et al.* 2016; Zheng *et al.* 2017).

3.3.2.2. NS2

NS2 is encoded by the same gene as NS1 in an overlapping reading frame and is one of the smallest influenza A proteins at only 121 residues in length (Paterson & Fodor 2012). NS1 and NS2 share the 18 N-terminal residues which in NS2 are spliced together with the alternative reading frame residues

to form the full protein (Lamb & Lai 1980). Its main function is in export of the vRNP complex to enable viral assembly at the membrane (O'Neill *et al.* 1998). It forms a complex with M1 and vRNP. Highly conserved NES sites have been shown to be essential for this function (Table 3.2). In addition, a cluster of serines have been shown to be phosphorylated with S24 being the primary functional phospho-acceptor and the most conserved residue of the three (Hutchinson *et al.* 2012). It has been suggested that this phosphorylation site can regulate nuclear export as it is adjacent to the two nuclear export sites (Reuther *et al.* 2014). All motifs in NS2 are from part of the gene that overlaps with NS1 in an alternate reading frame.

Table 3.1. NS1 functional motifs. Experimentally determined modification sites and other interaction motifs used by NS1 to subvert the host.

Motif	Position	Function	Conservation	Evidence level	Reference
PDZ	226-230	Prevent cell apoptosis	Mixed	High	(Thomas *et al.* 2011)
SH2	89-92	Mediates PI3K interaction	High	Mixed	(Shin *et al.* 2007b)
SH3	164-167	PI3K activation	High	High	(Shin *et al.* 2007a)
SH3 2	212-215	PI3K activation. Strain specific.	Low	High	(Heikkinen *et al.* 2008; Ylösmäki *et al.* 2015)
NLS1	35-41	Nuclear localisation	High	High	(Greenspan *et al.* 1988)
NLS2	216-221	Nuclear and nucleolar localisation	Low	Low	(Melén *et al.* 2012)
NES	137-146	Nuclear export	High	High	(Han *et al.* 2010; Li *et al.* 1998)
PKR	123-127	Prevent eIF2 activation	High	High	(Min *et al.* 2007)
Sumoylation	70	Regulate interferon response	Low	Medium	(Santos *et al.* 2013)
Sumoylation	219	Regulate interferon response	High	Medium	(Santos *et al.* 2013)
Phospho	42	Regulate dsRNA binding (Motif PKCa)	High	High	(Hsiang *et al.* 2012)
Phospho	48	Unknown, no virus attenuation from motif loss	Low	Low	(Hsiang *et al.* 2012)
Phospho	49	Regulate binding to TRIM25 and dsRNA (Motif PKB)	High	High	(Kathum *et al.* 2016)
Phospho	80	Regulating binding with RIG-1. (Motif GSK3)	Medium	High	(Zheng *et al.* 2015)
Phospho	195	Regulate dimerization (Motif PKA)	High	Medium	(Hutchinson *et al.* 2012)
Phospho	215	Unknown, no attenuation from loss	Low	Low	(Hsiang *et al.* 2012)

3.3.2.3. NP

Nucleoprotein is one of the most highly expressed influenza A proteins along with M1 and the coat proteins (Kummer *et al.* 2014). It is a 498 amino acid protein encoded from gene segment 5 and is functionally important for the stabilisation of viral RNA during replication (Portela & Digard 2002). It forms a complex with the polymerase, the viral RNA and the proteins M1 and NS2 during export to achieve this (Pohl *et al.* 2016). Its functions include multimerisation and viral RNA binding, nuclear import and export both for monomers and for vRNPs during infection, and several ways of regulating these functions through modification sites (Mondal *et al.* 2015; Turrell *et al.* 2015). The key conserved motifs include the localisation motifs, with the N-terminal one having been determined to have the biggest functional impact (Neumann *et al.* 1997); the export sites which are thought to also be important for vRNP export alongside export sites of N2 and M1 (Yu *et al.* 2012); and many phosphorylation sites. The phosphorylation sites near the N-terminus have been shown to regulate localisation through the NLS and the remaining phosphorylation sites are at the multimerisation interface (Mondal *et al.* 2015; Zheng *et al.* 2015). The exact structural and functional details of these sites have not yet been elucidated.

Table 3.2. NS2 functional motifs. NS2 is a short protein and has few functional motifs, NESs are the two most important motifs as they regulate vRNP export and are essential for viral spread.

Motif	Position	Function	Conservation	Evidence level	Reference
NES 1	12-21	vRNP export	High	High	(Iwatsuki-Horimoto *et al.* 2004)
NES 2	31-40	vRNP export	High	High	(Huang *et al.* 2013)
Phospho	23	Low phosphorylation. Regulation of export	Low	Low	(Hutchinson *et al.* 2012; Reuther *et al.* 2014)
Phospho	24	Regulation of export	High	High	(Hutchinson *et al.* 2012; Reuther *et al.* 2014)
Phospho	25	Low phosphorylation. Regulation of export	Low	Low	(Hutchinson *et al.* 2012; Reuther *et al.* 2014)

3.3.2.4. M1

M1 is a 252 amino acid protein encoded by segment 7 (Shtykova *et al.* 2013). It and crucial for viral assembly at the membrane. M1 interacts with vRNP and NS2 as well as the cytoplasmic tails of HA and NA to assemble the viral genome and enable virion budding (Rossman & Lamb 2011). It interacts with the membrane and is thought to interact with the host cytoskeleton to achieve localisation and assembly (Avalos *et al.* 1997; Bedi & Ono 2019). It is the most numerous of the coat proteins, and possibly the most numerous of all the influenza proteins with about 3000 copies per virion (Kummer *et al.* 2014). Despite many known phosphorylation sites, the functions and regulatory aspects of these modi-

fications are poorly understood. The most well characterised motifs to date are the localisation and export sites that are conserved and functionally important for M1 targeting to the nucleus after translation and subsequent vRNP localisation to the cytoplasm (Li *et al.* 2015).

Table 3.3. NP functional motifs. Motifs and phosphorylation sites enable vRNP and NP shuttling and regulation.

Motif	Position	Function	Conservation	Evidence level	Reference
NLS 1	3-13	Unconventional NLS, essential for mRNA synthesis and transport.	High	High	(Li *et al.* 2015; Neumann *et al.* 1997)
NLS 2	198-216	Bipartite. Essential for viral survival,	High	High	(Li *et al.* 2015)
NES 1	24-49	Shuttling	High	Medium	(Yu *et al.* 2012)
NES 2	183-197	Shuttling	High	Medium	(Yu *et al.* 2012)
NES 3	248-274	Shuttling	High	Medium	(Yu *et al.* 2012)
Phospho	3	Nuclear shuttling regulation	High	High	(Arrese & Portela 1996)
Phospho	6	Nuclear shuttling regulation.(Predicted)	High	Low	(Hutchinson *et al.* 2012)n
Phospho	9	Nuclear shuttling regulation	High	High	(Zheng *et al.* 2015)
Phospho	165	Regulates oligomerisation	High	High	(Mondal *et al.* 2017; Turrell *et al.* 2015)
Phospho	188	Regulates nuclear export	High	High	(Li *et al.* 2018b)
Phospho	378	Regulates oligomerisation	High	High	(Mondal *et al.* 2017)
Phospho	402	Regulates oligomerisation	High	High	(Mondal *et al.* 2017)
Phospho	407	Regulates oligomerisation	High	High	(Mondal *et al.* 2015, 2017)
Phospho	413	Regulates oligomerisation	High	High	(Mondal *et al.* 2015, 2017)
Phospho	457	Unknown	High	High	(Hutchinson *et al.* 2012)
Phospho	472	Unknown	High	Low	(Hutchinson *et al.* 2012)
SUMO	4	Trafficking and growth	High	High	(Han *et al.* 2014)
SUMO	7	Trafficking and growth	High	High	(Han *et al.* 2014)

3.3.2.5. M2

M2 is a 97-residue protein and the second protein encoded by segment 7, alongside M1 (Pielak & Chou 2011). It forms a tetrameric proton channel in the viral coat. It exists in low copy numbers and is crucial for viral entry into host cells through the release of the viral RNPs (Cady *et al.* 2009). A well-studied LC3 binding motif has been described and it is thought to regulate host cell autophagy by binding and moving LC3 to the membrane (Beale *et al.* 2014). Several sites have been observed to be phosphorylated through mass spectrometry, but only serine 64 is conserved and consistently phosphorylated (Hutchinson *et al.* 2012). The exact function of this phosphorylation has not yet been established however.

Table 3.4. M1 functional motifs.

Motif	Position	Function	Conservation	Evidence level	Reference
NLS 1	100-106	Classic NLS. Essential for localisation	High	High	(Li *et al.* 2015)
NLS 2	77-79	Basic stretch essential for localisation	High	High	(Li *et al.* 2015)
NES 1	59-68	Essential for function and export	High	High	(Cao *et al.* 2012)
Phospho	5	Unknown. Possible regulation of lipid binding	High	High	(Hutchinson *et al.* 2012)
Phospho	9	Unknown. Possible regulation of lipid binding	High	High	(Hutchinson *et al.* 2012)
Phospho	37	Unknown	High	Medium	(Hutchinson *et al.* 2012)
Phospho	108	Regulate localisation	High	High	(Hutchinson *et al.* 2012)
Phospho	168	Unknown	High	Medium	(Hutchinson *et al.* 2012)
Phospho	169	Unknown	High	Medium	(Hutchinson *et al.* 2012)
Phospho	195	Unknown	High	Medium	(Hutchinson *et al.* 2012)
Phospho	225	Unknown	High	Medium	(Hutchinson *et al.* 2012)
Sumoylation	242	Important for M1 vRNP interaction	High	High	(Wu *et al.* 2011)

Table 3.5. M2 functional motifs.

Motif	Position	Function	Conservation	Evidence level	Reference
Lig_Lir_Gen	88-94	LC3 binding preventing autophagy.	High	High	(Beale *et al.* 2014)
Glycosylation	20	No observed effect	High	High	(Wu *et al.* 2017a)
Phospho	64	Main phosphorylation site. Highly conserved. Function unknown	High	High	(Holsinger *et al.* 1995; Hutchinson *et al.* 2012)
Phospho	65	Redundant phosphosite. Phosphorylated in some strains.	Low	Low	(Hutchinson *et al.* 2012)
Phospho	71	No observed functional importance.	Low	Low	(Hutchinson *et al.* 2012; Reuther *et al.* 2014)
Phospho	82	No observed functional importance.	Low	Low	(Holsinger *et al.* 1995; Thomas *et al.* 1998)
Phospho	93	No observed functional importance.	Low	Low	(Holsinger *et al.* 1995; Thomas *et al.* 1998)

Table 3.6. H1 glycosylation sites. Positions refer to the mature H1 protein after cleavage of the N-terminal signal sequence. Positions in parenthesis correspond to the full-length sequence

Motif	Position	Function	Conservation	Evidence level	Reference
Glycosylation	10(27)	Stalk glycosylation site	High	High	(Cruz *et al.* 2018)
Glycosylation	11(28)	Stalk glycosylation site	High	High	(Cruz *et al.* 2018)
Glycosylation	23(40)	Stalk glycosylation site	High	High	(Cruz *et al.* 2018)
Glycosylation	54(71)	Head glycosylation site	Medium	High	(Cruz *et al.* 2018)
Glycosylation	87(104)	Head glycosylation site	High	High	(Cruz *et al.* 2018)
Glycosylation	125(142)	Head glycosylation site	Medium	High	(Cruz *et al.* 2018)
Glycosylation	160(177)	Head glycosylation site	Medium	High	(Cruz *et al.* 2018)
Glycosylation	287(304)	Head glycosylation site	High	High	(Cruz *et al.* 2018)
Glycosylation	481(498)	Stalk glycosylation site	High	High	(Cruz *et al.* 2018)

Table 3.7. H3 glycosylation sites

Motif	Position	Function	Conservation	Evidence level	Reference
Glycosylation	22	Stalk glycosylation site	High	High	(Alymova *et al.* 2016)
Glycosylation	38	Stalk glycosylation site	High	High	(Alymova *et al.* 2016)
Glycosylation	45	Head glycosylation site	Low	High	(Alymova *et al.* 2016)
Glycosylation	63	Head glycosylation site	High	High	(Altman *et al.* 2019)
Glycosylation	81	Head glycosylation site	Low	High	(Altman *et al.* 2019)
Glycosylation	90	Head glycosylation site	Low	Low	(Altman *et al.* 2019)
Glycosylation	122	Head glycosylation site	High	High	(Altman *et al.* 2019)
Glycosylation	126	Head glycosylation site	High	High	(Altman *et al.* 2019)
Glycosylation	133	Head glycosylation site	Medium	High	(Altman *et al.* 2019)
Glycosylation	144	Head glycosylation site	Medium	High	(Altman *et al.* 2019)
Glycosylation	158	Head glycosylation site	Low	Low	(Altman *et al.* 2019)
Glycosylation	165	Head glycosylation site	High	High	(Altman *et al.* 2019)
Glycosylation	246	Head glycosylation site	High	High	(Altman *et al.* 2019)
Glycosylation	285	Stalk glycosylation site	High	High	(Altman *et al.* 2019)

3.3.2.6. HA (strains H1 and H3)

HA is the most abundant of the coat proteins. It has a stalk domain that anchor within the viral membrane and a globular head domain that faces the external environment (Stegmann 2000). Antibodies frequently target antigens on the globular head domain. HA is also the protein responsible for binding to host cells through recognition of specific glycans (Stegmann 2000). Its glycosylation pattern modulates both binding of host cells and recognition by the host immune system.

Table 3.8. N2 Glycosylation sites. Numbering is based on the full length protein.

Motif	Position	Function	Conservation	Evidence level	Reference
Glycosylation	61	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	70	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	86	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	93	N/A	Low	Medium	(York *et al.* 2019)
Glycosylation	146	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	200	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	234	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	245	N/A	Low	Medium	(York *et al.* 2019)
Glycosylation	329	N/A	Medium	Medium	(York *et al.* 2019)
Glycosylation	367	N/A	Low	Medium	(York *et al.* 2019)
Glycosylation	402	N/A	High	Medium	(York *et al.* 2019)

Table 3.9. N1 glycosylation sites. Numbering is based on full length protein.

Motif	Position	Function	Conservation	Evidence level	Reference
Glycosylation	44	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	58	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	68	N/A	Medium	Medium	(York *et al.* 2019)
Glycosylation	88	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	146	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	235	N/A	High	Medium	(York *et al.* 2019)
Glycosylation	365	N/A	Medium	Medium	(York *et al.* 2019)
Glycosylation	434	N/A	Medium	Medium	(York *et al.* 2019)
Glycosylation	454	N/A	Medium	Medium	(York *et al.* 2019)

3.3.2.7. NA (strains N1 and N2)

NA is the second most abundant coat protein and it is essential for influenza A cell entry. After HA binding to sialic acid on host cell surfaces, NA cleaves the sialic acid allowing the virus to enter the cell (McAuley *et al.* 2019). NA is glycosylated mainly for structural and functional reasons, and is not known to have variable glycosylation to the extent of HA (Kim *et al.* 2018).

3.3.3. Motif loss frequency in virus evolution is shaped by codon choice

Through the simulations in chapter 2, I found that different codon choices within a motif consensus sequence drastically alters the evolutionary dynamics. This impacted the frequency at which motifs were lost in the simulations. In section 3.3.1 of this chapter I also found that the substitution rate at the codon level, in the observed sequence evolution of influenza, corroborates the observations from the

simulations for motif codon positions. Both of these observations indicate that the evolution of real functional motifs in influenza would be similarly affected by the expected loss probability of the nucleotide sequence used to encode the motif.

To what extent these dynamics occur in motifs over the course of real influenza strain evolution, and how much drift and selection skew the observed motif evolution dynamics have yet to be explored. Therefore, to investigate to what extent codon choice and motif sequence space impacts the real evolutionary dynamics of known motifs and putative motifs in the real sequences I analysed how frequently mutations within these motifs and putative motifs cause them to be lost. The expectation would be that under neutrality, the probability of loss and loss frequency are linearly correlated as a consequence of the codon usage and robustness of a motif (Figure 3.6). However, when selection acts on a motif essential for the function of the virus, it would tend to be more often preserved in strains and sequences than expected given the probability calculation. In contrast, if a motif carries a negative fitness impact, the expectation would be that more strains lose the motif than predicted.

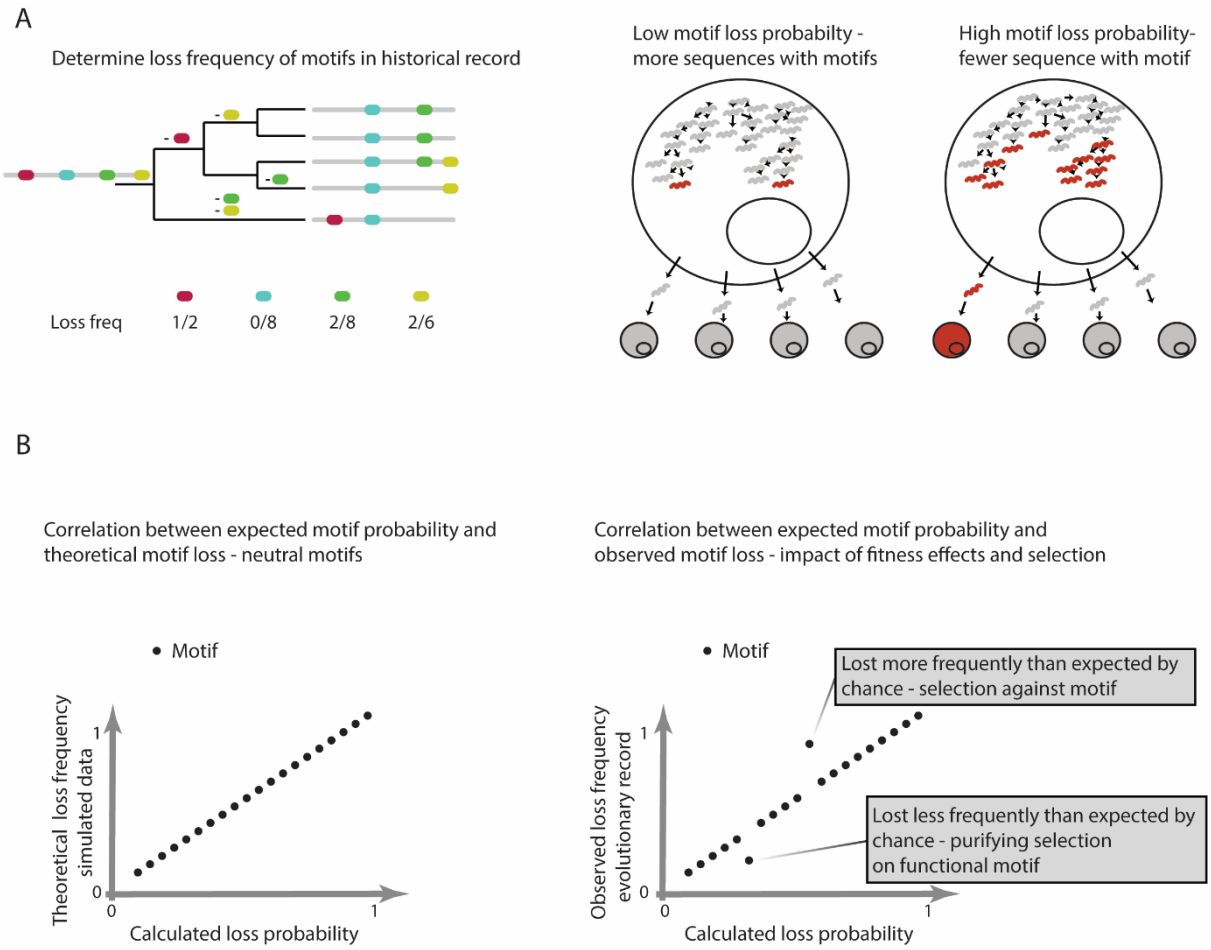


Figure 3.6. The expected outcome of mutation probability on motif loss frequency over sequence evolution. A) Loss frequency is determined through the number of motif loss events compared to the number of possible motif losses given the strain and nodes with a motif, i.e. the red motif is lost in the first branch point in the lineage, meaning no subsequent loss event can happen. Motif probability is expected to correlate with loss frequency due to the population outcome of sequences with motifs. B) Comparison between the expected correlation between probability and loss frequency in neutral evolution or when selection acts on functional motifs.

To investigate this, I used a dataset of known functional motifs from influenza protein NS1 (summarised in section 3.3.2). I also included a set of putative motifs in NS1 that have consensus sequence matches in at least 20% of the strains in the dataset. These putative motifs have no experimental evidence backing them up, but are simply sequences matching the consensus motif that occur in accessible parts of the viral proteins under conditions that imply they could have a functional role (these motifs are outlined in section 3.5.2.1). I would expect that some of them are likely to be functional unknown motifs and some are spurious sequence matches with no motif function. This should provide a dataset showing various levels of selection and random mutation accumulation. The dataset of known functional motifs is likely to be under strong selection and proteins without these sequences are unlikely to be as fit as viruses with the motif. The balance between selection and mutation rate in determining the ultimate evolutionary outcomes can thus be assessed to some degree.

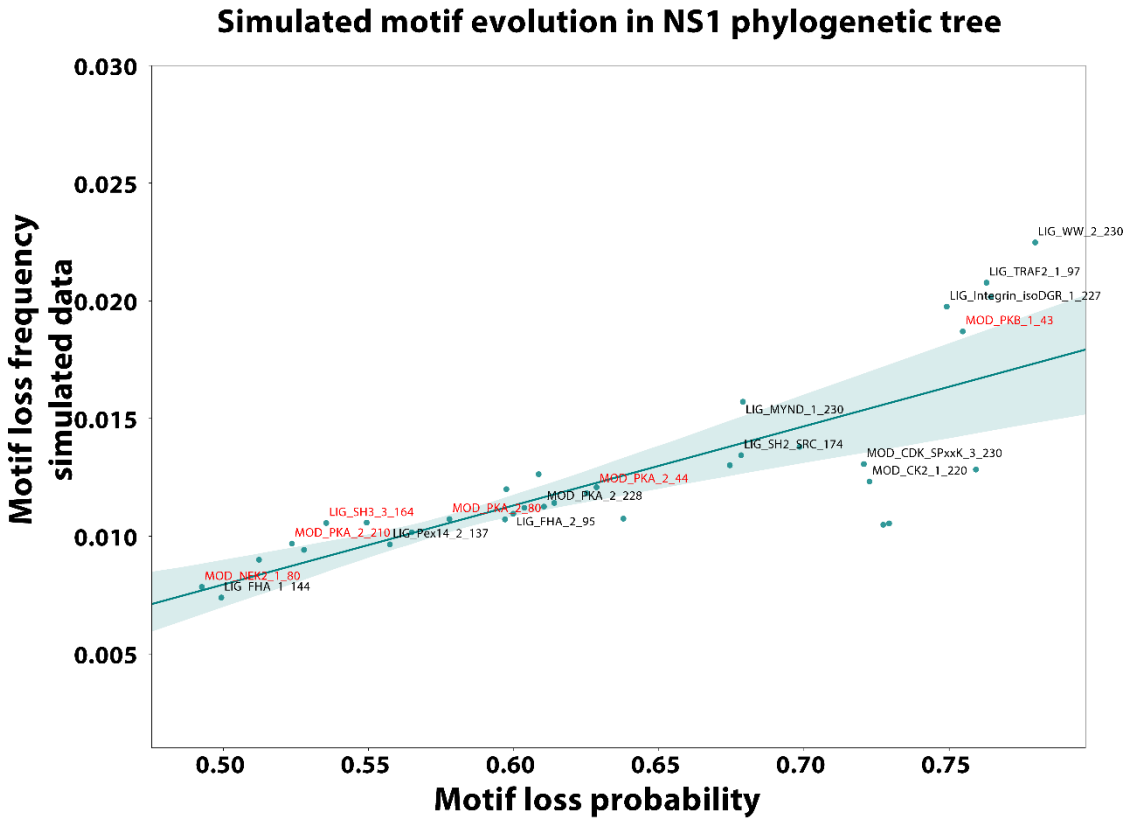


Figure 3.7. Simulated motif evolution in NS1. Predicted loss probability (x-axis) was compared with simulated motif loss using pyvolve (y-axis). The expected loss probability very closely correlates with observed loss frequency in this dataset suggesting under neutrality, the factors that inform loss probability, namely codon space and mutation rate, can closely predict evolutionary outcomes. Some motifs of interest are labelled, in red are experimentally verified motifs (for details see methods section 3.5.2.1).The blue line is best fit through least squares and the shaded area indicates the 95% confidence interval. R^2=0.63, p=2*10^{-8}.

I firstly wanted to get a baseline for the expected loss frequency if no selection was acting on the sequences. Using the set of functional and putative motifs from NS1 I first defined the most used se-

quence for each motif in the viral record. Using these sequences as the starting sequence I then simulated evolution in pyvolve using the same method as for the simulations in chapter 2 (see section 2.4.1.2) (Spielman & Wilke 2015). The tree had the same topology as the determined NS1 tree used in previous simulations, consisting of 8663 strains, and I used the substitution rates estimated from influenza sequence history through the observed substitutions. The loss probability was then calculated for each ancestral strain by the method described in 2.3.3. I determined the loss frequency of each motif by the same method as in chapter 2 (see methods section 2.7.5). This involved determining when an ancestral strain had a motif that matched the consensus sequence which was subsequently not present in the descendant strain due to mutations causing an amino acid substitution. This resulted in a very clear linear correlation between the predicted probability and simulated motif loss as expected under neutrality (Figure 3.7).

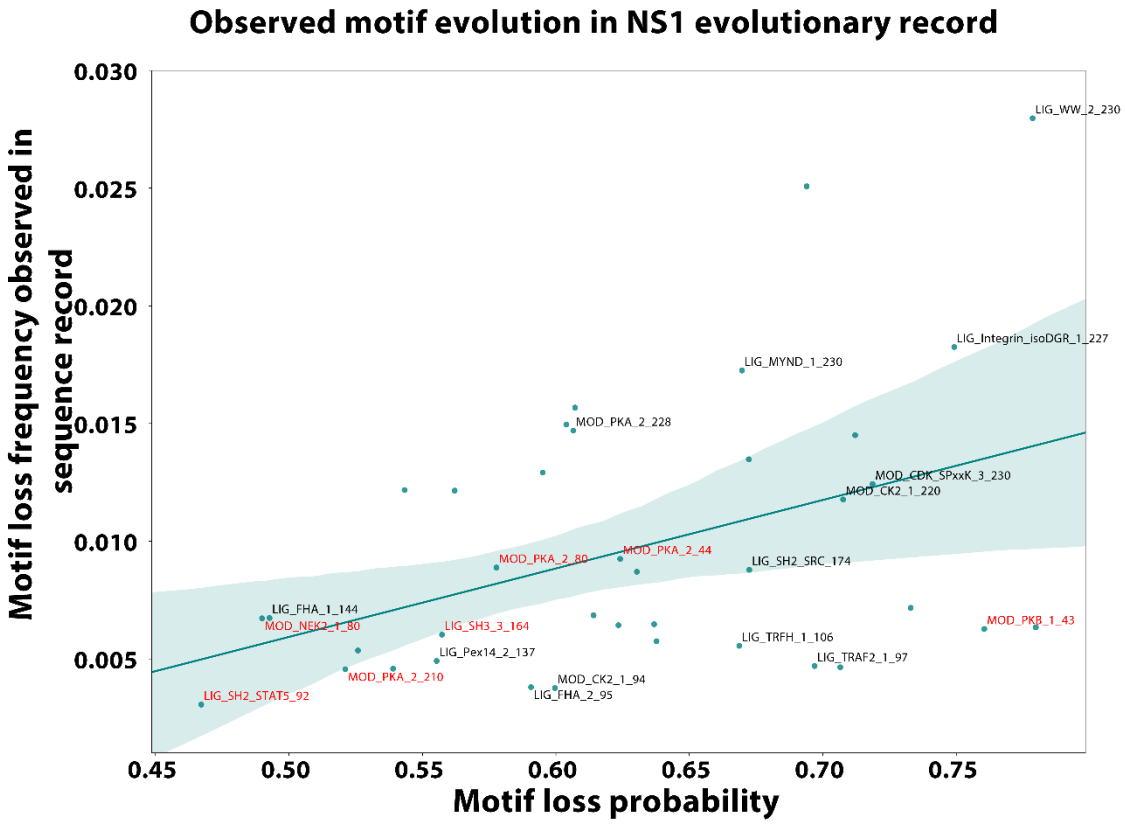


Figure 3.8. Observed motif evolution in NS1. Calculated loss probability based on codon usage and motif sequence space compared with instances of mutations that cause the loss of the motif. There is a clear overall trend between different motif types showing that different motif types have different predicted loss probability due to codon choice and motif consensus which correlates significantly with how often critical loss mutations occur in these motifs. In addition, within several motifs of similar types their loss probability also correlates with loss frequency suggesting that loss probability (inverse of robustness) describes the properties that influence motif evolutionary dynamics accurately. Some motifs of interest are labelled, in red are experimentally verified motifs (for details see methods section 3.5.2.1).The blue line is best fit through least squares and the shaded area indicates the 95% confidence interval. R^2=0.17, p=0.009.

I then went on to do the same analysis for the motif sequences in the viral sequence record, determining frequency of loss over the tree using the same approach as for the simulated data (Methods, section

3.5.2.3). There is a significant correlation between the predicted motif probabilities and their loss frequency in the sequence record (Figure 3.8). Since the predicted probability is based on the codon space known for the motif consensus, and the relative mutation rate these results suggest that motif evolutionary dynamics are predictable to some extent, given knowledge of these factors. Interestingly the line of best fit is almost identical to the predicted linear correlation for the simulations under neutrality, suggesting the dataset is reasonably distributed between motifs with different selection pressure and fitness.

Another interesting observation is that the known functional motifs that carry significant fitness loss and attenuation if lost, crucially SH2_STAT5_92, PKA_2_210, SH3_3_164 and PKB_1_43 (as established by experimental observations, discussed in 0), all fall below the rate expected at neutrality, indicating these regions see fewer mutations than expected. A possible explanation for this observation is that there is selection for the motif pattern there, as it is a known functional motif. It is worth noting that other selection pressured for unrelated reasons could also contribute to the lack of mutations tolerated in this region.

To more accurately compare the observed mutational dynamics with the dynamics expected under neutral evolution I plotted the loss frequency from the simulated data against the observed loss frequency from the historical record (Figure 3.9). As observed in Figure 3.8, motifs with known function in the virus are biased compared to the expected such that they are lost less, which is likely to in part be due to selection specifically for the motif sequence pattern. Generally, if there was no selection the correlation would be expected to fall on the line x=y. Many of the putative motifs included here appear to be very close to this neutral motif loss rate. Overall, more motifs appear to be lost less frequently than expected, which could indicate selection for motif function in some instances. However, it is clear that within viruses, many factors contribute to selection pressures on many levels. In addition to functional motifs, viral sequences are limited on both the RNA level and protein level due to interaction and structural constraints. This is likely to be a large part of any selection observed on residues. High conservation of residues that happen to conform to a motif consensus would thus result in false positives when comparing to expected loss under neutrality. The overall correlation does suggest that the probability of loss at a specific motif site plays a significant part in the evolutionary outcome, however it will be important to further disentangle the contribution of all the different aspects of viral sequences.

I have also included some motifs that would be expected to be in the wrong functional context in NS1, either structurally or within the cell environment, to compare the outcome (e.g.integrin_isoDGR). Most of these motifs are lost the same as or more frequently than expected neutrally. The hypothesis for how motif sequences with low fitness emerge in the first place could be one of two reasonable scenarios: the first is that they emerge through drift, despite having lower fitness, and subsequent strains

lose the motif with high frequency; the second is that changes to conditions or context could make neutral motif patterns suddenly detrimental, causing their loss from many strains.

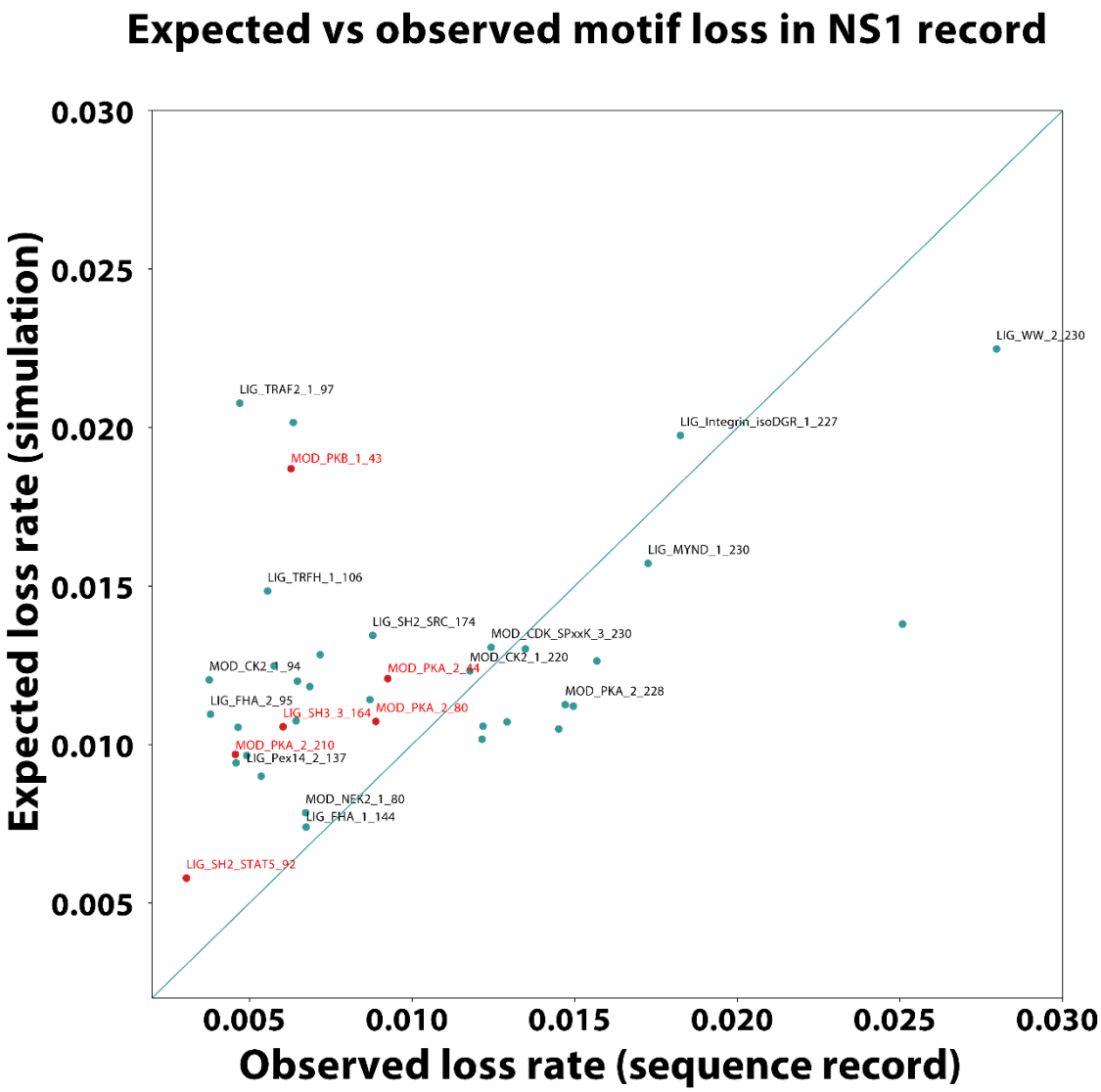


Figure 3.9. Observed motif loss rate compared to expected neutral loss rate in motifs in NS1. A shift left of the x=y line (blue) suggests less loss than expected and a shift right suggests higher loss than expected. In red are some known functional motifs that all appear shifted left to different degrees. There is an enrichment in motifs lost less than expected. This may be due to motif sampling in this dataset or a consequence of viral sequence under selection.

Several putative motifs that are not known to be functional, appear to be lost much more rarely than would be expected when comparing to the simulated neutral sequence. These would be potential targets for future motif investigations to determine if they are under purifying selection due to being functional motifs, or if their low loss frequency is a consequence of other selection pressures that happen to coincide with the motif sequence pattern. Overall, these findings make a strong case for the im-

portance of a mutation and codon based probability metric in the understanding of evolution of specific protein features such as motifs, and it will be worth exploring how well this kind of analysis can be used to predict new important motif targets for future characterisation and study.

3.3.4. Codon bias in phosphorylated serines and threonines

The observations thus far show that firstly, different codons have different substitution spaces resulting in differences in probabilities of amino acid substitutions during evolution; and secondly, that motifs using different codons (with different probabilities) experience different levels of motif loss. This implies that motif positions using the most optimal codons for the relevant motif consensus will on average be more robust to mutations. This would tend to lead to preservation of motifs and could result in higher viral fitness.

To explore if there is such a codon bias for optimising motif robustness I wanted to investigate codon usage at known phosphorylation sites in influenza. Phosphorylation is a key PTM in the majority of influenza proteins and many known phosphorylation sites are present in the influenza proteome (Hutchinson *et al.* 2012). In addition, phosphorylation sites can be assigned in absence of exact motif identity through the experimental identification of phosphorylated residues. Thereby a larger dataset could be used for these residues than by comparing just a single kinase motif, adding more power to the analysis. For that reason, I focused specifically on the codon usage at the key serine/threonine position for all known phosphorylation sites. The simulations and predictive model calculations suggest a hierarchical list of serine and threonine codons, with some being more robust to mutations in a motif codon space than others. These observations are also reflected in the substitution rates and targets for the different codons that were found in section 3.3.1 where e.g. AGC and AGT codon had much higher substitution rates in the record, and most frequently substituted into asparagine. One would expect to see the more robust codons being favoured at functionally important phosphorylation sites to reduce motif loss and thereby maximise fitness.

I first determined the codons used at all serine and threonine sites across the influenza proteins, excluding the known phosphorylated residue positions. This provided a baseline for the codon usage at Ser/Thr sites across influenza proteins, as sequences and organisms experience codon bias for a number of known reasons (Belalov & Lukashev 2013; Komar 2016). I then compiled the codons used at all the known phosphorylated serines and threonines from all the influenza proteins with the assumption that both serines or threonines can be phosphorylated in these contexts (Rust & Thompson 2011). The dataset included 31 phosphorylated residues across NS1, NS2, M1, M2, NP and PA with between 9000-11000 strains for each protein sequence (see methods for details). I determined the codons used by each strain in the phylogeny.

There is a clear difference in codon usage between the control set and the phosphorylated set of serine and threonine residues (Figure 3.10). The codons used at conserved phosphorylated residues have a

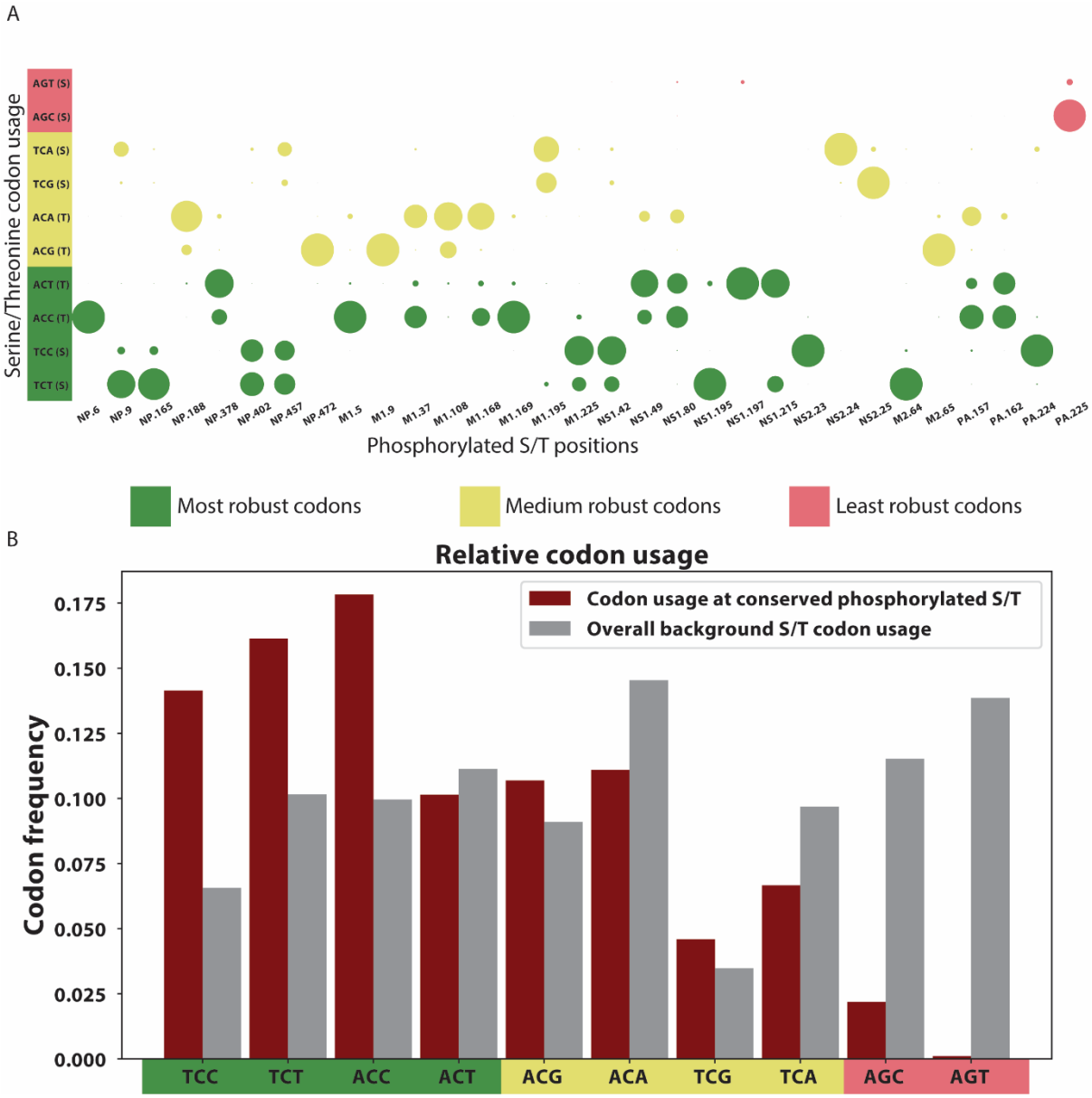


Figure 3.10. Codon usage at conserved and functional phosphorylated Ser/Thr residues. A) Data from all the known phosphorylation sites in all influenza A proteins. Each circle shows the proportion of codons at the phosphorylated position in all the strains in the dataset for each protein. There is a clear overrepresentation in the green "high robustness" area for codons at phosphorylation sites. The red area shows a marked depletion of codon usage in phosphorylation sites compared to other Ser/Thr sites. B) Summary data for each Ser/Thr codon showing clear enrichment for robust codons and depletion of non-robust codons. χ^2 (9, N = 1134) = 151.7, p << .0000001.

significantly different distribution than the codons used at other sites under both more and less conservative assumptions of independence between viral strains. Under the assumption that codon state is independent between strains, all strains are counted as independent entries: $\chi 2$ (9, N = 1330803) = 152717, p < .0000001. Under a more realistic assumption of strain relatedness where codon state depends on strain ancestry partially, and where higher weight is given to related strains that predominantly favour specific codons according to the level of conservation of the codons, $\chi 2$ (9, N = 1134) =

151.7, $p < .0000001$. Finally, under the most conservative assumption where individual strain codon choices are ignored (i.e. ignoring the level of conservation of codon choices across many strains) and the majority codon choices at phosphorylated sites are used to define the codon state (effectively asking out of the 31 phospho sites, how many times would we expect TCC, ACC, AGC etc. to be in the majority), $\chi 2$ (9, N = 313) = 20.5, $p < .02$. The functional phosphorylation sites significantly avoid using the less robust codons, most notably AGT and AGC from serine, which are predicted to be by far the least robust. Instead, they favour the codons that tend to be the most robust, predominantly serine codons TCC and TCT and threonine codons ACT and ACC, thereby maximising robustness and by extension potentially maintaining higher viral fitness. The combination of nucleotide mutation rates and available functional codon space make these codons less likely to be lost through random mutations. The background codon usage across serine and threonines show no bias for codons robust within the Ser/Thr space. In fact, codons AGC and AGT are the two most frequently used serine codons generally across the influenza proteins (Figure 3.10, grey bars). I hypothesise that the functional overlap in many instances between Ser/Thr and other polar residues such as asparagine, cysteine and sometimes alanine make up a functional set in non-phosphorylation contexts that make these residues more robust in structural residue contexts. In other words, because of the similarity in biophysical characteristics between e.g. asparagine and serine (size and polarity) they are often interchangeable in residues that are not phosphorylated. This hypothesis is supported by the data from section 3.3.1 that shows AGT and AGC being the most frequent codons in Ser/Thr to get substituted to asparagine. This finding suggests that the functional set of codons that make up linear motifs is an important consideration when trying to understand the evolutionary pressures and potential trajectories of viral sequences. This also indicates that codon choice in motif contexts is a sequence based indicator for motif functionality and selection. Having this information alongside detailed mutation rate data for viruses will be essential to better understand and importantly predict outbreaks and evolutionary innovations and pressures in influenza.

3.3.5. Codon bias in HA and NA glycosylation sites

N-linked glycosylation sites are highly important post-translational modification sites in the coat proteins haemagglutinin and neuraminidase of influenza a strains (Tate *et al.* 2014; Vigerust & Shepherd 2007). The evolution and functional characteristics of these sites have been studied in a range of influenza strains. Changes to the glycosylation state is an important mechanism through which the virus can avoid recognition by masking antigenic regions of the coat. Changes to glycosylation sites have been linked to more severe symptoms as well as increased mortality during infection (Peng *et al.* 2019; Wu *et al.* 2017a). In addition, changes in glycosylation can reduce the effect from vaccination attempts (Allen & Ross 2018). Importantly, the cell systems used for studying virus function and in developing vaccines are known to induce selective pressure and result in sequence changes that alters the virus function and ultimately the efficacy of vaccines (Chen *et al.* 2018a; Wu *et al.* 2017b; Zost *et*

al. 2017). There is a lot to be gained from understanding the properties that contribute to the evolutionary changes at these sites.

The HA protein consists of two structural regions, the stalk and the head (Gamblin & Skehel 2010). Both of these are glycosylated, but to different effect. The stalk is glycosylated primarily for structural and functional reasons within the coat, and rarely interacts with antibodies (Daniels *et al.* 2003). As a consequence, it has more highly conserved glycosylation sites and less evolutionary variation (Liu *et al.* 2016). The head domain, on the other hand, is the main site for antibody binding, and glycosylation sites here are used by the virions to prevent antigen recognition (see Figure 4.4 (p. 126) for an example structure of glycosylated H3). Glycosylation sites in the head region tend to be less conserved than those in the stalk region, and new glycosylation sites tend to emerge more frequently in different strains (Altman *et al.* 2019). The NA protein, on the other hand, is much less prevalent on the surface of the virus particle and as such is not as common a target for antibodies during infection (Chen *et al.* 2018b). Overall glycosylation sites in NA are more stable and conserved over influenza history.

In this section I have investigated the codon usage in N-linked glycosylation sites within the stalk and head domains of HA as well as the NA proteins for a range of different strains with different glycosylation sites. Glycans are added to influenza coat proteins in the ER by the protein oligosaccharyltransferase, which recognises the consensus sequence N[^P][ST] during protein translation and transfers the glycan on to the asparagine (Schwarz & Aebi 2011).

HA and NA proteins have different levels of glycosylation site conservation and they experience different selection pressures through different levels of antibody targeting. This is likely to result in different fitness outcomes due to glycosylation loss. In contrast to the generally negative fitness outcome resulting from phosphorylation site loss, glycosylation site loss and turnover is therefore likely to be more beneficial in some instances. Therefore, this could lead to different codon usage patterns for the Ser/Thr position in glycosylation sites compared to phosphorylation sites.

To investigate the codon usage, I analysed the strains with known glycosylation sites and looked specifically at the Ser/Thr position of the motif as that is likely to be have the biggest impact on robustness. The asparagine has the choice of only two codons and the robustness difference between the two is very small (but could still have an impact).

In HA from H1 and H3 there is a distinct bias in the codon usage profile at Ser/Thr sites in glycosylation motifs compared with the overall codon usage (Figure 3.11). Again, comparing statistics for the different assumptions as for the phosphorylated codon choices (p. 91), assuming full independence between strains resulted in a highly significant difference, χ^2 (9, N = 1152392) = 70769.3, p << .00000000001. Under the partially independent (most realistic) assumption the difference was also

highly significant with χ^2 (9, N = 972) = 71.7, p < .00000001, however the most conservative assumption is more ambiguous with, χ^2 (9, N = 450) = 16.5, p = .05 likely due to the smaller dataset available for unique glycosylation sites.

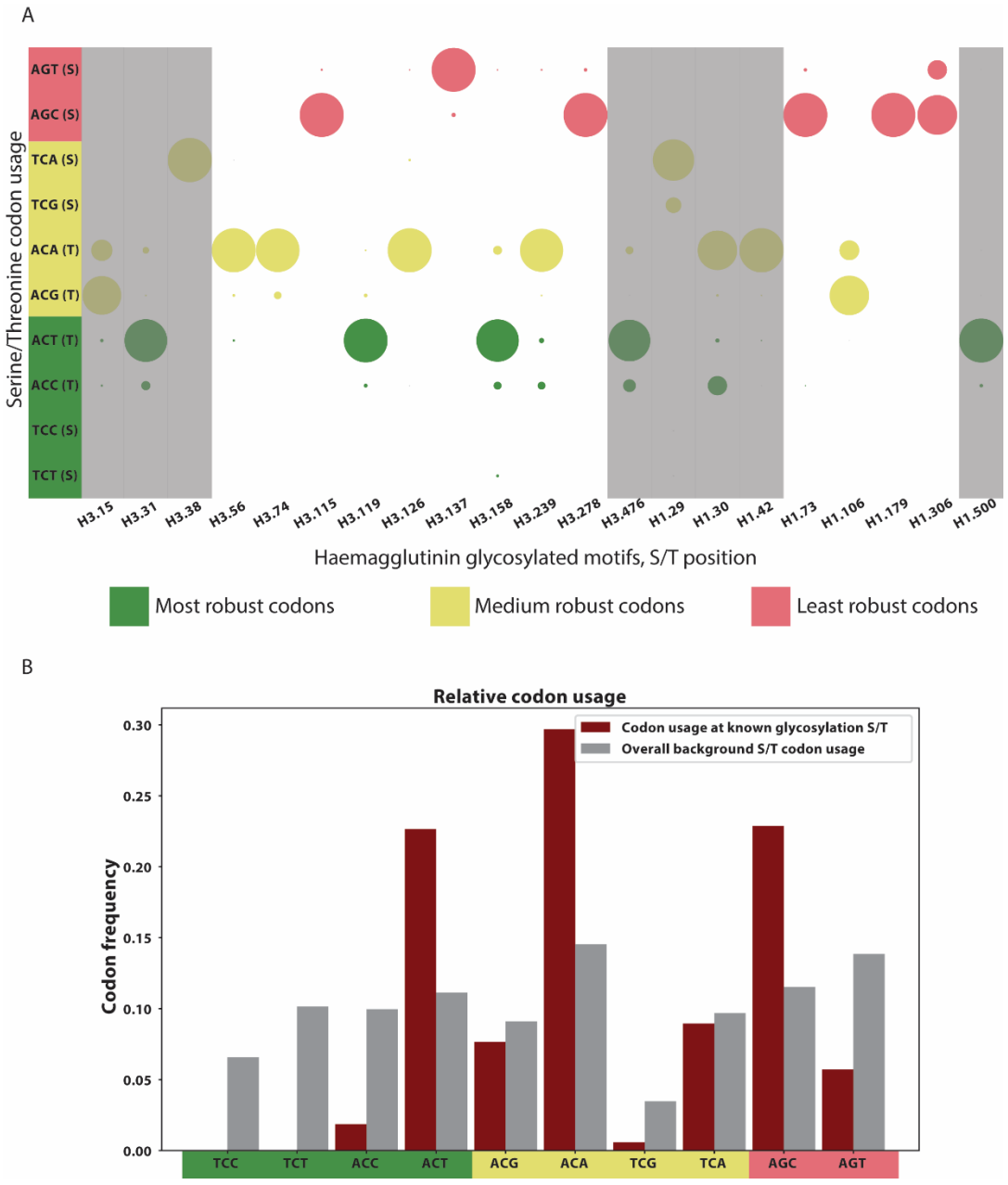


Figure 3.11. Codon usage at Ser/Thr positions in HA glycosylation sites. A) Codon usage at Ser/Thr positions in N-linked glycosylation motifs in H1 and H3 with colours highlighting robustness. Highlighted in dark are the glycosylation sites in the stalk which on average are more conserved. These are not seen in this dataset to use the least robust AGC and AGT. Overall at globular head sites there is an enrichment of low robustness codons. B) Summary data comparing codon usage against the expected usage in influenza. χ^2 (9, N = 972) = 71.7, p < .00000001.

In all, this suggests codon usage is significantly different at Ser/Thr at glycosylation sites. Comparing directly to codon usage at phosphorylation sites, here there is a depletion of highly robust codons and instead an enrichment of codons with medium or low robustness χ^2 (9, N = 249) = 81.49, p <

.00000001. Interestingly, the more conserved glycosylation sites in the stalk domain do not use AGC or AGT at all and instead favour more robust codons in this dataset. In comparison, AGC and AGT use is enriched in glycosylation sites in the globular head domain where conservation is low, and there is overall lower use of the more robust codons. Whether this is a consequence of selection for higher mutability or a result of lower conservation and frequent gain-loss cycles is difficult to say.

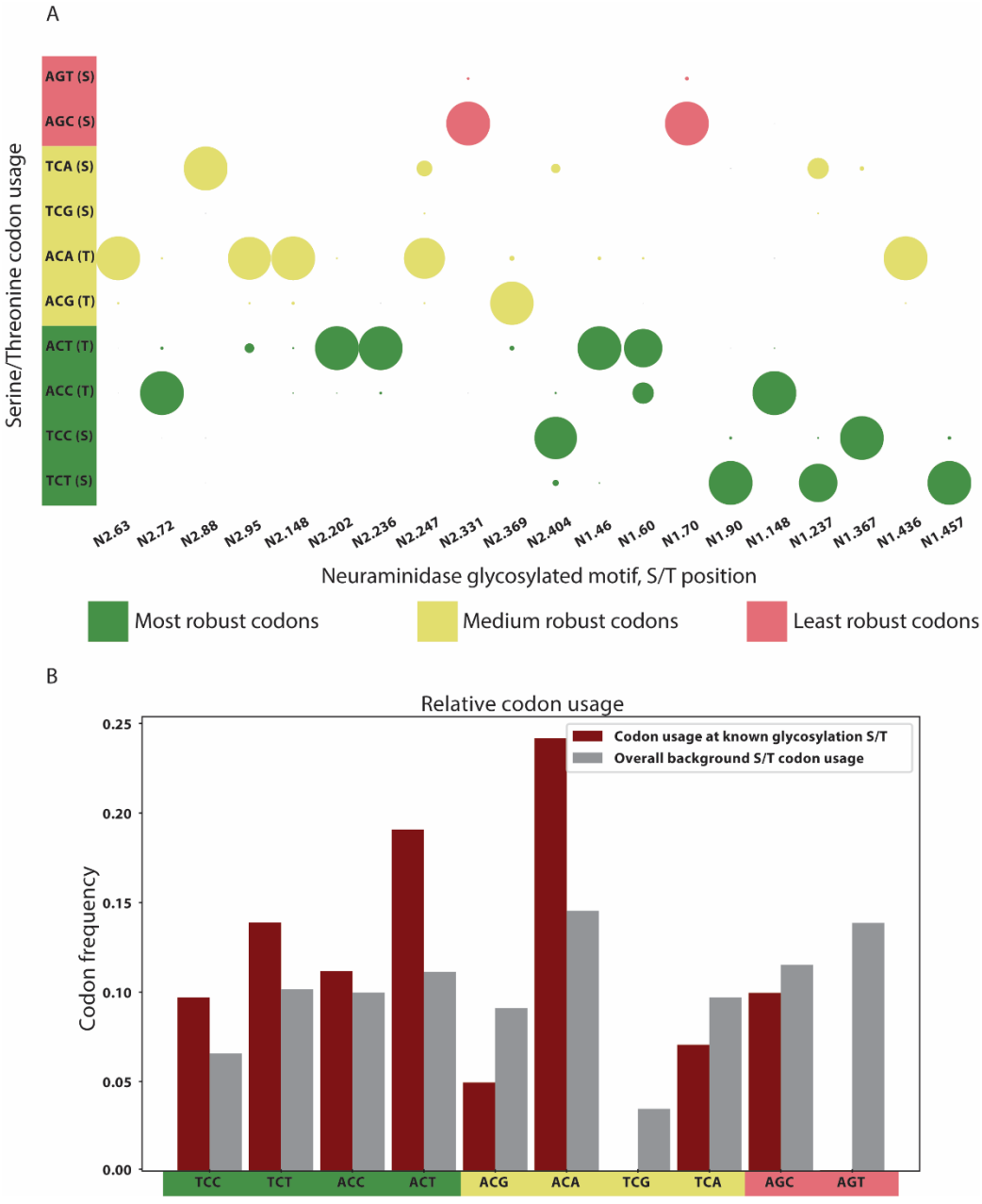


Figure 3.12. Codon usage at Ser/Thr positions in NA glycosylation sites. A) Codon usage at Ser/Thr positions in N-linked glycosylation motifs in N1 and N2 with colours highlighting robustness. These glycosylation sites favour robust Ser/Thr codons over less robust. B) Summary data comparing codon usage against the expected usage in influenza, $\chi 2$ (9, N = 995) = 33.14, p < .0001.

Nonetheless, a consequence of this codon usage is likely to be higher variation and mutability, which is consistent with the changes induced when adapting to differential immune responses.

By contrast, in neuraminidase N1 and N2 the codon usage profile is depleted towards the low robustness codons and highly enriched for robust and medium robust codons in a pattern resembling that of phosphorylation sites (Figure 3.12). Under the assumption of strain independence this difference is significant compared to the expected background, χ^2 (9, N = 1171117) = 32936.59, p < .00000001 , and for the more realistic, partially dependent assumption it is also significantly different, χ^2 (9, N = 995) = 33.14, p < .0001. However, under the most conservative assumption of complete strain dependence on ancestral state, there is not a significant difference, χ^2 (9, N = 399) = 8.9, p > .05. Although this assumption is biologically unrealistic, it still suggests there is somewhat more similarity in codons used in NA with the background usage, compared to HA. Comparing HA and NA codon usage directly, there is again a significant difference in codon usage distribution, χ^2 (9, N = 183) = 36.6, p < .00001. Overall, this would suggest there is more selection towards preserving the glycosylation sites which would result in overall lower variation and loss through mutations. This is consistent with the lower pressure from the immune system and the lower variation over time for glycosylation sites in NA. The codon usage profile for NA is more similar to the one for phosphorylation Ser/Thr sites but still very different from the codon usage at general Ser/Thr sites as well as Ser/Thr site in HA. This suggests Ser/Thr sites in NA avoid the less robust codons, leading to higher robustness to mutations, and could indicate that specific codons are indeed under purifying selection.

The consequences of glycosylation variation in especially the head domain of HA has been studied in many different strains and it has been established that this variation is important in viral fitness and evolution (Altman *et al.* 2019). However, given the less robust codon usage at those sites I would also expect that HA heterogeneity within a single cell and host would be high for the head domain. It is possible that having a heterogeneous population of HA with different glycosylation patterns in the coats of individual viral particles could have fitness implications for the infection as a whole. This is something that has not been addressed before. There is increasing evidence for the extreme heterogeneity of influenza in the cell population of single host, and it is likely that changing mutational properties through codon choice ultimately impacts this heterogeneity (Russell *et al.* 2018). It will be of great interest to take these results further by investigating the protein coat heterogeneity in single infections and determine the fitness properties of changes in this aspect.

3.3.6. Codon usage in functional and conserved influenza motifs

The previous two section on Ser/Thr sites in phosphorylation and glycosylation motifs clearly indicate that there is a preferred codon usage bias to preserve important functional residues in motifs. It would stand to reason that other essential motifs similarly maximise their mutational robustness to improve viral fitness. Since there are very few known motifs of each motif type in influenza the analysis I did to assess codon usage in Ser/Thr will not be applicable, since it requires several independent motif sites of similar type with the same amino acid limitations. For most other motifs in influenza only one

or two instances are known which includes PDZ, SH2 and the instances of NLSs and NESs (See section 3.3.2). To infer whether these conserved motifs predominantly favour certain codons I instead analysed the codon usage within the same motif over all the strains that have the motif. If there is no bias towards either codon I would expect silent mutations between allowed codons to occur throughout the evolutionary tree and between all the different strains. However, if there is a bias towards more robust codons I would expect to see a signal for the strains to use a smaller range of the more robust codons.

For this analysis I selected conserved and well characterised motifs from the literature (see section 3.3.2). The motifs used were nuclear localisation sequences 1 and 2 from NP; the nuclear localisation and nuclear export sequences from NS1, alongside the C-terminal PDZ motif and two phosphorylation sites (S42 - PKCα and S195 – PKA); the M1 nuclear localisation and nuclear export sequences; and lastly a glycosylation site and an LC3 interacting binding motif from M2.

To determine if these motifs show any motif-centric codon bias across the key sites I first determined the codon usage at the key motif sites across all the strains in the dataset. For the codon usage distribution at each position in a motif I then determined the weighted loss probability. The weighted loss probability for each motif, given its codon usage, was then compared to the expected loss probability given the overall codon usage distribution for the relevant amino acids. A codon that through selection uses more robust codons on average, would thus have a lower weighted loss probability than the expected background. As an example, if a key motif site requires lysine and the background codon use at lysine positions is 50% AAA and 50% AAG, the weighted probability is 0.5*P(AAG)+0.5*P(AAA). If a motif site uses 20% AAA and 80% AAG the weighted probability thus would be 0.8*P(AAG)+0.2*P(AAA). For each motif this was estimated at each defined residue position to yield the total weighted motif probability (for details, see Methods 3.5.3.3).

To evaluate the bias in the motifs in my dataset I thus compared the loss probabilities between these sets and plotted the log ratio of the motif compared to background probability (Figure 3.13, panel 2). A positive score thus indicates enrichment of robust codons and a negative score denotes enrichment of non-robust codons. I also compared the absolute robustness of each motif since it is possible that the expected background codon usage is also the most robust. This score is calculated as ln(Observed-Lowest/Highest-Observed), yielding a positive score for more overall robust codon use and a negative score for more overall non-robust codon use (Figure 3.13, panel 1).

Overall, these functional motifs are biased for robustness, both in absolute mutation probability terms, and compared to the expected codon usage in influenza (Figure 3.13). Surprisingly, the only motif not using robust codons is the LC3 binding motif in M2 (M2_LIG_LIR_GEN ([E,D]xx[W,F,Y]xx[I,L,V]) which is biased towards highly non-robust codons despite being highly conserved and functionally important. However, this motif is encoded in an overlapping reading frame with M1 which is likely to result in overlapping selection pressures for codon choice between the two proteins.

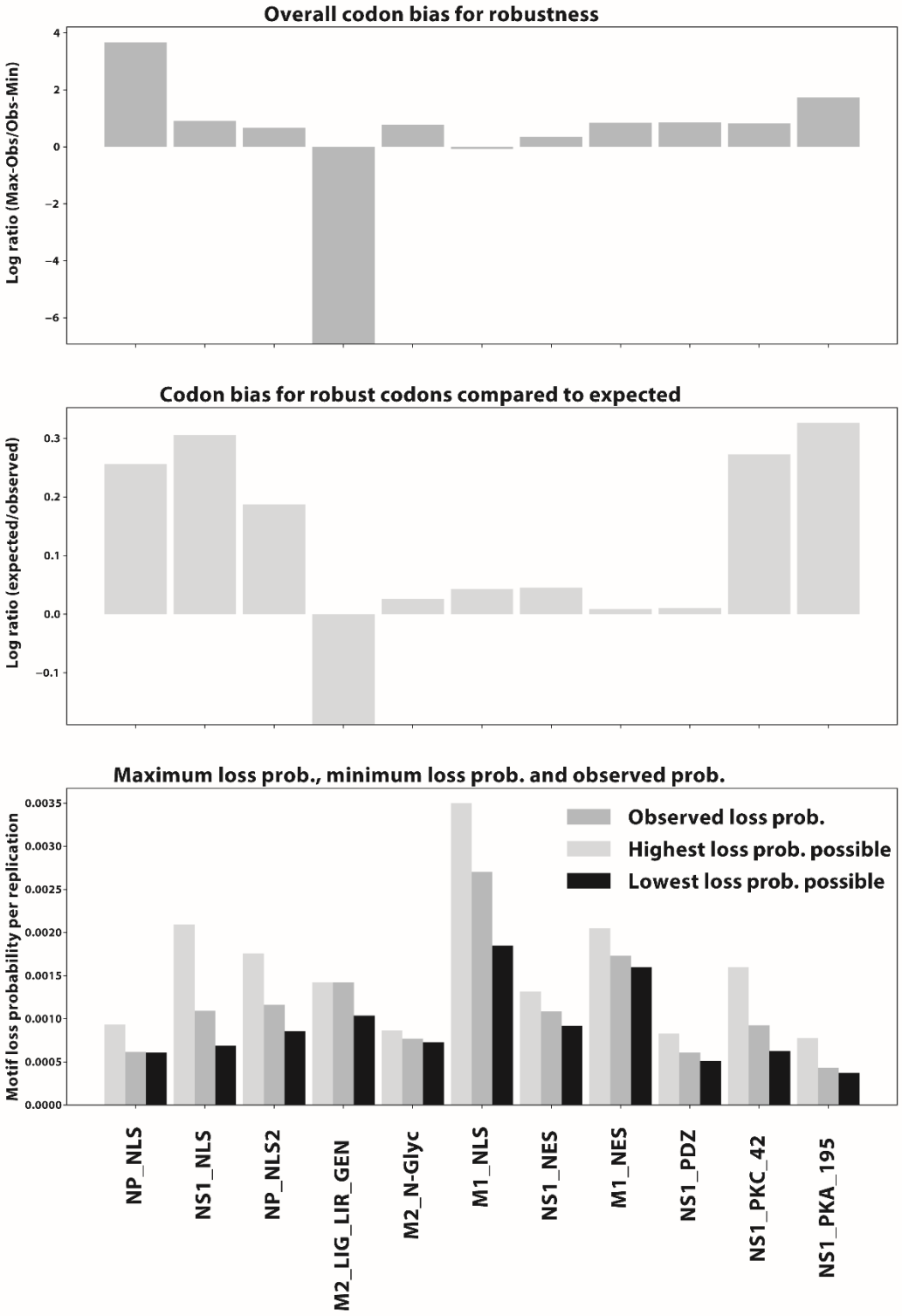


Figure 3.13. Codon usage bias for robustness in conserved functional motifs in influenza. Top panel: the bias towards robustness and away from non-robustness. Log ratio of codon robustness and codon loss. A positive value indicates motifs favour robust codons. Middle panel: Codon bias compared to prevailing codon use in influenza. The log ratio of the observed versus expected indicates if motifs favour codons that are more robust but less used in influenza. A positive value indicates bias towards robustness. Bottom panel: Comparing the loss probability for the least robust (light grey), most robust (black) and observed (dark grey). The closer the observed is to the most robust (lowest loss) the more optimised its codon use.

The codon usage at M1 NLS indicates neither robust nor non-robust codon usage in absolute terms, such that different codon usage could improve robustness. However, compared to the expected codon usage for those positions it turns out to be more enriched than expected if codon choice was random. In contrast, M1 NES and NS1 PDZ motifs use predominantly the codons expected by chance, but these are close to the most robust ones and no significant fitness gain could be made by using other codons. These observations illustrate some of the potential conflicting selection pressures across influenza sequences and highlight the complexity of this optimisation problem through evolution. To get an idea of the codon usage more directly I also visualised the codon distribution for each relevant site in the nuclear localisation motifs in NS1 (Figure 3.14). I plotted the total codon count for each codon observed at each position across NS1 sequence history. Nuclear localisation sequence 1 is highly conserved and uses a classical consensus sequence made up by 4 basic residues (35,37,38,41) and a hydrophobic residue (36) (Melén *et al.* 2007). Residues 35,38 and 41 have been suggested to be the most important for importin α binding and nuclear transport in some strains (Melén *et al.* 2007). The data reflects this selection through the much more limited substitutions allowed at these positions. This import motif uses highly robust codons given the theoretical space of allowed codons in the different positions, in particular at positions 35, 37 and 38.

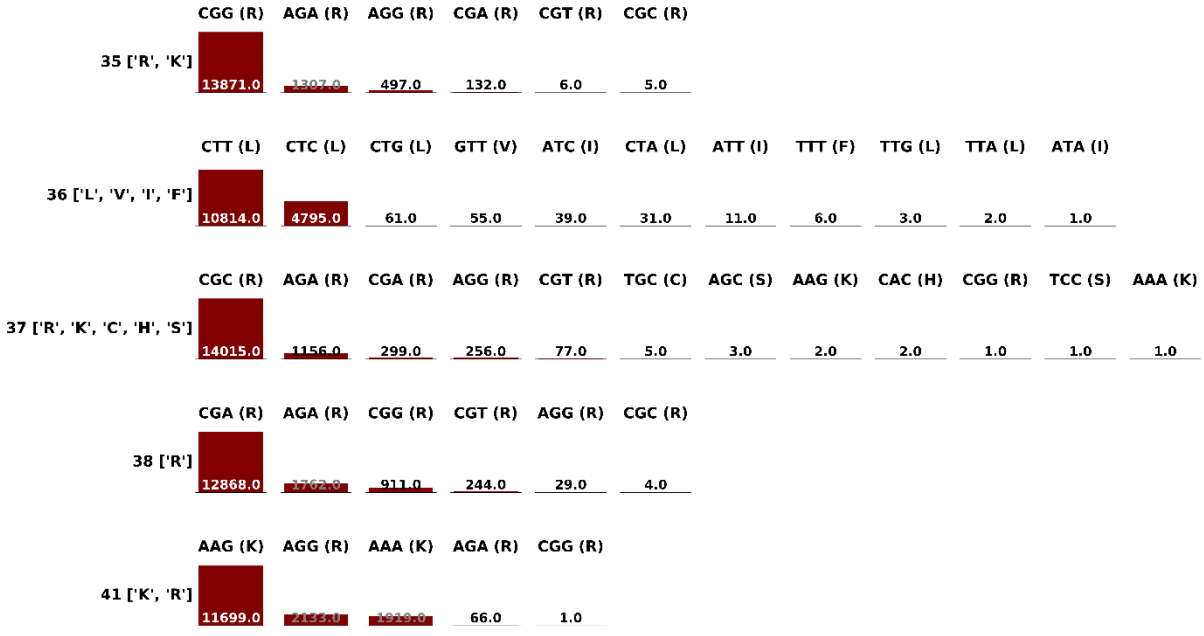


Figure 3.14. Codon usage at all positions in the nuclear localisation motif of NS1. At each position, each bar displays the total count of that codon observed across the phylogenetic tree of NS1 including reconstructed nodes. Overall, robust codons are preferred at positions 35, 36, 37 and 38. Position 36 is expected to be less important for importin binding than the basic residues. Residue 41 uses codons with low robustness predominantly, assuming lysine and arginine are equally fit in that position.

Assessing robustness within generally hydrophobic positions such as 36 in NS1 is a bit less straight forward. In general, when all hydrophobic residues are allowed and functional (LVIFM) there is minimal mutational differences between them as they all share the same replaceable 1st 2nd and 3rd positions

and thus all the same mutations are functionally silent or non-silent. However, if a limited set of the hydrophobic residues are allowed e.g. [LIV] or [LFI] different codons will generally have different robustness (depending on nucleotide specific mutation rates).

Within this hydrophobic position the predominant use of relatively robust leucine codons indicates a potential fitness impact from robust codon choices, however this is subject to further analysis and interpretation. It appears as if selection pressures on individual codons are lower here which is consistent with a wider variety of codons used and the structure of the genetic code around the codons for hydrophobic amino acids having higher inherent systemic robustness.

Overall, these preliminary results indicate that the observations made for codon choices and robustness in Ser/Thr sites might similarly hold true within whole motifs, i.e. that functional, conserved motifs have a codon bias for more robust codons. Selection appears to optimise viral fitness at key functional motifs by making them more robust to mutations, and thus less likely to be lost through replication. As our definitions of motif residues and contexts improve, this hypothesis can be expanded on and further tested on a wider range of motifs and organisms.

3.3.7. Influenza A favours robust stop codons

The observations of codon bias in functional sets of amino acids and codons in motifs led me to also look at robustness and bias in the stop codons. In many ways the three stop codons are organised similarly to a motif and therefore are likely to share some of the mutational properties of motifs. Stop can be encoded by three different codons that can differ in both the second and third position. The mutational overlap is different for the different codons and they use nucleotides with different mutation rates.

Stop codons have historically been ignored in the vast majority of sequence and codon bias studies, however a small number of published papers have found that in many organisms as well as in highly expressed household genes there is purifying selection towards TAA and usage is therefore biased towards that codon (Belinky *et al.* 2018; Korkmaz *et al.* 2014). The authors however conclude that there is still no explanation for why TAA should be under purifying selection. In this section I propose a simple mechanistic explanation for this selection pressure based on mutational robustness similar to what has been presented here for motifs.

Firstly, I wanted to determine if influenza also has codon usage bias at stop codons. I analysed all stop codons used by the different influenza proteins in all of the strains available (Figure 3.15). Influenza heavily favours the stop codon TAA over TAG or TGA. This has not been previously described in influenza but corroborates data from *E. coli* and other bacteria where stop codon usage has been looked at. In influenza the proteins NP, NS2, M2, HA and NA use almost exclusively TAA. Proteins PB1 and PB2 use exclusively TAA in strains that infect humans, but interestingly TAG in strains that infect

birds. PA uses almost exclusively TAG across hosts. M1 and NS1 are the only two proteins to use predominantly TGA. Interestingly, both M1 and NS1 are encoded through overlapping reading frames with M2 and NS2 respectively. This causes the stop codon to be in a part of the protein which also encodes residues in alternate reading frames which is likely to have other selection pressures on the combined codons in both reading frames.

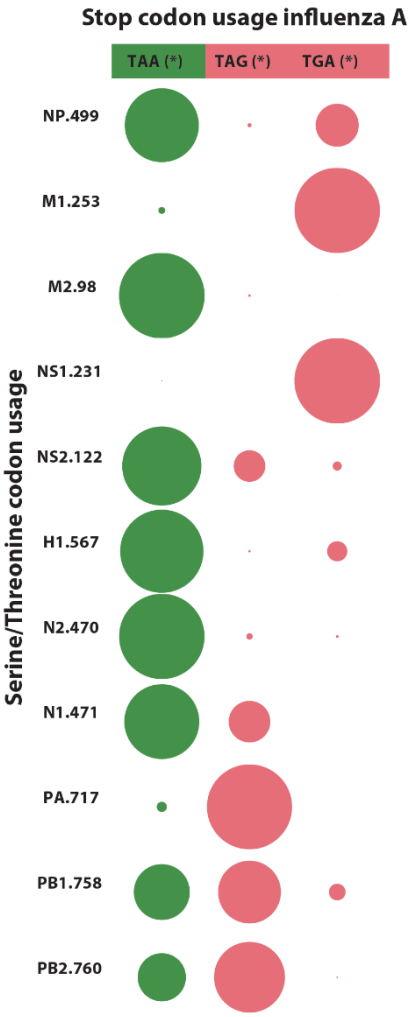


Figure 3.15. Codon usage at stop codons in influenza A proteins. There is a strong bias for TAA codons that are more robust to loss (green). M1 and NS1 diverge from this trend, however their stop codons are in overlapping reading frames with M2 and NS2 respectively giving rise to conflicting fitness pressures

I applied the probability calculation to investigate which of these codons would be the most robust given the available codon space and the nucleotide specific mutation rates. TAA comes out as the most mutationally robust of the three stop codons. There is no difference between TAG or TGA. TAA is 2-fold more robust with an estimated loss frequency of $3*10^{-4}$ per replication compared with $6*10^{-4}$ per replication for TAG and TGA. Given the typical population size and number of replications per cell during infection these rates are likely to cause a significant difference in the number of viral particles with genes lacking stop codons during viral proliferation, which may directly impact fitness.

To explore how the populations dynamics are likely to be affected by stop codon choice I ran simulations in the same way as for motifs in chapter 2. Sequences with TAA had significantly less stop codon-loss and a more fit population, assuming spontaneous stop-codon loss is detrimental to viral fitness. When using TAA, proteins gave rise to 50% fewer mutated progeny overall, compared to the other two stop codons. Overall, in sequences using TAA about 0.2% of progeny experienced stop codon loss when considering a single protein. In sequences using TAG or TGA this number was 0.5%. In addition, since there are 10 stop codons per viral particle the fitness outcome of optimisation can be compounded. If we consider that each influenza particle needs all 10 intact stop codons to be functional, using all TAG leads to progeny loss in over 5% of particles compared to 2%-2.5% for using TAA exclusively. This difference represents a dramatic fitness change resulting from a single codon difference across proteins. The increased number of faulty proteins is also likely to reduce fitness of the entire viral quasispecies, as the protein load, potentially reduced growth rate, and the incorporation of faulty proteins in virions also reduce successful viral spread, in addition to the direct fitness effect from the non-viable gene segments. It has been established previously that very small fitness differences are enough to have effects on synonymous codon choices through selection. In *Salmonella,* the selective

disadvantage when using suboptimal codons (for the optimal rate of translation) was shown to be around $0.2\text{-}4*10^{-4}$ which thus equates to ~0.02-0.4% slower growth compared to the optimal codon choice (Brandis & Hughes 2016). As the expected effect from stop codon choice is >2% better growth from viable progeny alone, it is likely to provide enough of a fitness benefit to be selected for. In addition, with the added expected growth disadvantage resulting from the expression of mutated genes likely to be on the order of 0.2% (due to expression from the mutated gene segments) as a result of mutations, that would likely compound and further improve the selective advantage of using less mutable stop codons. This suggests the differences seen here are likely to significantly influence stop codon choices in viral proteins to reduce the load of faulty proteins and increase the overall successful progeny produced.

It is important to consider that stop codons also have different termination efficiency, and that stop-codon readthrough is higher when using TGA and TAG compared to TAA (Dabrowski *et al.* 2015). This is likely to further increase the selection advantage from TAA when stop codon readthrough is detrimental to the infection. The effect from different readthrough propensities is likely to be smaller than the effect from mutational robustness in influenza, as basal readthrough rates are expected to be on the order of 0.01-0.1%, but can be higher depending on sequence context and sometimes functionally important readthrough in some viruses (Csibra *et al.* 2014; Floquet *et al.* 2012). However, little is currently known about influenza specific termination efficiency.

The impact of this stop codon bias and its fitness effect is likely to be relevant in many organisms and could be an additional important insight into the effects of codon choice on fitness. This is corroborated by the finding that many organisms show purifying selection for TAA. In the context of eukaryotes and disease it is also possible that stop codon bias can impact protein stress and misfolding as well as aggregation loads in cells which may be a contributing factor to disease severity.

3.3.8. Using codon choice and relative conservation to predict unexplored functional motifs in influenza

In chapter 2 and 3 I have described a series of observations that add information about the functionality of motifs based on sequence and mutation rate data. Firstly, the observation that conserved, functional motifs tend to favour a more limited set of more robust codons as predicted by the nucleotide specific mutation rate and the available codon space for each motif. Secondly the observation that by simulating expected motif loss dynamics over phylogenetic trees it is possible to estimate how frequently motif-like sequence patterns are lost when not under selection in the tree. Comparing this loss frequency to the observed loss frequency of known functional motifs informs whether the motifs are lost less than expected which can indicate selection of a functional motif. These observations are factors that can be taken into account when predicting whether a motif-pattern identified in sequence searches is likely to be a non-functional instance compared to a functional motif

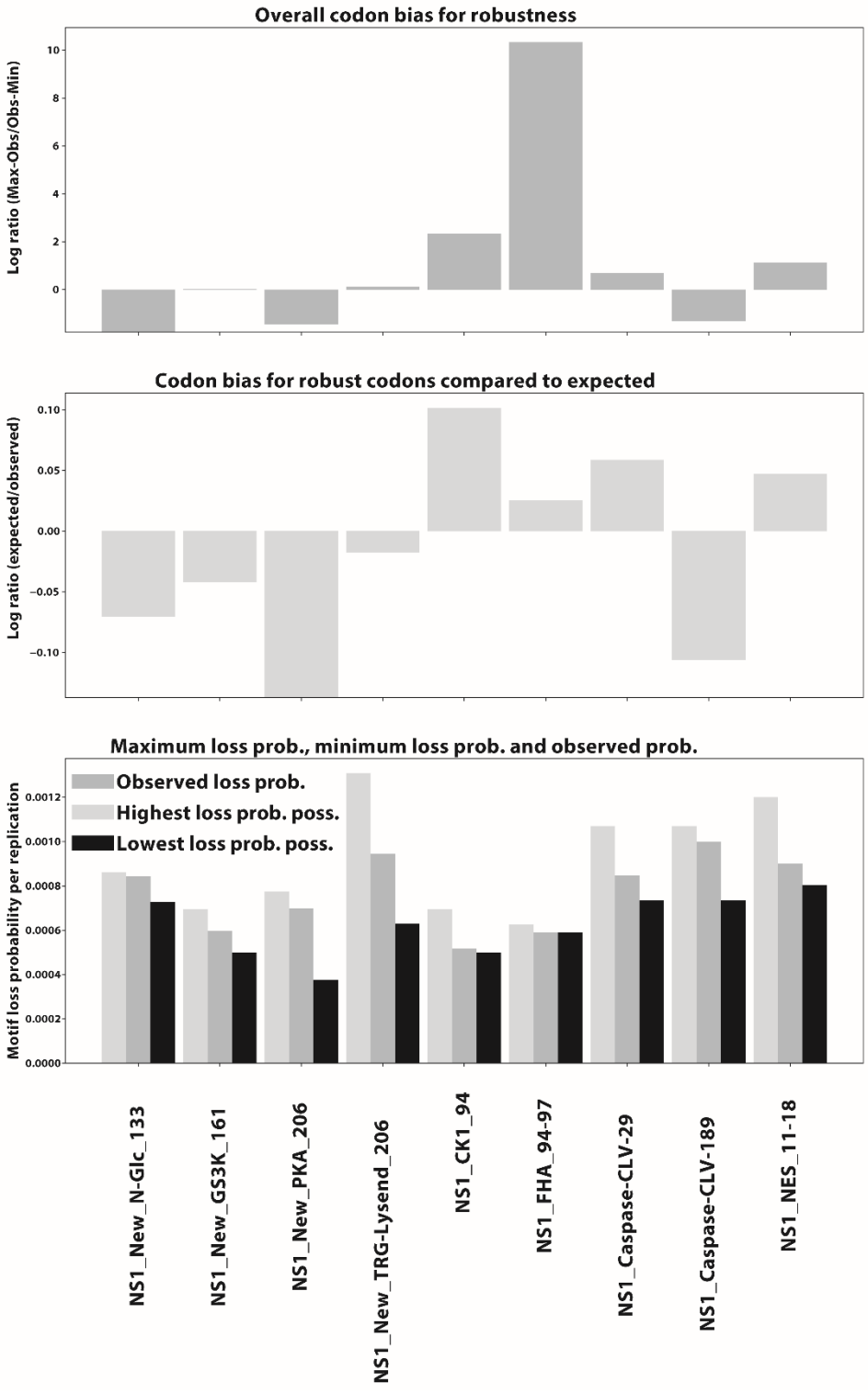


Figure 3.16. Assessing codon usage at putative motif sites to determine functionality through selection pressure. Top panel: the bias towards robustness and away from non-robustness. Log ratio of codon robustness and codon loss. A positive value indicates motifs favour robust codons. Middle panel: Codon bias compared to prevailing codon use in influenza. The log ratio of the observed versus expected indicates when motif favour codons that are more robust but less used in influenza. A positive value indicates bias towards robustness. Bottom panel: Comparing the loss probability for the least robust (light grey), most robust (black) and observed (dark grey) codon choice. The closer the observed is to the most robust (lowest loss) the more optimised its codon use.

that is under selection. Codon choice has the benefit over existing methods based simply on sequence conservation that it does not rely on related sequences to determine if the motif is using robust codons; only the motif definition and a nucleotide specific mutation rate is needed. Comparing expected loss in a tree with observed loss requires a sequence history as we often have for many viruses, but is more informative than simply looking at conservation thresholds since motifs that are only conserved in a subset of strains or are recent in the history of the virus can be more accurately analysed. Conservation scores would be expected to be different for different types of motifs to indicate significance, but by instead comparing expected rates of change of a motif with the observed changes, selection on a specific motif can be inferred irrespective of motif size and complexity.

Since motifs influence many aspect of influenza protein function, having accurate predictions of potential new motifs can be very helpful in designing experiments to investigate virus-host interaction dynamics. To generate a list of high confidence predictions of putative motifs for current human NS1, I have here collected a list of motifs found through simple sequence matches but have not yet been described or tested experimentally. They have varying degrees of conservation across all strains but all are at least conserved in some branches of influenza A. I performed the same analysis as in section 3.3.6, when looking at codon bias in conserved motifs. I have determined the codon usage across strains for all the motifs and compared the weighted loss probability to the expected background (Figure 3.16). The majority of these putative motif matches show no codon bias towards robustness which could be an indication that it is simply a spurious match. In comparing to the expected background, similarly most motifs use less robust codons than would be expected by chance in influenza. However, two motifs stand out as showing bias towards robust codon use. A CK1 binding motif and an FHA binding motif, which overlap in the sequence. The CK1 motif is a phosphorylation motif and FHA is a domain that that recognises phosphorylated threonines. This codon bias suggests these sequences could be good targets for further investigations. They also offer an intriguing switch through overlapping motifs, that could regulate complex protein behaviour.

In addition to looking at the codon usage bias, I also compared how frequently these sequences are lost over the phylogeny, and compared it to the expected loss for those sequences. Both CK1 and FHA are lost less than would be expected, again indicating the specific combination of residues is being selected for. However, several of the other motifs with no codon bias are also lost less than expected. Depending on the complexity of the motif consensus it is thus likely that several of the predictions are false positives, despite appearing to be selected for simply due to selection for those residues at those positions in a non-motif context. Encouragingly though, some of the putative motifs included here, that should not be functional in NS1 including N-Glc and TRG LysEnd are lost as frequently or more frequently than expected. These motifs are very unlikely to be functional in NS1 given its functional context. These results indicate that this approach can help inform which motifs or sequence patterns

are under selection and be used to prioritise likely functional motifs in conjunction with other methods.

3.4. Discussion

In this chapter I have explored the evolutionary dynamics of functional short linear motifs in influenza A. I have looked at the impact of codon choice and prior sequence on the mutational landscape of virus strains and how those mutations impact motifs and importantly modification sites in particular. I have found that among residues in motif contexts, there is a significant codon bias where more robust residues are favoured and less robust residues are used less than would be expected. These findings suggest that the predicted fitness impact of certain codon choices, as established in chapter 2 through simulations, could be used to inform the expected impact on real codon choices in viral strains. In particular, the codons at positions in motifs that confer the highest motif binding selectivity and fitness appear to be under the most intense pressure to reduce the detrimental impact of mutations. They consequently tend to use codons that are more robust to retaining the functional motifs. This pattern is seen very clearly among the Ser/Thr sites within phosphorylation motifs where phosphoregulation is important for overall viral fitness.

Interestingly, this level of codon bias is also found for the stop codons. The observation that TAA codons experience purifying selection was described recently (Belinky *et al.* 2018). However, the authors had no explanation for the underlying cause for this selection. The simulations and calculations here support a model where mutational robustness is maximised through stop codon choice. Using TAA reduces mutational loss of stop codons in mRNAs and descendant vRNAs which could lead to increased infectious fitness. It is also plausible that this effect is compounded by the termination efficiency of the stop codons, as TAA tends to have the least readthrough frequency. The organisation of the stop codons in codon space, like motif codon spaces, lead to differences in the number of functionally silent mutations. The codons also have nucleotide variations at both position 2 and 3 leading to different mutational outcomes and different mutational frequencies depending on which stop codon is used. Interestingly in NS1 and M1 a less robust TGA is favoured however, these stop codons are in an overlapping reading frame to NS2 and M2 respectively. In addition, using TGA might also contribute to them being lost more frequently in different strains as suggested by the many variations in C-terminus length in these proteins. In those instances, the frequency of loss could have contributed to an opportunity for innovation of new functionality (e.g. the new NLS formed in the NS1 extended sequence) (Melén *et al.* 2007).

The pattern at other important motif associated sites also indicate selection for robustness, albeit more weakly. The pattern at motif associated basic sites which is common in Nuclear Localisation sequences for example show a bias towards robust codons, however less robust codons are more commonly seen than in e.g. Ser/Thr or stop codons. There are multiple possible reasons for this; the most common arginine codon in mammals is AGA and AGG and it presumably confers some translational benefits and it is consequently the most used codon in influenza A too, despite being among the least robust (Komar 2016). In localisation sequences at conserved Arg/Lys positions, AGA appears to be used less frequently than elsewhere in the viral genome, based on preliminary observations, suggesting there is a bias for more robust codons there too. However, non-robust codons are still seen in highly conserved residues in some of the NLSs. Another potential factor that can impact this is the reduced selection pressure on individual residues in the basic sites in NLSs. Most influenza NLSs are made up of 4-6 basic residues but most experiments have reported functional nuclear localisation with as few as 3 basic residues suggesting that this redundancy acts as additional layers of functional robustness (Kosugi *et al.* 2009).

In general, evolution achieves robustness in multiple ways. Overall, robustness simply means to reduce the functional impact of mutations. Selecting robust codons is one way to do this, however influenza uses other well-established modes of achieving robustness. As discussed above for NLSs, having additional residues can increase functional redundancy. This is also seen at many phosphorylation sites. It is well established that sequences of serines and threonines in close vicinity can act as redundant phosphosites increasing robustness. Indeed, in influenza, groups of 2-4 Ser/Thr are very common and in fact, I have observed that several of the residues at these redundant sites tend to use less robust codons whereas non-redundant sites favour more robust codons. There are not enough of these sites to confirm whether this is a statistically robust observation, or simply a coincidence currently, but it would be an interesting further analysis to make as more data becomes available. These are alternative ways of reducing mutational effects and increasing fitness and they can likely buffer each other's impact to reach an optimal functional threshold given the population dynamics in influenza infections.

In addition to codon bias for robustness as explored here, there are previously characterised sources of codon bias in sequences that will both synergistically enhance the effect seen from motif robustness or sometimes act counter to the selection for motif robustness. Different codons have been shown to translate at different rates which is a known source of codon bias towards more optimal codons (Komar 2016). As the sequences exist as RNA as well as proteins, there will also be important pressures to maintain nucleotide binding sites and secondary structures that can affect fitness given particular codons and nucleotide sequences used (Carlini *et al.* 2001). In influenza viruses as well as several other viruses, reading frames may overlap in which case the complex interaction between selection at several levels of sequence and function will be very difficult to predict. This work provides an im-

portant first insight into this new phenomenon of codon bias for the mutational robustness of important linear motifs, however future work will have to look more closely to try to disentangle the selective effects from all these different sources.

High mutational robustness is also not always a property that will confer increased fitness. For highly evolvable sequences like HA and NA in influenza A it has previously been argued that high mutability is maintained, as robustness of certain features might lead to higher conservation of exposed antigens, allowing the immune system to neutralise the virus. Glycosylation sites are known to be important for changing and masking sites recognised by antibodies thereby allowing the virus to escape detection. Glycosylation sites in these proteins are rarely well conserved. In my analysis of known glycosylation sites in HA, the majority appeared to actively use less robust codons. If that is a consequence of high mutability or if it is actively selected for remains to be determined, but this observation agrees with previous studies that show that high mutability is favoured at these functional sites in particular in HA (Plotkin & Dushoff 2003). This prospect is interesting from an evolution perspective as it suggests that evolution can use the codon space at functional sites and modulate evolvability for each feature depending on codon choice, thus allowing a single protein to have regions both of robustness and high evolvability depending on selection pressures. The fact that fast evolving RNA viruses are able to fine tune the evolvability of their sequence in this manner is an intriguing idea and it remains to be seen if it is restricted to organisms that have similar properties to influenza or if this is also seen in other organisms and systems such as cancer and other pathogens. This idea of modulating the evolvability and robustness of individual features is in contrast to how these properties have been studied so far; generally these properties are applied to whole proteins when analysing their evolutionary consequences. The observations here can open up new ways of thinking about evolution of protein features and also may impact the way we understand how systems properties change and evolve.

Extending the findings here to thinking about how new motifs might evolve it is intriguing to consider the possibility that sequences can be selected for their rate of evolving a new motif such that the motif is consistently present in only a small part of a population. This could enhance the fitness for the overall infection through interacting with particular host systems without ever being visible in the sequence of the progenitor strain. This could theoretically evolve given the right codons being chosen such that the combined mutational frequencies lead predictably to motif emergence, however it might be unlikely to exist. These kinds of equilibria are common in populations of organisms in nature when considering competing behavioural phenotypes for example, as in the case of evolutionarily stable strategies and game theory (Maynard Smith 1974; Smith & Price 1973). Whether this can happen in the selection for low frequency emergence for particular sequence features remains to be seen.

To further investigate the role of prior sequence in the evolvability and innovation of new motifs, in the next chapter I will explore when and where motifs evolve in influenza A and attempt to predict the motif innovation landscape in current circulating strains.

3.5. Materials and Methods

3.5.1. Mutational outcomes analysis

3.5.1.1. Datasets

To determine the mutational outcomes for the different codons in influenza, alignments were downloaded from the curated dataset by Worobey *et al*, (2014), as in the analyses from chapter 2 (see Materials and Methods 2.7.1). For this analysis I used the alignments for NS1, M1 and NP, which were filtered to exclude sequences with gaps, ambiguous positions and identical sequences. The final size of the filtered alignments were 9921 strains in M1, 8663 in NS1 and 9592 in NP.

3.5.1.2. Building phylogenetic trees

Phylogenetic trees were constructed in RAxML using the same approach as described in section 2.7.2 (Stamatakis 2014). The nucleotide sequence alignments described above were used as input. Due to influenza reassortment, it is more accurate to treat each protein as a separate genome and create individual phylogenies (Worobey *et al.* 2014). For each alignment the following command line argument was used:

raxmlHPC-PTHREADS-AVX -T 8 -m GTRGAMMA -s ../Input-Alignment.fasta -f a -p 12345 -x 12345 -# 100 -n Outputfile

3.5.1.3. Ancestral sequence reconstruction

To determine the mutational outcomes of the different codons I needed to be able to track sequences across a phylogenetic tree and determine when mutations have happened in a directional manner. To construct trees that incorporate the known times during which the influenza strains were active, and to use this information to reconstruct accurate ancestral sequence states the python software TreeTime was used (Sagulenko *et al.* 2018). TreeTime uses a pre-existing tree, a sequence alignment and a file containing the known time each strain was active in the population. The RAxML trees were used alongside the nucleotide alignments. TreeTime creates an improved tree by incorporating the time information, and also reconstructs accurate ancestral sequence for the known nodes. TreeTime uses a maximum likelihood approach to infer divergence times of the strains, which is improved by the timestamp information, and uses a general time reversible substitution model estimated from the alignment for sequence reconstruction. The command line argument used for the trees and alignments was:

timetree_inference.py --aln FLUALIGNMENT.fasta --dates FLUYEARS.csv --tree FLU-RAXML-TREE --gtr infer --reroot best --optimize_branch_length --verbose 6 --Tc skyline

To ensure that the reconstructed sequences were viable I checked that they all translated without internal stop codons. If any did I pruned that lineage from the tree using DendroPy v.4.4.0 (Sukumaran & Holder 2010). I did random spot checks of the reconstructed sequences using NCBI BLAST to determine if they were reasonably reconstructed sequences given the time period for that state. 100% of all sequences verified using this method (~30) were highly similar (>99.8%) to sequences around the same time period, and many matched 100% with sequences in the correct year, which indicates that the reconstructions are generally reflecting sequence history.

3.5.1.4. Determining codon substitutions

To determine how each of the 61 codons have been substituted in influenza sequence history I traversed the TreeTime generated tree on a codon by codon basis. For each codon I determined each instance of when a descendant (child) node had a mutation compared to the closest ancestral sequence and I assigned which codon was in the descendant compared to the ancestor. For each codon class I thus created a matrix of substitution counts, and in the end had a 61 by 61 matrix with a total count of each instance of a substitution that had occurred for each codon. To get comparable substitution rates I also determined each instance when the codons did not substitute between ancestor and descendant as a normalisation factor. The number of substitutions observed for each codon was divided by the number of times that codon did not substitute, to get a relative substitution rate to compare between the codon classes. The final frequency presented is the ratio multiplied by 100 000, to get the number of substitutions per 100 000 codons to get comparable numbers in a good range.

3.5.2. Motif evolution across a phylogenetic tree

3.5.2.1. Datasets

For this analysis the NS1 alignment and TreeTime inferred tree as described above were used.

Motif consensus sequence definitions were downloaded from ELM as a CSV file containing all classes defined there as of 2018 (Gouw *et al.* 2017). Motifs with variable length definitions were either left out of the document or modified to different versions for each possible length where viable. This was to enable accurate probability calculations, which is not possible for variable sequence lengths currently.

The motifs used in this analysis are listed in Table 3.10. These sequences include known functional motifs and also putative motifs identified as consensus sequence matches. The putative motifs were selected manually to reflect a range of motif classes and a range of levels of conservation. All putative motifs were conserved in between ~20-80% of strains.

Table 3.10. Dataset of known and putative motifs. Motifs were identified through regular expression matches in at least 20% of strains. Several motifs have been described as functional in the literature.

MOTIF	PEPTIDE	POSITION	RNA
CLV_C14_CASPASE3-7	ELSDA	(25, 30)	GAACTGAGTGATGCC
CLV_PCSK_FUR_1	RLRRD	(34, 39)	AGACTCAGAAGAGAT
MOD_PKB_1	RGRGNTLGL	(43, 52)	AGGGGAAGAGGCAATACTCTTGGTCTA
MOD_PKA_2	GRGSTLG	(44, 51)	GGAAGAGGCAGCACACTTGGA
CLV_PCSK_SKI1_1	KILKE	(66, 71)	AAGATTCTGAAAGAA
MOD_GSK3_1	SSETLRMT	(72, 80)	TCCAGCGAGACACTTAGAATGACA
MOD_NEK2_1	LKMTMV	(76, 82)	CTTAAAATGACCATGGTC
MOD_PKA_2	LRMTIAS	(76, 83)	CTTAGAATGACAATTGCATCT
DOC_PP2B_LXVP_1	LKMP	(76, 80)	CTTAAAATGCCG
DOC_USP7_MATH_1	ALASR	(83, 88)	GCACTTGCTTCGCGG
CLV_PCSK_SKI1_1	RYITD	(87, 92)	CGGTACATTACCGAT
LIG_SH2_STAT5	YITD	(88, 92)	TACATTACCGAT
MOD_CK2_1	TDMSIEE	(90, 97)	ACCGATATGAGCATAGAGGAA
LIG_FHA_2	DMTIEEL	(91, 98)	GACATGACTATTGAGGAATTG
LIG_TRAF2_1	SIEE	(93, 97)	AGCATAGAGGAA
DEG_APCC_DBOX_1	SRDWLMLIP	(98, 107)	TCAAGAGATTGGTTAATGCTCATTCCC
TRG_LYSEND_APSACLL_1	DWLMLI	(100, 106)	GATTGGTTAATGCTCATT
LIG_TRFH_1	YMLMP	(102, 107)	TACATGCTCATGCCA
MOD_N-GLC_1	KNITLK	(125, 131)	AAAAACATCACGTTGAAA
LIG_PEX14_2	FSVLF	(133, 138)	TTCTCTGTCCTATTT
LIG_FHA_1	LETIVLL	(140, 147)	CTAGAGACTATAGTATTGCTA
MOD_PRODKIN_1	AEISPIP	(157, 164)	GCTGAAATATCTCCCATTCCT
MOD_GSK3_1	AEISPIPS	(157, 165)	GCTGAAATATCTCCCATTCCTTCT
LIG_SH3_3	SPIPSMP	(160, 167)	TCTCCCATTCCTTCTATGCCA
DOC_USP7_MATH_1	PIPSM	(161, 166)	CCCATTCCTTCTATG
LIG_SH2_SRC	YEDV	(170, 174)	TATGAGGATGTC
MOD_PLK	NDNSIRA	(187, 194)	AATGATAACTCAATTCGAGCG
MOD_PKA_2	IRASENI	(191, 198)	ATTCGAGCGTCTGAAAATATA
MOD_CK2_1	AWRSSNE	(201, 208)	GCTTGGAGAAGCAGTAATGAG
MOD_CK1_1	SNETGGP	(205, 212)	AGTAATGAGACTGGGGGACCT
DOC_PP2B_LXVP_1	LPLP	(206, 210)	CTTCCACTCCCT
LIG_INTEGRIN_ISODGR_1	NGR	(208, 211)	AATGGGAGA
MOD_PKA_2	GRPSLPP	(209, 216)	GGGAGACCTTCACTACCTCCA
MOD_CDK_SPXXK_3	PPLTPKQK	(211, 219)	CCTCCACTTACTCCAAAACAGAAA
DOC_WW_PIN1_4	PPLTPK	(211, 217)	CCTCCACTTACTCCAAAA
LIG_WW_2	PPLP	(211, 215)	CCTCCACTCCCT
LIG_MYND_1	PPLPP	(211, 216)	CCTCCACTCCCTCCA
DOC_CKS1_1	PLTPKQ	(212, 218)	CCACTTACTCCAAAACAG
DOC_USP7_UBL2_3	KQKRK	(216, 221)	AAGCAGAAACGGAAA
MOD_SUMO_FOR_1	IKSE	(225, 229)	ATTAAGTCAGAA

3.5.2.2. Simulating motif loss

I also determined the expected motif loss rate by using the same tree topology (NS1) and then simulating all the nucleotide sequences for the motifs above, over that topology. I used Pyvolve as outlined in chapter 2. In this instance I used the state_freqs:

freq (A): 0.317235
freq (C): 0.200558
freq (G): 0.243714
freq (T): 0.238493

and I used the custom_mu for the relative mutation rates:

rate A <-> C: 1.792412
rate A <-> G: 9.542882
rate A <-> T: 0.796239
rate C <-> G: 0.263115
rate C <-> T: 9.340186
rate G <-> T: 1.000000

These are the rates based on the estimates from the phylogenetic tree rather than the biochemical mutation rates as measured. They are symmetric meaning some information is lost making it impossible to differentiate e.g. A-to-C vs. C-to-A. Due to the size of the tree and the number of sequences simulated, 50 independent simulations were run.

3.5.2.3. Determining loss rates

For both the motif dynamics in the sequences record, and in the simulated datasets motif loss was calculated in the same way. It also uses the same approach as for determining the motif loss in simulations used in section 2.7.5. At each node in the tree I determined motif presence or absence through a regular expression sequence match in the correct position in the alignment. I then traversed the tree and determined every instance when a mutation between an ancestral state and the descendant caused a motif to be lost. I normalised the number of motif losses by the number of times the ancestor and the descendant both had the motif, i.e. non-loss events, in a similar manner to what I have described before in 2.7.5 and 3.5.1.4

3.5.2.4. Calculating loss probabilities

The probabilities were calculated as defined in section 2.3.3, which I will reiterate here. For an initial codon from Ser/Thr e.g. AGC to substitute into a non-Ser/Thr codon e.g. AAT, the probability is calculated as the product of the probabilities for all the nucleotide substitutions for each codon outcome, i.e.

$$P(AAT|AGC) = P(A|A) * P(A|G) * P(T|C)$$

To determine the loss probability, I thus determined the probability of gaining a non-Ser/Thr codon. All the probabilities for the individual codons outside Ser/Thr codon space are added together i.e.:

$$P(NotSerThr|AGC) = P(AAT|AGC) + P(AAC|AGC) + P(CGC|AGC) \ldots$$

The total probability at each defined motif position is ultimately multiplied to get a motif loss probability.

In this instance the substitution probabilities were determined by the substitution frequencies from the phylogenetic tree. This is an alternative method to the per-replication probability, developed in collaboration with my colleague Greg Slodkowicz. It calculates the probability of any given substitution after a specified amount of branch length in the phylogenetic tree. It is based on a markov-model in which a rate matrix, Q is determined from the phylogenetic tree based on the rates for each nucleotide substitution (Yang 2014). The probabilities of all nucleotide transitions can then be determined by exponentiating the Q matrix

$$P(t) = e^{Qt}$$

This results in a P matrix of the probabilities of all nucleotide substitutions after branch length t in the phylogenetic tree.

3.5.3. Determining codon usage bias

3.5.3.1. Datasets

For phosphorylation Ser/Thr:

The filtered alignments from before for NP, NS1 and M1 were used. In addition, specific alignments for the overlapping reading frames of M2 and NS2 were constructed using the coding region definitions from NCBI. Only sequences of full and equal length were included to improve alignment quality. This still included >95% of NS2 and M2 sequences form the original alignments. These alignments were the same length as NS1 and M1 respectively The alignment for PA was also included (10651 sequences).

For glycosylation Ser/Thr:

Sequences for H1 (H1N1), N1 (H1N1) and N2 (H3N2) were manually downloaded from the NCBI Viral Genomes Resource (Brister *et al.* 2014). All coding sequences spanning the years 1900-2015 were downloaded, excluding any identical sequences. The fasta files were then filtered to only include sequences of the same length and without gaps or unknown nucleotides. H1, N1 and N2 residues were then respectively aligned using MAFFT (Katoh & Standley 2013). The final alignments contained N1: 8020, N2: 9009, H1: 9783 sequences. The H3 alignment was downloaded from Worobey *et al*, (2014) and contained 4026 sequences.

For stop codons:

The same alignments as described above were used (NS1, NS2, M1, M2, NP, PA, H1, H3, N1, N2) with the addition of PB1 and PB2 alignments from Worobey *et al*, (2014).

For set of whole functional motifs:

The alignments described above for NS1, NP, M1 and M2 were used to assess codon usage in the set of functional motifs (described below).

The functional motifs analysed were:

NP_NLS
NS1_NLS
NP_NLS2
M2_LIG_LIR_GEN
M2_N-Glyc
M1_NLS
NS1_NES
M1_NES
NS1_PDZ
NS1_PKC_42
NS1_PKA_195

3.5.3.2. Determining codon use at Ser/Thr sites and stop codons

To examine the relative codon usage at phosphorylation [ST] and glycosylation [ST] I manually assigned the positions known to be functional motif sites.

For phosphorylation: I assigned the positions in the proteins NS1, NS2, M1, M2, NP and PA respectively. For each protein I used the relevant alignments determined above. I then counted the occurrence of each codon within Ser/Thr codon space. I summed all non-Ser/Thr codons as a separate "other" category to be able to determine overall Ser/Thr vs non-Ser/Thr use. I plotted the relative codon usage as circle areas proportional to relative codon usage. To determine the total codon bias over all Ser/Thr sites at phosphorylation motifs I summed both the total counts of codons for each position of each type, and also calculated a weighted codon usage where all sites were weighted by dividing the codons by the total Ser/Thr codons at the site. This way the codon bias at each site is shown and comparable, without sites with higher conservation and therefor higher codon count overall biasing the outcome. Since these were functionally important and thus conserved sites however, the outcomes in bias are virtually identical. I also determined the background codon usage across Ser/Thr sites in general in all of these alignments, excluding the known phosphorylated positions. This is assumed to be reflective of the underlying codon biases in influenza in general, and thus important to compare to identify what expected codon use normally would be. To determine if the codon usage at phosphorylated sites was statistically significant I determined contingency tables for the different residue locations and the phosphorylation state. I used three levels of independence with the strain data to test more and less realistic cases. Complete independence assumed each strain was a unique data point and

the codons used at each residue in each strain were recorded. Partial independence (most biologically realistic) assumed that inheritance of codons would contribute to codon usage in strains, but that codon choices that persist across a large number of strains contain additional datapoints. In this scenario sites that are highly conserved across many strains were counted once for every 1000 related strains they were present in. In the third scenario complete dependence in strains was assumed, and each site was only counted once, and the majority codon usage was assigned to the site. This disregards any information from conservation among strains, and does not weigh the data towards highly conserved codons in sites. The significance level was then tested using chi-square.
For glycosylation sites: The same process as for phosphorylation sites was used, with the difference that since glycosylation sites are well defined, I first filtered all the sequences in the alignment by those containing the full motif. Thus, I only looked at codon usage at sites with the full motif. The same statistical approach as for phosphorylation sites was used.

For stop codons the same method was used: For all the alignments I determined which codon was used in the stop positions and plotted the codon usage proportional to circle areas for each protein.

3.5.3.3. Determining overall codon bias in conserved motifs

To determine if the codon usage overall in these sites is biased towards codons that are more robust, I first determined all strains with the full motif in each relevant protein. I then determined the codon usage distribution at each position within the motifs. For example, for a [RK][ST] hypothetical motif, I determined the codon usage distribution within the motif space, e.g. AAA: 0.2, AAG: 0.1, AGG: 0.4, CGC: 0.3 codon usage across the RK position in all strain and similarly for the ST position. I then calculated the weighted loss probability given the codons used. That is, the probability for AGG to mutate into non-RK contributed 0.4 of the total, and the probability of AAA to mutate into non-RK contributed 0.2 etc. I also determined the motif loss probability if each position used 100% of the codon with the lowest loss-probability or if it only used the codon with the highest loss probability. Finally, I determined what the weighted motif probability would be if the overall codon distribution seen across the relevant amino acids in all influenza proteins was used.

To compare the observed probability and the expected probability I calculated the natural log (ln) of the expected loss probability/observed loss probability. This yields a positive value if the observed codon choices are less likely to mutate and lose the motif (lower loss probability than observed) and a negative value if the observed loss probability is higher than the expected. Thus a positive value indicates robust motifs and a negative value indicates non-robust. I also compared the absolute robustness of each motif independent of the codon use in influenza overall. This score was calculated as ln(Observed-Lowest/Highest-Observed), which results in a positive score for more overall robust codon use and a negative score for more overall non-robust codon use.

Chapter 4

Predictability of motif evolution in influenza

4.1. Overview

So far I have looked at the relationship between codon choice and motif loss in influenza. However, one of the interesting observations from chapter 2 was that the gain of new motifs also depended on codon choice in the simulations. Predicting evolutionary outcomes is generally almost impossible because it requires knowledge of all the selection pressures and fitness effects acting on a sequence. In motifs the sequence complexity is reduced since they are part of linear sequence. We also have a better understanding of the functional aspects and limitations of motif sequences based on the motif consensus. This allows us to better understand the possible outcomes of mutations in sequences as I have established in this thesis. In this chapter I explore how and when viruses gain new motifs. I investigate how motifs are sampled in sequence through mutations, thereby generating the underlying functional innovation that selection can act on. I then look specifically at instances of influenza history where motif evolution has been studied over time. I find that motif probability can aid in the prediction of new glycosylation site evolution. Phosphorylation motifs are also sampled and shaped by the underlying probability of motif evolution, suggesting these methods can inform the possible evolutionary outcomes of motif patterns in general. The functional consequences of motif evolution in influenza are important and the increased understanding of the evolutionary dynamics of these motifs are discussed in the final section.

4.2. Introduction

The observations made in chapter 3 strongly indicate that motif codon space and mutation rates impact mutational outcomes and by extension the fitness of the virus. These findings were predicted by the simulation outcomes showing that there could be a potentially significant fitness effect from differences in codon choice in the motif space, through the variable loss of key functions. The other observation from the simulations was that non-motif containing sequences evolve new motifs at drastically

different rates depending on prior codon choice in a sequence as well. Evolution of new motifs in influenza could thus similarly be influenced by codon choice indicating that evolution of motifs could be predictable.

4.2.1. Predicting evolutionary outcomes

Fundamentally any given mutation is random and cannot be predicted, however due to mutational biases, sequence landscapes and similar selection pressures, convergent outcomes are often observed in evolution (Larter *et al.* 2018; Stoltzfus & McCandlish 2017). Convergent evolution can happen on relatively short time scales. Surprisingly frequently, similar solutions evolve in many unrelated pathogens through convergence. A common example of this is the evolution of interactions with PI3Kinase by a large number of pathogens (Diehl & Schaal 2013).

The dialogue around evolutionary prediction in the field has mostly been focused around the large scale population shifts and not on changes and biases on the sequence level (Lässig *et al.* 2017). Cancer evolution is an example where prediction has been discussed extensively, and while there are common targets and pathways that are involved in many cancers, predicting which change or mutation will drive it is almost impossible (Lässig *et al.* 2017; Lipinski *et al.* 2016). A contributing factor is likely to be our limited understanding of the determinants of mutation rates and of the effects of mutations at the vast amounts of sites in these systems. However, the mutational landscape of sequences is important for future evolutionary outcomes. A classic example of this is from the long term evolutionary *E. coli* experiment conducted by Richard Lenski and colleagues (Good *et al.* 2017). Here, one of the populations evolved the ability to metabolise citrate after more than 30000 generations (Blount *et al.* 2008). They showed that an exaptation, or a non-adaptive mutation, primed the lineage for evolution of the citrate metabolising ability. Generations of populations following the "priming" mutation readily also evolved to metabolise citrate, whereas other linages or prior generations could not easily evolve this function under the experimental conditions.

It is only very recently that an interest in mutational biases have begun to be explored in adaptive evolution, since previously it had been argued that any biases would be negligible compared to natural selection (Stoltzfus & McCandlish 2017; Yampolsky & Stoltzfus 2001). To what extent evolution can be predicted, and how different aspects of mutation rate bias shape sequence evolution is still a fundamental question in evolutionary biology.

4.2.2. Importance of motif emergence for functional innovation

Motifs are a resource for pathogens, enabling rapid evolutionary innovation and adaptation in host cells (Chemes *et al.* 2015; Davey *et al.* 2011; Hagai *et al.* 2014). They could be argued to occupy a unique sequence niche, being some of the smallest functional units in the cell with a disproportionate impact on complex functionality compared to their size in protein space (Van Roey *et al.* 2014).

Changes to single phosphorylation sites and glycosylation sites in influenza can drastically impact how the virus life cycle is timed and how it affects the host (Mondal *et al.* 2017; Vigerust & Shepherd 2007; Zheng *et al.* 2015). Mutations to localisation signals can change the entire active environment of the protein very rapidly (Li *et al.* 2015; Tynell *et al.* 2014). Importantly, motifs also enable viruses and viral proteins to evolve similar functionality convergently through the interaction with conserved domains used in host signalling systems, which can be achieved through e.g. PDZ motifs (Chemes *et al.* 2015). Having better models for how, when and where different motifs and functions evolve in response to certain pressures would have implications for disease modelling and our ability to tackle the disease.

4.2.3. Chapter summary

In this chapter I establish if the predictions from the simulations in chapter 2 are reflected in the evolutionary record in influenza. That is, if putative and real motif emergence can be inferred from prior sequence based on mutation rate bias and the overall underlying probability through the codon space for the motif. I then analyse and predict the sites where motifs were likely to evolve in historical strains of influenza and compare to the subsequent evolution observed in the record, in particular for glycosylation sites in influenza coat proteins. I also determine the future potential motif landscape based on currently circulating strains. By considering functional and structural context alongside motif evolutionary potential I can then also classify the motif hotspots based on probability of functional impact and likelihood of the motif gain event. For both phosphorylation sites and glycosylation sites I show that these regions of potential motif innovation are important considerations for future evolutionary trajectories. Having a hotspot map is an early step towards better understanding future evolutionary dynamics of motifs and how they might react on short evolutionary scales to drastic changes in environment and pressures. This change in selective pressures is particularly relevant when thinking about targeted treatments such as vaccines. Vaccines by design exert a large selection pressure on viral sequences that may cause a shift in the prevailing strain and phenotype in response. A prior idea of the likely evolutionary trajectory of a strain could facilitate better vaccine design, through a combination of better informed strain choices that are less susceptible and careful selection of the protein targets of the vaccines ultimately put into production.

4.3. Results

4.3.1. Mutation induced sampling of new motifs in influenza A proteins

Mutations are a key source of functional innovation in influenza (Shao *et al.* 2017). Point mutations will occasionally make new motifs emerge in influenza sequences. These new motifs can immediately

have a functional impact and change regulation within a protein, or they can be a latent source of functional innovation for the virus (Fortuna *et al.* 2017; Petri *et al.* 1982). Sequences with a motif that has emerged in the wrong functional context can thus be primed for a new environment, which can be of importance for adaptation in viruses such as influenza where host shifts are common. This is known as an exaptation, and is a common mechanism of evolutionary innovation in biology (Whittington *et al.* 2018). The probability of sampling a new motif should be determined by the prior sequence, the codon space and the mutation rate, however other restrictions on structure and function will limit this potential evolvability. A more in depth look at when and where mutations lead to motif emergence would be of value to improve our understanding of motif evolvability and innovation potential.

To gain a better understanding of these evolutionary events I analysed the evolutionary history of influenza A nucleoprotein (NP) to determine how frequently different motifs are sampled (Figure 4.1). To do this I have used the pipeline developed in chapter 2 to identify motif pattern matches. Since phosphorylation sites are the most well characterised motifs in influenza proteins, and in particular NP, I have restricted my analysis specifically to phosphorylation motifs (Hutchinson *et al.* 2012; Mondal *et al.* 2015). I look specifically at the evolutionary emergence of the recognition motifs by the kinases CK1 and CK2, PKA and GSK3. These are common eukaryotic kinases that phosphorylate influenza proteins (König *et al.* 2009; Meineke & Rimmelzwaan 2019; Sierra *et al.* 1998). They are ubiquitous kinases that are expressed constitutively in human cells. They also all use short recognition motifs, which means influenza can sample putative motifs easily through few mutations.

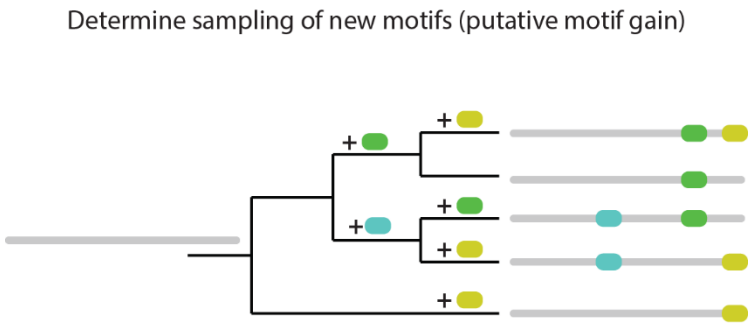


Figure 4.1. Sampling of motifs over evolutionary time. Mutations will stochastically give rise to new motif consensus sequences that can give rise to new functions in sequences. By identifying novel motif patterns over a viral phylogeny I determine when motifs have been sampled.

In this analysis I have determined all the instances of new motif emergence of these kinase sites in NP. Since the viral record contains ancestral sequence states I was able to determine which mutations occurred to yield new motifs. Thus I have compared the codons, amino acids and nucleotides that are most frequently involved in putative phosphorylation site emergence.

Emergence of new phosphorylation sites is a quick way by which proteins can alter the regulation of proteins and interactions and the timing and localisation of various events (Holt *et al.* 2009). In several influenza A proteins, importantly NS1, NP and M1 phosphorylation has been found to be used to regulate oligomerisation, RNA binding, protein-protein interactions and nuclear import (Hsiang *et al.* 2012; Turrell *et al.* 2015; Zheng *et al.* 2015).

These sites frequently get sampled in strains through point mutations. Whether these sampled motifs end up being phosphorylated is unknown. However, their emergence through point mutations are important data points to analyse the underlying innovation potential influenza has for new phosphorylation sites. It will also allow me to determine the sequence biases involved in sampling new motifs. Overall, these four classes of phosphorylation motifs have been sampled 285 times through independent mutational events at 40 different sites in the NP protein. Some sites were sampled only in a single strain in the dataset used for this analysis (9592 strains ranging from 1902-2012) and the site that emerged the most times independently evolved in 42 independent lineages. The ancestral sequence prior to the motifs evolving were all very high probability sites for getting the specific mutations that would lead to the kinase recognition motif. Since they were recent ancestors to these strains that is to be expected, but it still highlights that the information gained from a nucleotide and codon centric sequence level can inform regions were motifs have a higher chance of emerging.

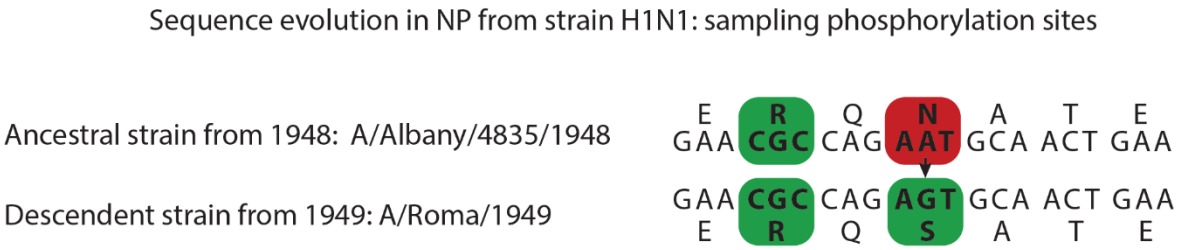


Figure 4.2. Example of a sequence gaining a new PKA motif. The ancestor in the 1948 sequence lacked an Ser/Thr site. In a strain in 1949 an A-to-G mutation (high mutation rate) substituted asparagine to serine yielding a new PKA motif.

For PKA evolution with consensus xRx[ST][^P]xx, the majority of sites emerge as a result of a mutation yielding a serine or threonine rather than a gain of arginine (Table 4.1, yellow highlight). This is likely a result of more total Ser/Thr codons (10) than arginine (6), as well as the mutation rates and mutational space favouring gain of Ser/Thr. The amino acids and codons that evolve into new Ser/Thr sites are predominantly asparagine and isoleucine codons with a single instance of methionine in this dataset (Table 4.1). All these codons are among the top predicted to spontaneously mutate into codons within serine or threonine which is reflected in the probability scores for the ancestral sequences. The nucleotide mutations that resulted in the motif gain are biased to A-to-G and T-to-C mutations, which are the two most frequent substitutions, both as determined by the history in the phylogenetic tree as

well as by results from experimental mutation rate determination (Table 4.1, orange highlight). They are both estimated to be around $2*10^{-4}$ mutations per replication. An illustration of these typical changes from a 1948 ancestral sequence, where asparagine (AAT) mutated into serine (AGT) through an A-to-G substitution, highlights how the sequence history of viruses can be used to infer these evolutionary events (Figure 4.2).

Table 4.1. Sampling of PKA sites in influenza A NP. Rows in yellow highlight motif gain through a new serine or threonine. Orange highlights when the mutations giving rise to a motif is one of the mutation classes in influenza with high mutation rates. Blue highlights high probability motif emergence in the ancestral sequence (threshold 0.015).

Position	Codon Change	AA change	Nucleotide change	Independent Strains	Probability
46	[AAA->AGA]	[K->R]	[A->G]	9	0.069
88	[AAG->AGG]	[K->R]	[A->G]	6	0.077
17	[AAT->AGT, AAT->ACT]	[N->S, N->T]	[A->G, A->C]	9	0.082
193	[ATC->ACC]	[I->T]	[T->C]	1	0.072
197	[ATC->ACC]	[I->T]	[T->C]	4	0.077
97	[GGC->AGC]	[G->S]	[G->A]	1	0.059
53	[ATC->ACC, ATT->ACT]	[I->T, I->T]	[T->C]	3	0.070
115	[ATT->ACT]	[I->T]	[T->C]	3	0.052
5	[AAA->AGA]	[K->R]	[A->G]	2	0.061
119	[GCG->TCG, GCT->TCT]	[A->S, A->S]	[G->T]	2	0.066
96	[ATA->ACA]	[I->T]	[T->C]	2	0.063
101	[AAG->AGG, ATG->ACG]	[K->R, M->T]	[A->G, T->C]	2	0.050
2	[CAA->CGA]	[Q->R]	[A->G]	2	0.074
72	[AAT->ACT]	[N->T]	[A->C]	1	0.083
219	[TGC->TCC]	[C->S]	[G->C]	1	0.016
148	[AGA->ACA]	[R->T]	[G->C]	1	0.029
137	[TGG->AGG]	[W->R]	[T->A]	1	0.085

For GSK3 with consensus xxx[ST]xxx[ST] I would expect these to follow similar trends to the [ST] site of PKA. Probabilities in the ancestral sequences are very high overall, as also seen in PKA. The most frequent substitutions to Ser/Thr codons at either of the key positions are isoleucine, methionine, asparagine and arginine AGA and AGG (expected to be higher than the remaining arginine codons) (Table 4.2). On the nucleotide level the expected high frequency mutations are again overrepresented with A-to-G and T-to-C but also several G mutations which have a measured mutation rate around $5*10^{-5}$ and are the second highest overall after the previously mentioned two mutations (Table 2, orange highlight). In fact, of all the substitutions in this dataset to mutate into new phosphorylation sites, >87% are either G-to-any or T-to-C or A-to-G which are only half of all possible substitutions. This highlights the impact of mutation rates in predicting high probability sites for sampling new motifs.

An additional interesting observation is that at several sites (e.g. 88, 122) several independent lineages have evolved the motif through the same high probability substitution.

Table 4.2. Sampling of GSK3 sites in influenza NP. Orange highlights when the mutations giving rise to a motif is one of the mutation classes in influenza with high mutation rates. Blue highlights high probability motif emergence in the ancestral sequence (threshold 0.015).

Position	Codon Change	AA change	Nucleotide change	Independent Strains	Probability
19	[GCA->TCA, GCA->ACA, GCT->TCT]	[A->S, A->T, A->S]	[G->T, G->A]	13	0.086
207	[AAT->AGT, AAC->ACC]	[N->S, N->T]	[A->G, A->C]	11	0.089
122	[GGT->AGT]	[G->S]	[G->A]	4	0.083
88	[ATT->ACT, ATC->ACC, ATA->ACA]	[I->T, I->T, I->T]	[T->C]	4	0.076
46	[GGC->AGC]	[G->S]	[G->A]	2	0.074
37	[ATA->ACA]	[I->T]	[T->C]	4	0.062
101	[ATG->ACG]	[M->T]	[T->C]	1	0.053
167	[AGA->ACA, AGG->ACG]	[R->T, R->T]	[G->C]	2	0.029
163	[ATG->ACG]	[M->T]	[T->C]	2	0.061
11	[CGC->AGC]	[R->S]	[C->A]	2	0.015
172	[GGT->AGT]	[G->S]	[G->A]	1	0.085
166	[AGG->ACG]	[R->T]	[G->C]	1	0.018
162	[TTG->TCG]	[L->S]	[T->C]	1	0.068
153	[CCC->ACC]	[P->T]	[C->A]	1	0.091
61	[AGA->AGT]	[R->S]	[A->T]	1	0.035
1	[GGC->AGC]	[G->S]	[G->A]	1	0.067

Finally, for CK1 and CK2 recognitions sites with consensus either [ED]xx[ST]xxx or xxx[ST]xx[ED] there is also a preference for the motif to emerge through gain of [ST] rather than [ED]. Again, that is likely due to the larger codon space for Ser/Thr but also due to the mutation rates for the relevant codons and nucleotides, and this is consistent with the calculated probabilities and expectations for this motif. Asparagine and glycine are the two predominant substitutions into aspartate or glutamate in this dataset. These residues are high probability residues, but these substitutions are also very likely helped by the similarity in biophysical properties of the amino acids, which creates synergy between motif sampling at these residues. Interestingly, glutamine (Q) frequently substitutes to glutamate (E) despite being very low probability due to the C-to-G mutation required. This can be understood through the similarity in biophysical properties, but is one of the only observed instances of low probability sites evolving the motif across these three kinases (Table 4.3, grey highlight).

Overall there appears to be a large innovation potential for phosphorylation sites in influenza based on the data from this NP analysis alone. Sequence space and mutation rates result in regions where motifs can emerge with high probability providing an opportunity for functional innovation in influenza.

Table 4.3. Sampling of CK1 and CK2 sites in influenza A NP. Rows in yellow highlight motif gain through a new serine or threonine. Orange highlights when the mutations giving rise to a motif is one of the mutation classes in influenza with high mutation rates. Blue highlights high probability motif emergence in the ancestral sequence (threshold 0.015).

Position	Codon Change	AA change	Nucleotide change	Independent Strains	Probability
213	[ATT->ACT, ATT->AGT, GGT->AGT, ATC->ACC, AAT->AGT, ATT->ACA]	[I->T, I->S, G->S, I->T, N->S, I->T]	[T->C, T->G, G->A, A->G, T->A]	42	0.086
23	[GCA->TCA, GCA->ACA, GCT->TCT]	[A->S, A->T, A->S]	[G->T, G->A]	13	0.092
81	[GCG->ACG, GCA->ACA, GCC->ACC, GCT->ACT, GCT->TCT, GCG->TCG]	[A->T, A->T, A->T, A->T, A->S, A->S]	[G->A, G->T]	45	0.090
105	[ATC->ACC, ATT->ACT, ATT->AGT, ATC->ACT]	[I->T, I->T, I->S, I->T]	[T->C, T->G, C->T]	37	0.072
121	[AAT->AGT, AAT->ACT]	[N->S, N->T]	[A->G, A->C]	17	0.081
17	[AAT->AGT, AAT->ACT]	[N->S, N->T]	[A->G, A->C]	9	0.080
50	[GGC->AGC]	[G->S]	[G->A]	2	0.077
113	[AGG->AGT]	[R->S]	[G->T]	1	0.014
120	[AAC->AGC, AAT->AGT]	[N->S, N->S]	[A->G]	2	0.080
41	[AAA->GAA, CAG->GAT]	[K->E, Q->D]	[A->G, C->G, G->T]	2	0.066
72	[AAT->ACT]	[N->T]	[A->C]	1	0.085
184	[GGA->GAA]	[G->E]	[G->A]	1	0.084
5	[CAG->GAG]	[Q->E]	[C->G]	1	0.004
219	[TGC->TCC]	[C->S]	[G->C]	1	0.016
137	[AAT->GAT]	[N->D]	[A->G]	1	0.062
211	[GGA->GAA]	[G->E]	[G->A]	1	0.074
138	[AAT->AGT]	[N->S]	[A->G]	1	0.069
11	[CAG->GAG]	[Q->E]	[C->G]	1	0.004
39	[ATG->ACG]	[M->T]	[T->C]	1	0.058
71	[AGA->ACA]	[R->T]	[G->C]	1	0.030
79	[CCC->TCC]	[P->S]	[C->T]	1	0.086
143	[AAT->GAT]	[N->D]	[A->G]	1	0.066
147	[TAT->GAT]	[Y->D]	[T->G]	1	0.007

It is more than likely the case that many sampled motifs are either detrimental to fitness when phosphorylated or inaccessible for phosphorylation. However, characterising the potential and frequency for these sequences where kinase recognition sites and other motifs emerge can give us insights that help better predict and understand the evolution of regulation and new functionality. It is likely the case that most if not all phosphorylation sites currently used in these proteins emerged by this same mechanism of random point mutations. Changing the phosphorylation and motif landscape is an important mechanism by which the virus adapts to changing conditions, new hosts and also treatments. Knowing the underlying potential for change in various directions in evolutionary space can provide us with deeper insight into how to better treat viral infections and make it harder for viruses to evolve around treatments.

4.3.2. Motif emergence correlates with motif probability and suggests predictability of evolutionary trajectories

In the previous section I looked at how some phosphorylation sites are sampled over influenza strains in history, and it is clear that many putative motifs emerge spontaneously through point mutations. These putative motifs predominantly evolved at sites that were high probability, "primed", sites meaning that one or several nucleotide point mutations that have high mutation rates would lead to the emergence of the motif. In chapter 2 I also simulated the evolutionary dynamics of putative motif sequence patterns that are more or less primed to evolve given their codon choices. I found that the frequency of motif sampling over the different strains in a large phylogenetic tree, and indeed the population size of viruses with the new motif during individual infections all would be expected to correlate closely with the prior motif probability such that higher probability leads to a higher rate of motif emergence. The next question I wanted to ask was therefore if there is an overall correlation between the theoretical motif probability at individual sites in influenza proteins, and the emergence rate of putative motifs at these sites in influenza history. The implications of a correlation between a theoretical evolution probability and real evolutionary sampling would be of interest within the scope of the predictability of emergence for motifs at specific sequence locations. If the likelihood of putative motif evolution can be determined based on prior sequence, these tools could be used to evaluate the adaptations of viruses to changing conditions, and in particular to selection pressures imposed by human anti-viral efforts.

For this analysis I used the influenza protein NS1 as it is highly relevant from a motif perspective, having the most diverse set of known motif classes, including transportation, modification and binding sites (summarised in section 3.3.2). The structure of the protein also gives it a high potential for motif evolution and sampling since it has several disordered and exposed sequence regions (Carrillo *et al.* 2014; Hale *et al.* 2008). It is also the main protein involved in subverting the host defence machinery. This dataset consists of 8663 NS1 sequences spanning the years 1902-2012 from a range of strains and

backgrounds including H1N1, H3N2, H5N1 and H7N7. From the strain dataset I identified a range of putative motifs that have been gained and lost in a proportion of the strains. For all the strains without the motif I determined the average probability at the pre-motif site for that motif evolving. This average probability was used as the general probability for evolution for each motif. I then tracked mutations in the phylogeny to determine all instances of where a mutation between an ancestor and descendant gave rise to a new motif pattern. For each motif I determined the gain events that gave rise to the motif-sequence at the selected sites.

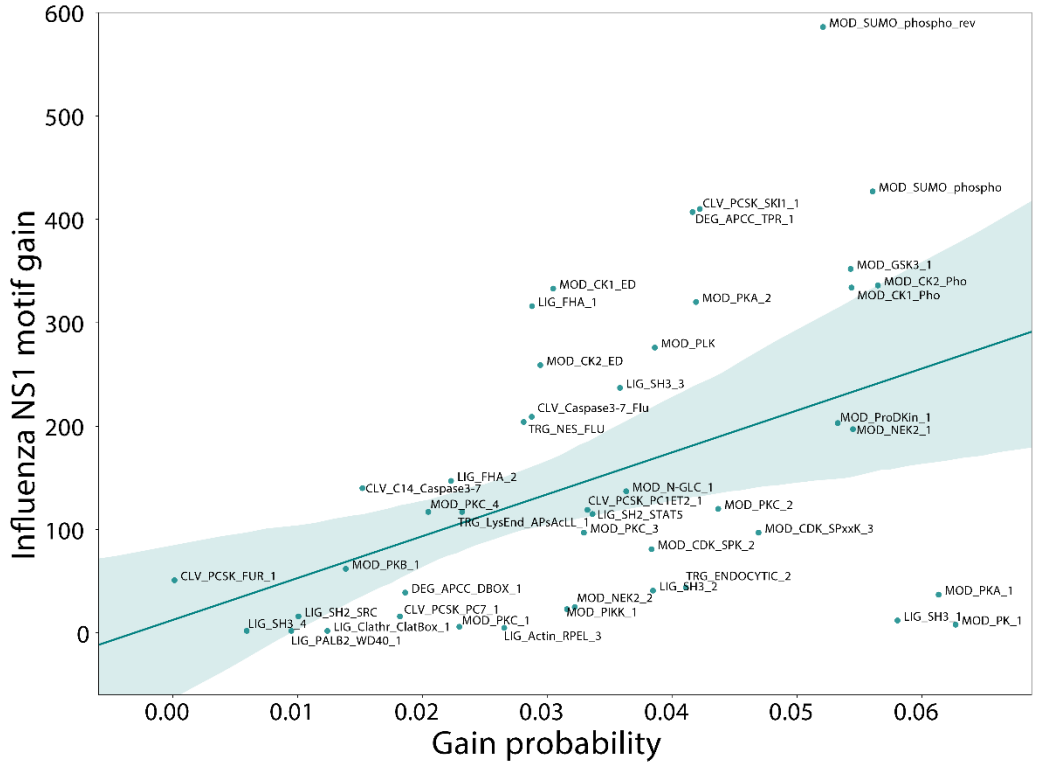


Figure 4.3. Motif gain at specific sites in NS1 is correlated with prior sequence gain probability. Overall there is a trend for functional and putative motifs to emerge more frequently at sites with high motif emergence probability in ancestral strains lacking the motif, indicating that a high emergence probability at a site informs evolvability. The blue line indicates best fit through least squares, p=0.002.

Overall, there is a significant correlation between the average gain probability and the gain events across the sampled putative motifs and sites in NS1 (Figure 4.3). Motifs that have a higher gain probability at ancestral sites evolve more often in the historical sequence record. The variation is quite large and the data relatively noisy which is not surprising given the nature of viral sequences with generally high selection pressures to conserve residues in many regions of the protein, as most mutations have been shown to be detrimental (Lyons & Lauring 2018). Thus many high probability putative motifs are likely to be near other conserved features leading them to never emerge in NS1 history thus far.

The motifs sampled in this dataset include several different kinase recognition sites, degrons, SUMOylation sites and several sites recognised by scaffolding domains such as WW, SH3 and FHA. These motifs are putative motif evolution events, and as such they are simply pattern matches, and most are thus likely not to carry any current functionality for the protein. The fact that these patterns evolve through mutations in a way that correlates to the probability of that pattern at the prior site carries important implications for the evolution of new functional motifs nonetheless. As motifs are simply recognisable patterns carrying no function prior to them emerging spontaneously, the evolution of new functional motifs should follow the same correlation. Alongside other methods such as structurally informed modelling and experimental binding assays, the prior sequence and probability thus carries predictive qualities for the emergence of new motifs in general. Two other important things to note are that this is in line with the predictions in chapter 2, suggesting the simulations are reasonable models for sequence evolution in nature, and the most crucial fact which is that the purely theoretical probability that relies simply on the nucleotide mutation rate and shape of the codon space can inform outcome probability, which can be useful when only limited data exists.

4.3.3. Investigating HA glycosylation evolution and predicting the future glycosylation landscape

Glycosylation of the coat protein HA is essential for both normal function and folding of the protein as well as in regulating the way the protein interacts with and infects host cells (Vigerust & Shepherd 2007). Importantly, it is also involved in altering the viral coat surface accessibility, which changes the way in which the immune system can deal with the viral infection. Antigenic sites can be blocked by glycosylation modifications, thereby preventing antibodies from binding to the virus (see Figure 4.4)(Peng *et al.* 2019). In the majority of human influenza strains the pattern of glycosylation shifts over time. The most common seasonal strains (H1N1 and H3N2) had very few glycosylation sites in the globular head domain when they were first introduced in the human population (H1N1 in 1918 and 2009 and H3N2 in 1968) (Altman *et al.* 2019). However, over time the number of glycosylation sites increased through mutation and subsequent fixation, likely as an adaptation to increased immunity in the human population (Tate *et al.* 2014). Knowing where glycosylation sites are likely to evolve would be important in our understanding of influenza adaptation to humans as a host. It could potentially allow us to better predict and tackle the influenza response to vaccinations as we could anticipate the glycosylation changes that are likely to shape the virus and thus choose strains with low potential for modifications or changes that would nullify the effect of vaccination. It has also been found that vaccination efficacy can be affected by changes to glycosylation sites in the cell systems used for vaccine production (Belongia & McLean 2019). Low efficacy vaccination attempts have resulted from glycosylation gain as an adaptation to hen's egg cells used in vaccine production, which compromised the vaccination in humans (Zost *et al.* 2017).

To investigate the evolutionary probabilities and outcomes of glycosylation sites I have looked at the history of both H1N1 and H3N2, the predominant seasonal influenza strains. These strains have the most well characterised glycosylation sites as they have been studied in detail through the 20th and 21st centuries. I have characterised the glycosylation site evolvability across ancestral strains and compared that with the subsequent known glycosylation site evolutionary events. I then also looked at the most recent strains to determine where the likely glycosylation hot-spots are for the future of these seasonal strains.

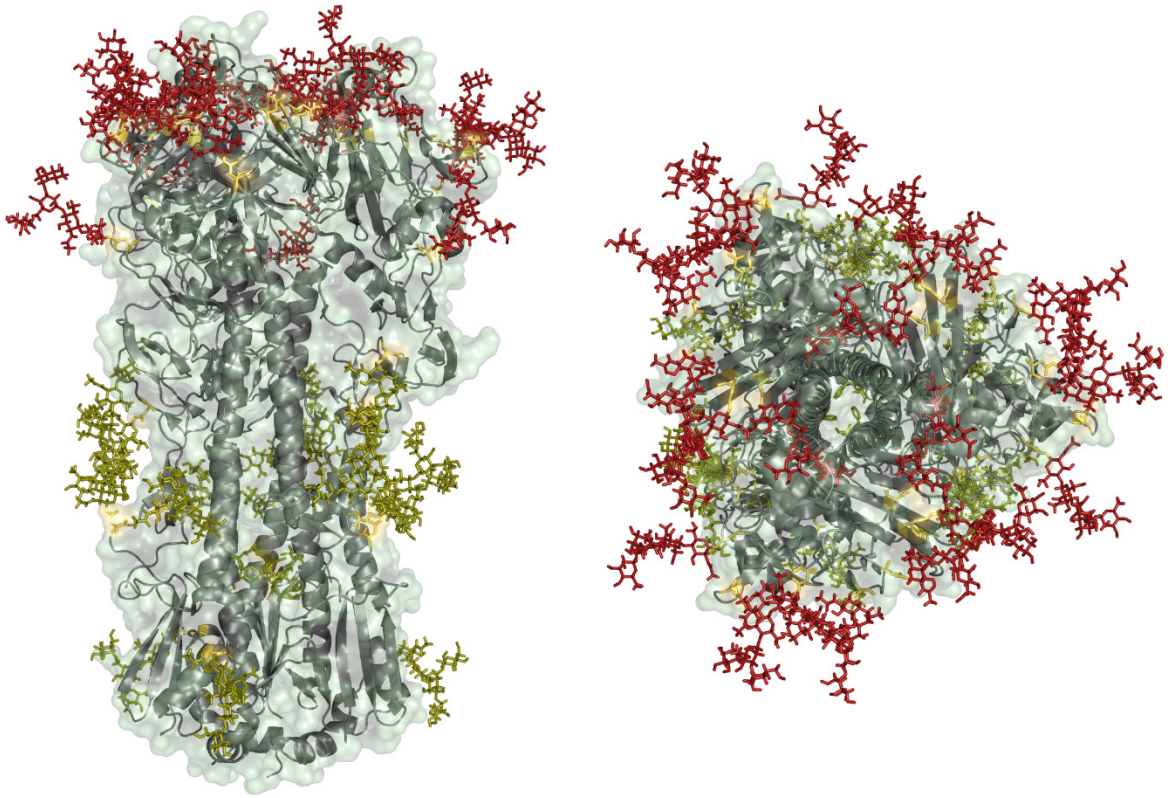


Figure 4.4. Glycosylated H3 trimer. In red are glycans attached to the globular head domain. In gold are glycans in the stem domain. Left shows a side view, highlighting the clustering of glycans around specific regions and the difference between the stalk and head domains. Right shows a top view to illustrate the abundance of glycans on the head domain. The figure was made from PDB 4O5N showcasing H3 from H3N2 strain A/Victoria/361/2011. Glycans were added using GlyProt (Bohne-Lang & von der Lieth 2005).

4.3.3.1. Evolution of glycosylation sites in H3N2

H3N2 emerged in humans during the pandemic in 1968 (Kilbourne 2006). It has since been one of the predominant seasonal flu strains and has gained and lost glycosylation sites across the globular head domain (Altman *et al.* 2019; Alymova *et al.* 2016). When it first entered the human population in 1968 it only had two predominant N-linked glycosylations in the globular head, at residues N81 and N165 (numbering based on the mature protein without the N-terminal signal sequence). In the time since, several new glycosylation sites have been sampled and fixed (and sometimes lost again) in different

strains over the world and the currently circulating strain has ~7 known glycosylation sites (Figure 4.5).

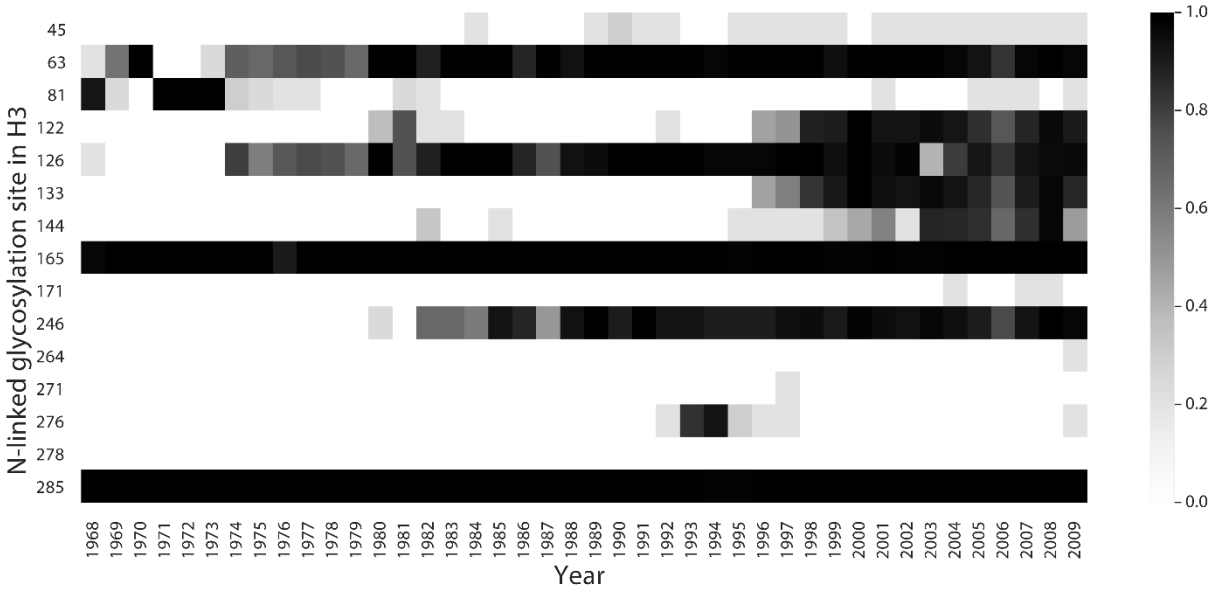


Figure 4.5. Timeline of glycosylated positions in H3 strains. The heat-map intensity displays the relative number of strains with a glycosylation site at the indicated position within the dataset in a given year. If all sequences in a given year have a specific site, it is thus black. This highlights that some sites such as 165 and 285 are absolutely conserved in H3, whereas other sites emerge and disappear at different time points and are more or less widespread. Some only appear in a few strains over a short time period.

To determine the evolutionary potential in glycosylation sites in the early strains, after introduction into the human population and the following pandemic, I used the pipeline as outlined in chapter 2 to scan through the nucleotide sequence and calculate the probability for each site to mutate into N[^P][ST] (Figure 4.6). Each residue was also filtered corresponding to the relative solvent accessibility in the protein, using the EBI PISA resource for precalulated accessibility values for the PDB structure 1ha0 (Krissinel & Henrick 2007). Glycosylation on the globular head is predominantly on surface exposed asparagines. I also filtered out the residues that make up the stalk domain. In addition, residues corresponding to the highly conserved receptor binding pocket were also excluded. After getting the probabilities I used a lower end cut-off to only keep the high and medium probability sites. The calculated and filtered probabilities were then plotted over each sequence position (Figure 4.7). The glycosylation state of the known sites (numbered in Figure 4.7 and Figure 4.8) have been confirmed through several experimental observations in some strains, and subsequently inferred in other closely related strains. The most striking observation is that the top seven probability predictions all have evolved at different points in H3N2 strains through mutational sampling (Table 4.4). Four of those 7

motif sites also became fixed in the viral population at different time points and are still functionally important glycosylation sites in circulating influenza strains.

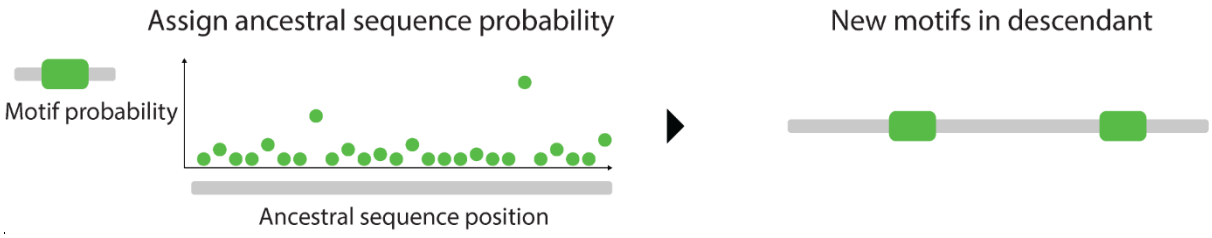


Figure 4.6. Predicting motif evolvability. Using the motif probability as described each position in a sequence can be assigned a probability of motif emergence through mutation. Comparing those values to the evolutionary record can then inform how much motif evolvability determines evolutionary outcomes.

The sites closer to the lower end cut-off point at around 0.02 are not sampled as frequently and few of those sites have evolved as favoured glycosylation sites. An exception to this is N126, which is in fact one of the key early glycosylation sites to be fixed after the outbreak of the 1968 influenza. N126 requires only a single point mutation (Codon ACT to AAT) however it is a rare mutation making the probability lower than even some double mutations (e.g. SKA to NKS/T at position 91, (Table 4.4)).

Table 4.4. Top predictions for glycosylation site evolution. Fields in green in "evolved in pop" indicate predicted sites that have gained glycosylation motifs in subsequent strains, and red indicate that the sites have never been gained. Green fields in "fixed in pop" indicate the glycosylation sites that were fixed in H3N2 at some point (i.e. the motif was present in the majority of strains for one or several seasons). Red fields indicate sites where glycosylation motifs were not fixed.

Position	Residues-Pre	Evolved in pop	Fixed in pop	Calculated probability
171	NDN			0.086
246	NSN			0.086
53	NNP			0.082
63	DCT			0.079
45	SSS			0.074
122	NEG			0.074
264	KSS			0.028
96	NCY			0.021
104	DYA			0.017
91	SKA			0.017
173	NFD			0.017
101	DVP			0.016
271	DAP			0.016
126	TWT			0.016

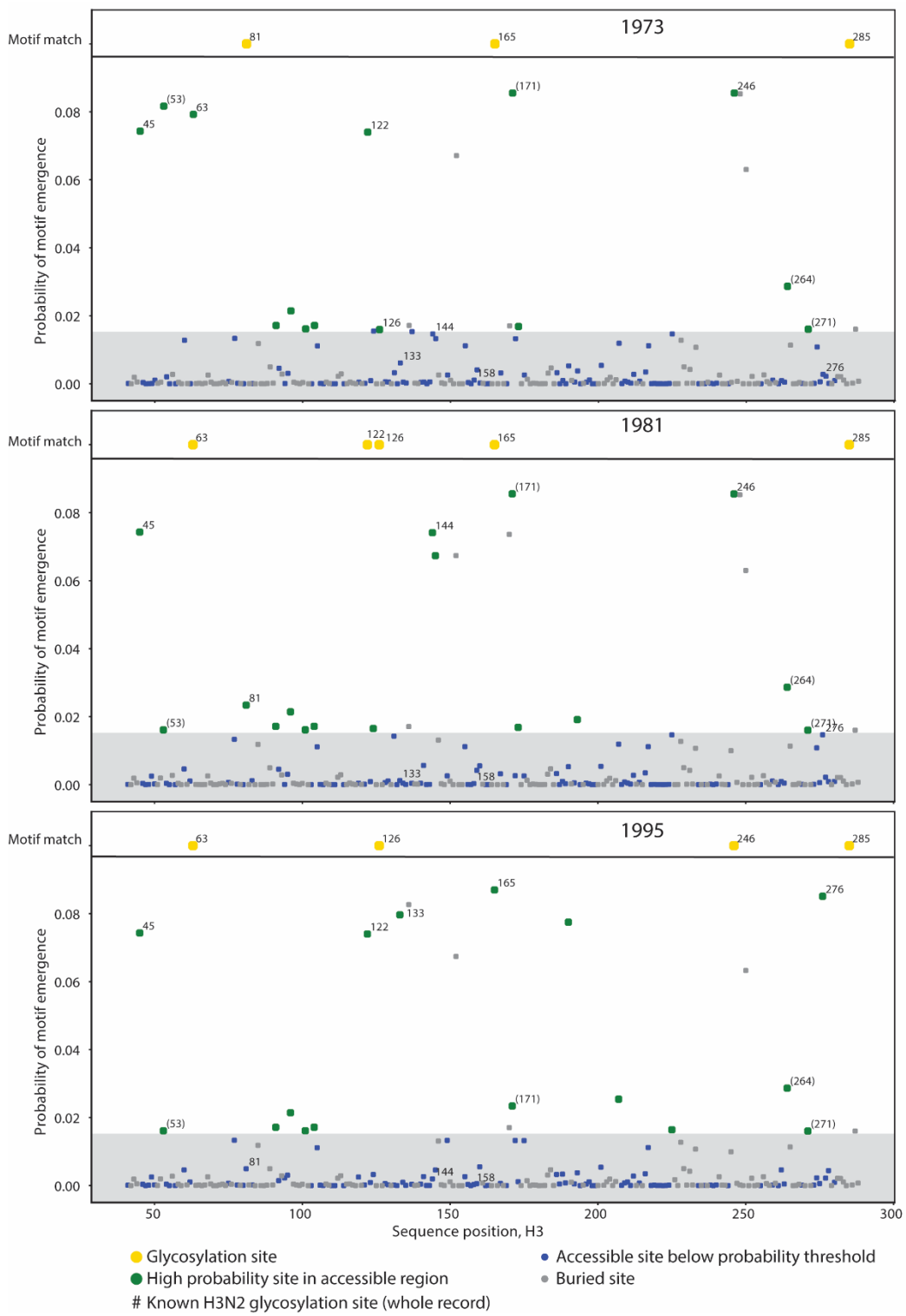


Figure 4.7. Probability of sites across the sequence of H3N2 to evolve new glycosylation motifs in three different years. Numbered sites indicate the key glycosylation sites across all of H3N2 history and numbers in parentheses indicate sites that have been glycosylation consensus sites in some strains, but not fixed. Between 1968 and 1973 things remain unchanged. The top predicted sites (green) include the sites that will become new sites by 1981, N63, N122 and N126. Also future key site 246 and 45 are high probability sites. Between 1981 and 1995 246 evolves into a new glycosylation site. 122 and 165 are also lost, but they remain high probability sites. 276 has also become high probability, and will be fixed in a single season and then lost.

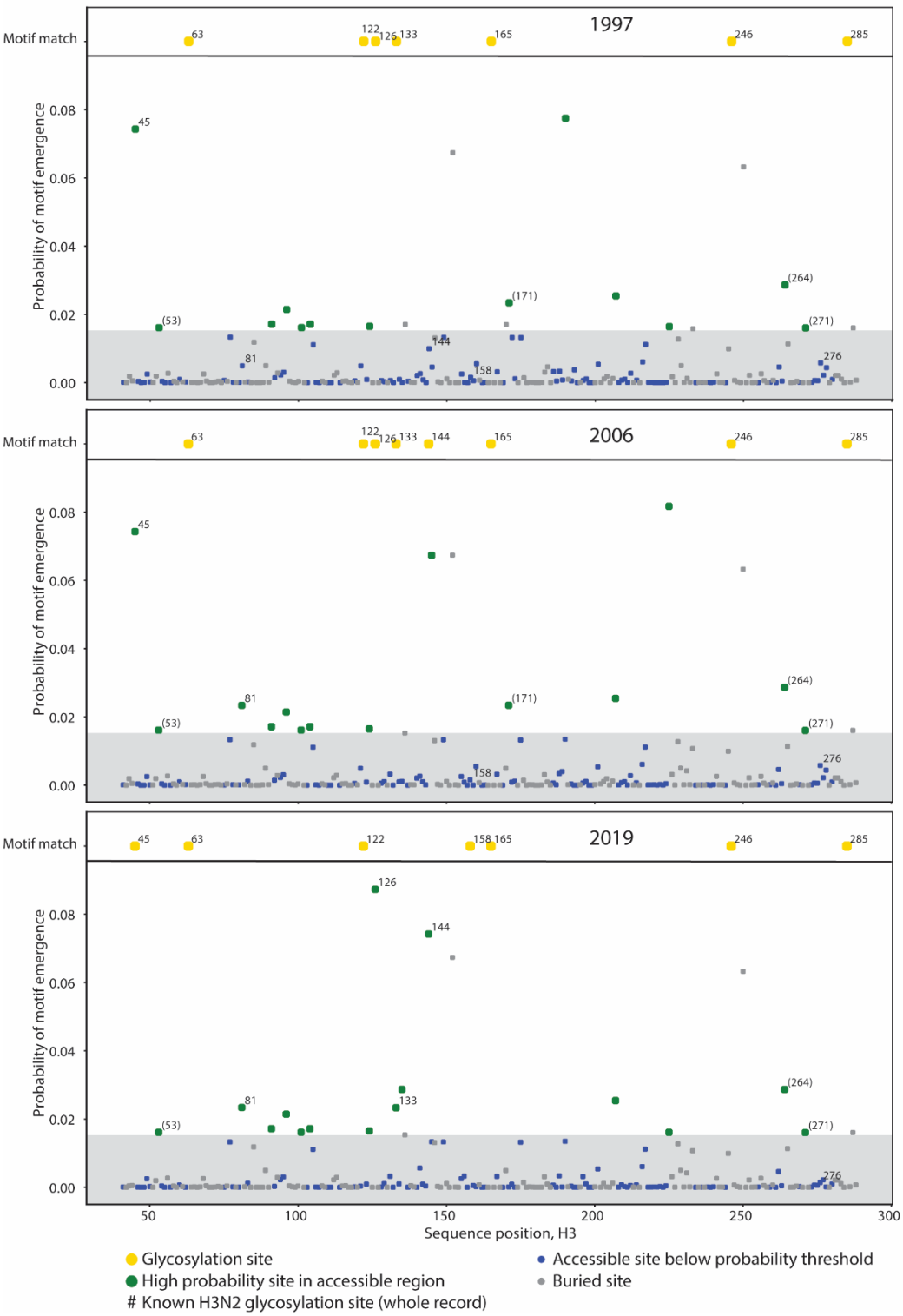


Figure 4.8. Probability of sites across the sequence of H3N2 to evolve new glycosylation motifs in three different years. Numbered sites indicate the key glycosylation sites across all of H3N2 history and numbers in parentheses indicate sites that have been glycosylation consensus sites in some strains, but not fixed. Between 1995 (Figure 4.7) and 1997 sites 122, 133 and 165 all are gained from high probability positions. Few high probability sites apart from 45 remain. By 2004 144 emerges, a site that has been fluctuating in probability across the different years (and was a top predicted site in 1981 (Figure 4.7). Finally, comparing the most recent 2019 strain sequence, 45 has become a glycosylated motif which has been a high probability site throughout. 126, 133 and 144 have been lost but remain high probability. Other 2019 high probability sites include 81, 53, 264 and 271, all of which have been seen emerge in past strains.

Between 1973 and 2019, as illustrated, the key observation is that new glycosylation motifs are predominantly gained from the sites with the highest gain probability as calculated (Figure 4.7 and Figure 4.8). Low probability sites tend to gain mutations that make them higher probability before then evolving into a glycosylation motif (see locations 133, 276 in 1981, 1995 and 1997 respectively) (Figure 4.7 and Figure 4.8). It is particularly striking to compare the predicted sites in 1981 and 1995 with the functional glycosylation sites in 2004 and 2019, where the vast majority of sites that evolved correlate with the set of high probability sites. Summarising these observations by looking at all strains that were circulating between 1968-1985, and their respective probabilities of evolving new glycosylation sites, it is striking to see that all but one of the subsequent glycosylated sites were predicted (Table 4.5). To test whether prior probability significantly correlates with subsequent evolution, I used a contingency table. This resulted in a highly significant correlation with a substantial difference between the probabilities in sites that later evolved, and sites that never evolved, χ^2 (1, N = 42) = 26.25, p < .0000001. Despite the majority of very low probability sites being filtered out, even within the lower probability sites above the cut-off, there was a significant difference in how many of the sites evolved. Overall, the probability calculation combined with some basic contextual filtering including protein domain and accessible surface area is enough to yield a small number of high quality predictions that closely reflect both the sampling frequency of new sites during infection and the evolutionary potential on the larger scale of influenza strains.

Table 4.5. Contingency table for potential glycosylation sites across H3 in strains circulating between the years 1968 and 1985. The cut-off used to consider a site predicted was 0.025. Only sites above the probability cut-off of 0.005 were included in the analysis. χ^2 (1, N = 42) = 26.25, p < .0000001. All but 1 of the glycosylation sites in currently circulating H3 strains are in the predicted sites here.

Evolved	**Yes**	**No**
Predicted		
Yes	13	1
No	2	26

Looking specifically at the sequence predictions in 2019 it thus seems possible to make predictions for glycosylation sites in future H3N2 strains (Figure 4.8). Surprisingly very few new sites have a high probability of evolving. Among the high probability sites, four are sites that are commonly glycosylated in circulating strains, or have been in the past (81, 126, 133 and 144). They remain high probability of being sampled whilst being absent in the currently circulating strain. Among the remaining lower probability predictions that are above the cut-off several are the same as for the earlier strains that have never evolved in the dataset analysed in this study, including several sequences between residues 90 and 110. This suggests these might simply be structurally conserved regions that happen to remain close to evolving the consensus for N-linked glycosylation, and indeed, most residues in that region contact or are near key residues at the receptor binding site. Their overall conservation suggests

that these are not likely candidates for future glycosylation. The remaining high probability for previously widespread glycosylation sites might simply be a remnant of the point-mutation that caused the loss of the site, but it would be interesting to investigate if strains with high probability evolution at these sites tend to circulate and remain in the population better as these sites allow re-emergence of favourable glycosylation sites in these spots when conditions change, or immune individuals are encountered. Overall, these results indicate that future H3N2 glycosylation sites are likely to simply be sites that have been common in the past. For completely new glycosylation sites not previously seen, I would expect some intermediate mutations would be required in the low probability regions to prime a new site for subsequent evolution, which we might see in strains in a few years.

4.3.3.2. Evolution of glycosylation sites in H1N1

H1N1 is an interesting case since it emerged in the human population in 1918 causing the deadly Spanish flu (Kilbourne 2006). In 1918 only one site in the globular head was glycosylated, N87. In the subsequent decades several new glycosylation sites emerged and glycosylation has been argued to have reached saturation in the seasonal H1N1, although with some glycosylation changes where new sites replace old ones (e.g. N127 to N125 transition) (Sun *et al.* 2011) (Figure 4.9). Even more interestingly, in 2009 the swine flu pandemic was caused by an H1N1 strain that was very similar to the 1918 strain, also having only a single (N87) glycosylation site (the same site as in 1918) and overall similar antigenic structure (Xu *et al.* 2010).

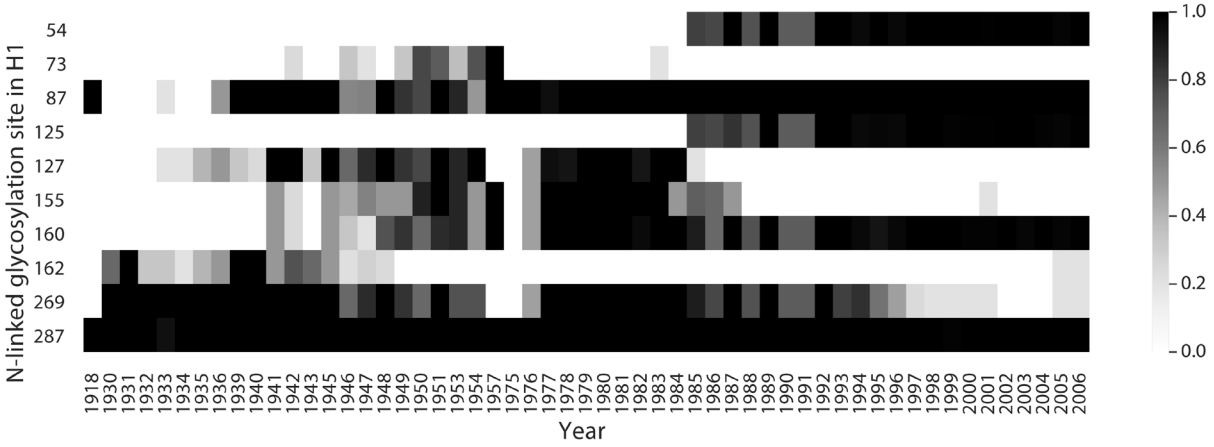


Figure 4.9. Timeline of glycosylated positions in H1 strains. The heat-map intensity displays the relative number of strains with a glycosylation site at the indicated position within the dataset in a given year. Some years between 1918 and 1930 have no sequencing data and have been left out. Seasonal H1N1 disappeared between ~1955-1975 due to several other pandemic strains becoming dominant in the world (H2N2 and H3N2).

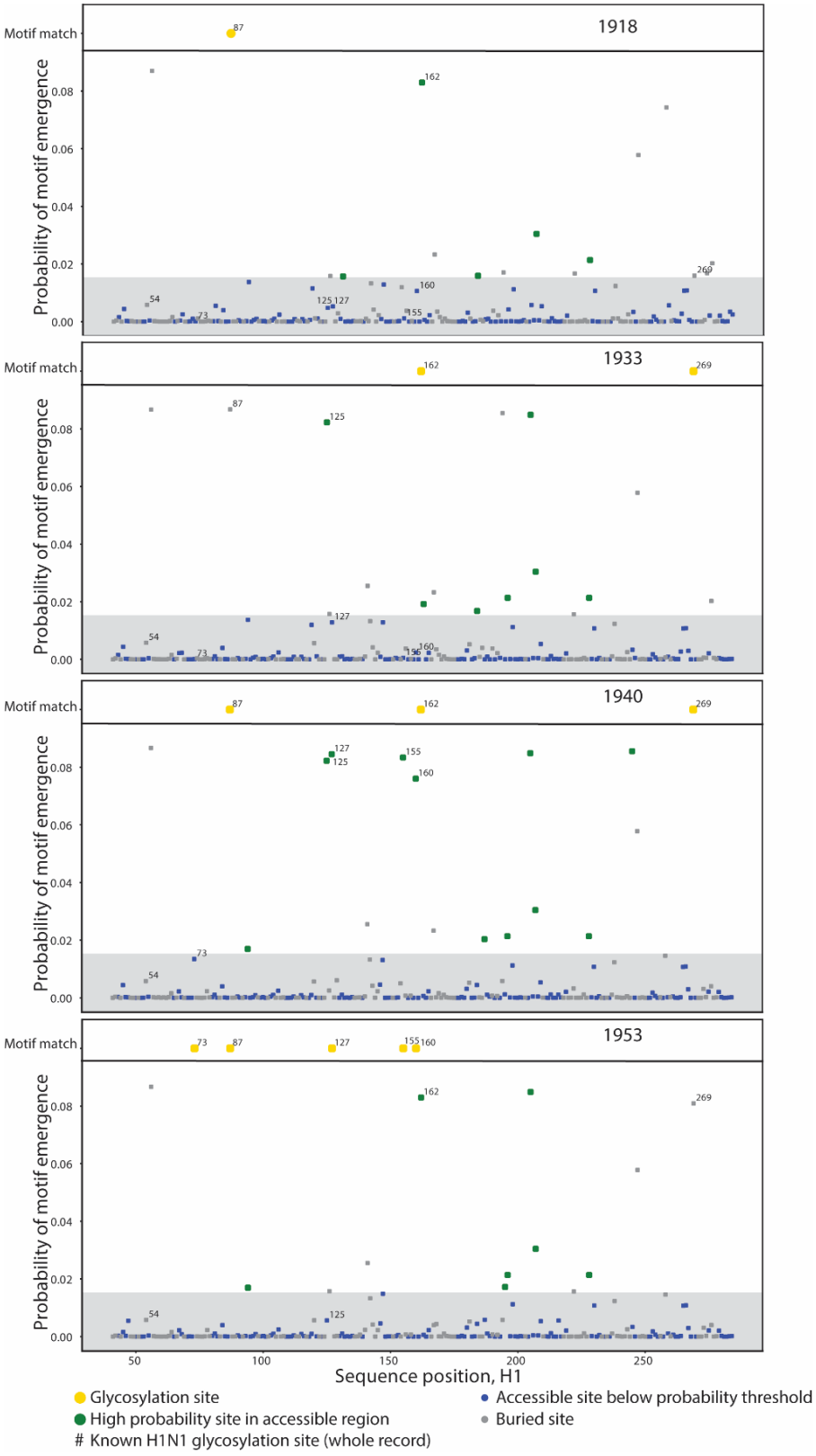


Figure 4.10. Probability of gain of glycosylation sites in HA of H1N1 spanning 1918 to 1953. The top predicted sites tend to be gained over a decade or so, N162 and N269 from 1918-1933, N87 from 1933-1940; N125, N127, N155 and N160 from 1940- 1953. As seen in H3N2 lost sites remain high probability and are often regained (e.g. N87, N269).

Overall, a sequence comparison between the 1918 and 2009 strains revealed that the sequence identity match was just over 94% (data not shown). This similarity is even higher than the sequence match between the 1918 strain and the H1N1 strain in 1933 which is the closest sequence available post 1918.

Doing the probability analysis on the 1918 strain and using the 1918 HA structure (PDB 3GBN) to assess relative accessible surface area yielded a relatively small number of highly predicted sites, and several at a medium probability, close to the cut-off (Figure 4.10). The two most important sites, N162 and N269 are both predicted as high probability in the 1918 strain, although N269 falls just under the cut-off for surface accessibility based on the published structure (Figure 4.10). These two sites will be the most important glycosylation sites until around 1940.

If we look again at the high probability sites subsequently around 1940 the landscape has shifted rather drastically and now several additional sites have emerged as very high probability sites. Within this set of sites are the majority of sites that will evolve over the subsequent 50 years, importantly N127, N125, N155 and N160 (Figure 4.10). These four sites are in fact all the sites that will be circulating in the H1N1 strains until after 1985, when N155 is lost and N54 is gained instead. N205, N207 and N228 (the two latter also present at 1918) also evolve in some of the strains in this dataset, however they are not important in the general viral population.

H1N1 captures many of the same dynamics seen in H3N2. It is clear that the glycosylation motif probability captures some interesting aspects of the evolutionary dynamics seen subsequently. It seems to predominantly be these high probability motifs that fix in the viral populations in the subsequent years. As seen in H3N2, motifs also tend to gain substitutions that make them higher probability before they subsequently evolve into glycosylation motifs, indicating that the likelihood of gain outcomes are essential for the evolutionary dynamics in viral sequence. As in H3, I determined whether the predicted residues were statistically correlated with the ones that subsequently evolved (Table 4.6). The predicted residues are significantly more likely to evolve than residues of lower probability, χ^2 (1, N = 24) = 13.7, p < . 0.0005. indicating this approach is able to identify the majority of the key residues that are able to evolve to become glycosylation sites in both H1 and H3 proteins.

Table 4.6. Contingency table for potential glycosylation sites across H1 in strains circulating between the years 1918 and 1955. The cut-off used to consider a site predicted was 0.025. Only sites above the probability cut-off of 0.005 were included in the analysis. χ^2 (1, N = 24) = 13.7, p <.0.0005. All the key glycosylation sites that evolved during this period are predicted as high probability sites in the analysis. Some additional sites that were sampled in some strains were also high probability, and 2 sites that were high probability did not evolve.

Evolved	**Yes**	**No**
Predicted		
Yes	9	2
No	0	13

Given the similarity between the 1918 and 2009 strains I wanted here to compare the two and determine if the same sites emerge as high probability in the current strain.

This information could be indicative of the future glycosylation evolution in current seasonal H1N1. Interestingly, in the 2015 season current H1N1 evolved a new glycosylation site at N162, which is the same residue as evolved first after 1918. To see if this was a high probability residue prior to that event I first looked at the 2014 strain (Figure 4.11).

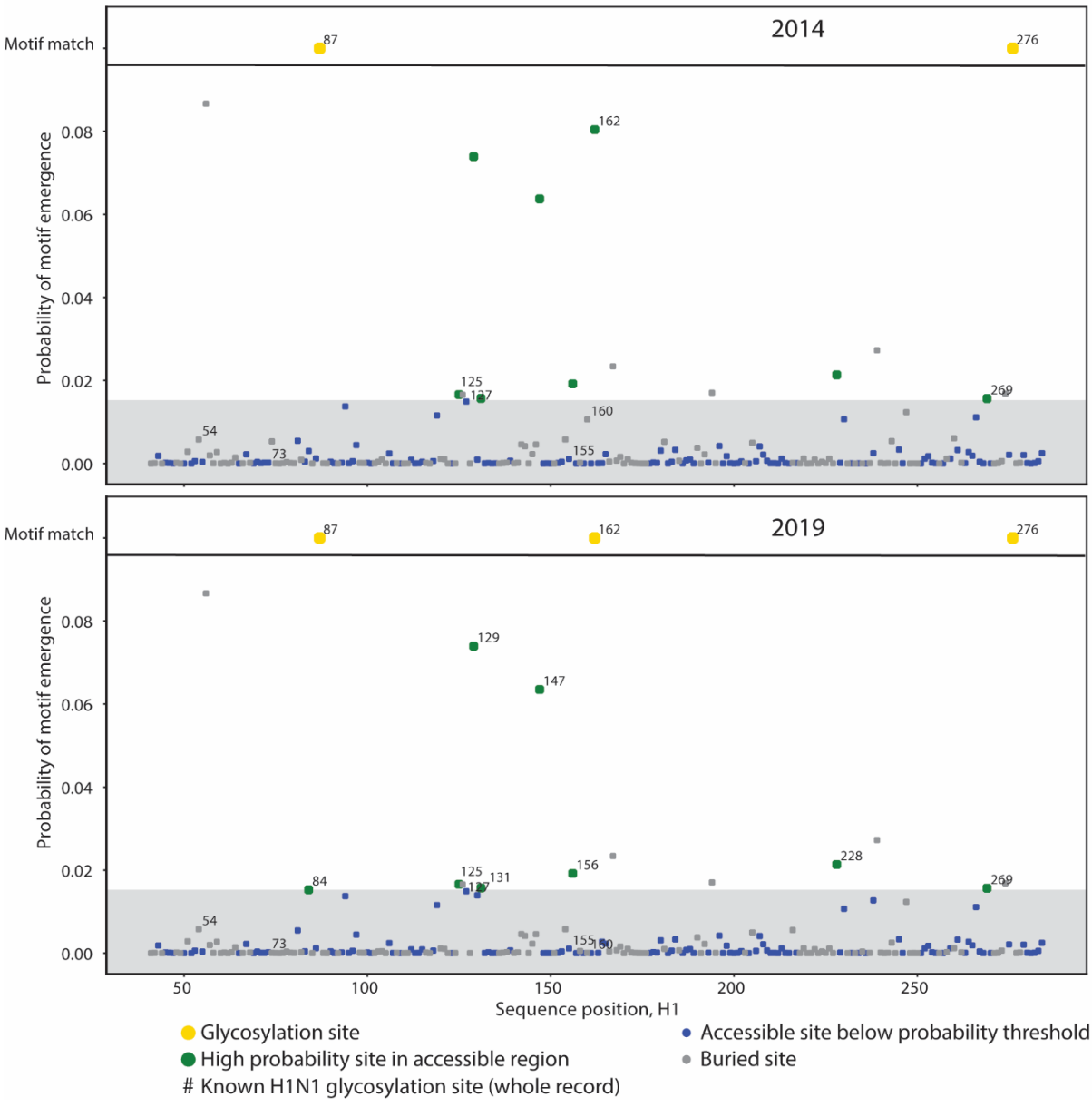


Figure 4.11. Probability of gain of glycosylation sites in HA of H1N1 post-2009. Similar high probability sites as in 1918 are seen, importantly 162 and 269. N162 emerged around the 2015 season. Other notable sites are 125, 127, 129 and 131 as they are all high probability and several of them have been key glycosylated sites in the past suggesting they are near an important antigenic region.

Interestingly, residue 162 is the site with the highest predicted probability prior to 2015. It is very striking that the probabilities and evolutionary history of 1918 and 2009 H1N1 are so similar. Comparing 1918 and 2019 strain probabilities, in the 2019 strain there are several high probability residues around 125, 127 and 129 that were not high probability in 1918. However, 125 and 127 were previously important glycosylation sites in seasonal H1N1 (between 1950s and 1980s), as mentioned. The fact that they are seen here as high probability sites already is a strong indication that they might evolve glycosylation in the next decade, which will be important for vaccination considerations. Important to consider is that vaccination efforts might likely shape evolution of glycosylation residues at these regions if vaccines target the antigenic regions otherwise protected at these sites. If the vaccines are designed to target regions without high probability glycosylation sites in the vicinity, I hypothesise that their success might be higher. In addition to these sites we also see residue 269 in this sequence as another high probability site. 269 was the second glycosylation to emerge after 162 in the 1930s and it will not be surprising to see this site emerge again in this currently circulating H1N1.

Overall, the use of this probability score adds to our ability to predict the evolutionary behaviour of these features in influenza proteins. These analyses have used fairly straight-forward contextual and structural filters, and it will be very exciting to take these results further and to implement other factors to improve the prediction quality. Improved contextual filters and more in depth understanding of the antigenic regions in influenza sequences would add valuable insights in combination with these probabilities. This has the potential to improve our ability to predict and deal with seasonal influenza from a new angle. It will be key to work alongside virologists and other experimentalists and to determine *in vivo* how these regions mutate and evolve and what effect different codon choices, and even different mutation rates per nucleotide, can have on the mutational outcomes of these features.

Determining if this type of analysis can be extended to other motif types will also be of interest. In the next section in this chapter I will do a similar analysis for phosphorylation sites in NP, to determine how these predictions work for motifs in more ambiguous structural and functional contexts.

4.3.4. Landscape of phosphorylation motif evolution in influenza A

The correlation between sites with high glycosylation motif potential in HA and the subsequent evolution of key functional motifs at those sites suggests that creating a potential motif landscape for evolution could be useful in other instances too.

For many motif classes in the influenza proteins the interactions and regulations are not as well understood as they are for glycosylation. Glycosylation is an unusually well behaved motif in that the consensus is very well characterised and the site can be recognised before the protein is folded in the endoplasmic reticulum (Yan & Lennarz 2005). Glycans are added to the peptide, which then folds,

meaning that most motif patterns end up being glycosylated (Aebi 2013). In contrast, for many phosphorylation sites consensus sequences are less well defined, partly because of many kinases recognising similar sequences and partly because it is not always clear if and when sites are functionally phosphorylated (Day *et al.* 2016; Landry *et al.* 2013; Levy *et al.* 2009; Ubersax & Ferrell 2007). This means that evaluating the functional impact of new motifs and their expected fitness impact is much more difficult. The better we understand the current set of interactions and the structural and functional dynamics of the influenza proteins the more informative a predicted motif space will be. Despite this it will be of interest to characterise sites where these motifs are currently primed to evolve. This information can then be used as we characterise further details of functionality in influenza. There can also be a benefit of knowing if primed sites that might be expected to emerge at a certain frequency but never do are present at important locations. This information could hint at regions where inadvertent phosphorylation has a large negative fitness impact which could also be functionally informative when combined with our current understanding of structure and function. In this section I have investigated currently circulating influenza A strain H1N1 in humans and determined the evolvability landscape for GSK3 phosphorylation sites in the protein NP. NP is a highly phosphorylated influenza protein, and GSK3 is a good candidate kinase for general influenza phosphorylation. This analysis can inform if there is a correlation between high probability sites and functional phosphorylation sites and to establish regions that are currently highly evolvable for the emergence of GSK3 sites.

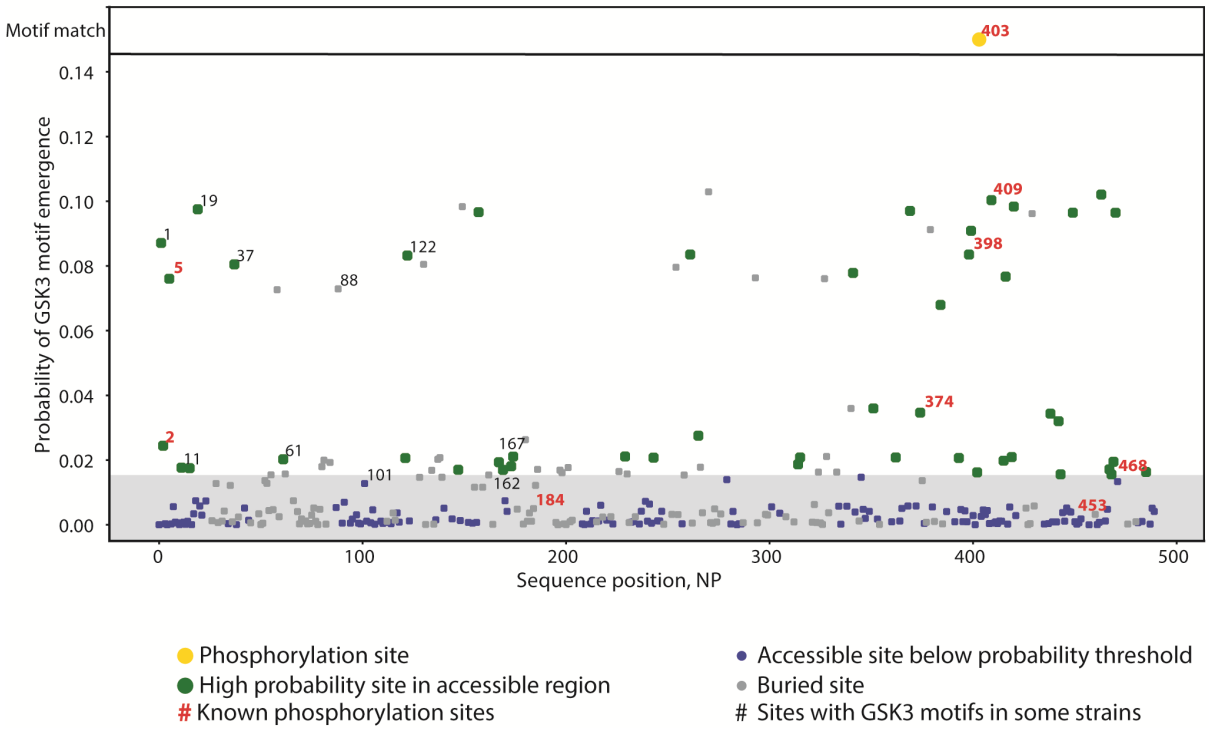


Figure 4.12. GSK3 evolvability landscape in NP from H1N1 in 2011. Red numbers in bold indicate known phosphorylated Ser/Thr sites (the Ser/Thr position being the red number+4 to account for the GSK3 consensus) and black numbers indicate the GSK3 motifs that have been sampled in NP history.

To perform the analysis, I used the same methodology as for the HA glycosylation sites in the previous section. The sequence was scanned for probability to evolve the motif consensus xxx[ST]xxx[ST] for each possible position in the protein sequence. The positions were then filtered for the solvent accessibility of the NP monomer, as phosphorylation sites are more likely to be present on the surface of the protein. In the final analysis a probability cut-off was also used to only include the high probability sites. I mapped the GSK3 motif sampling data from section 4.3.1 onto the predicted residues and also included the known phosphorylation sites described in section 3.3.2 (Figure 4.12).

There are several high probability sites for GSK3 motif evolution over the sequence, predominantly centred around the N-terminus and the C-terminus due to the accessible regions. Notably, the number of high probability phosphorylation sites is much larger than glycosylation sites in HA. The two main hot spots for new motif innovation potential are around residues 20-40 and residues 370-420. We can compare the evolutionary potential in this sequence to the results from section 1 in this chapter where I looked at motif sequence sampling across NP strains for among others GSK3 motifs (Figure 4.12, black numbering). Interestingly, near the N-terminus there are several overlapping motif sequences. At position 1, 11, 19, 37, 61 and 122 there are high probability motifs that have evolved in other related strains. These sites provide innovation potential in NP for new phosphorylation sites as the sampling and high probability enables new phosphorylation sites to be tested throughout infection. Surprisingly none of the high probability sequences around residues 400 have evolved in any of the strains in the current dataset. This is where the dimerization interface is which is a plausible reason for more restricted sequence sampling here. This is an indication that any mutations in this region are detrimental to the virus, most likely due to interfering with dimerization, and therefore, putative phosphorylation sites cannot be sampled. By contrast, the N-terminus is largely disordered, with overall fewer sequence restrictions, allowing exploration of sequence space to take place.

Interestingly, a majority of already phosphorylated residues have a high probability of gaining a GSK3 site as well, despite the majority being phosphorylated by other kinases. The only residue likely to be phosphorylated currently by GSK3 is at the motif at 403 (S407), although a kinase for this site has not been experimentally determined to date. The fact that phosphorylated sites have high probability of gaining other kinase recognition sites can be rationalised given that the Ser/Thr position requirement is already satisfied in the sequence. This provides an intriguing mechanisms by which a change in kinase recognition can change the timing or function of a protein without even changing the phosphorylated state of the protein. Instead, only the context and/or timing of phosphorylation is altered. This is a possibility that could play an important role in influenza adaptation in some contexts such as jumping to a new host species.

The high probability sites where no new phosphorylation sites have been observed in sequenced strains would still be expected to frequently gain sites through mutations during infection. If gain of phosphorylation at these sites carried a large fitness cost it could be expected that sequences would avoid those high probability codon choices. Determining the fitness cost of spontaneous gain of sites that are unfavourable for viral replication would be of great interest to better delineate the evolutionary mechanisms at play during motif evolution.

In conclusion, the evolvability of putative phosphorylation sites also depends in large part on the codon choices and the resulting probability of the sequence mutational outcomes. High probability sequences get sampled in history and are also highly likely to emerge around already existing phosphorylated residues. However, the overall picture of phosphorylation site evolvability is more convoluted than for glycosylation sites, with many additional factors at play. Phosphorylation sites are more impacted by surrounding residues and structural features that can change the phosphorylation dynamics. We currently do not have any examples of known phosphorylation sites in influenza that have evolved recently. All known sites are widely conserved. This is in part due to very limited phosphorylation studies in the different strains. To improve our understanding of phosphorylation site evolution, more data on phosphorylated sites in different strains would be an important step. Overall, an approach that considers the evolutionary probability of motifs alongside structural and functional features would likely be able to provide high quality evolutionary predictions for many different viruses, not only influenza. It would be of great interest to expand this approach in the future.

4.4. Discussion

In this chapter I have explored the predictability of motif evolution and sampling frequency of a range of motifs to determine how they evolved in influenza sequences in nature. I first found that new putative phosphorylation motif sites tend to emerge at high probability prior sequences through substitutions involving nucleotides with the highest mutations rates. This simply reflects the fact that high frequency mutations are more likely to result in high motif emergence rates allowing neutral motifs or motifs with small positive fitness effects to quickly evolve in a “first-come, first-served” manner. This corroborates the other observations that have been made in chapters 2 and 3 about the mutation frequencies of different nucleotides and their effect on the rate of various codon substitutions. I then looked specifically at the frequency of gain events for a range of motifs given average probabilities at different sites in ancestral influenza sequences and found that there is a positive correlation between how often putative motifs are gained and the evolvability of the ancestral sequence. This suggests there is predictability for putative motif evolution in viral sequences. Predicting functional outcomes based on sequence is especially important in viruses and viral evolution cycles as we try to predict

their evolution to aid future vaccination efforts. However, predicting motif evolution is also of interest in general in biology. Since motifs carry the potential of drastic functional and regulatory rewiring, being able to predict their evolution over time will allow us to gain a much better understanding of how complex regulatory and interaction networks emerge and change over time.

I further went on to look at the evolvability and predictability of glycosylation sites in HA proteins in H1 and H3 strains. Glycosylation is a particularly important motif for influenza and many other viruses, as it plays a role in structural stability of the coat, binding to host cells, and immune evasion (Alymova *et al.* 2016; Peng *et al.* 2019). Influenza is known to have very high variation in glycosylation states in the HA protein primarily, which is the main target for host antibodies (Altman *et al.* 2019). In both H1 and H3 there were few glycosylation sites when they entered the human population initially. I found that the probability of new glycosylation site emergence in those sequences correlated closely with the glycosylation motifs that subsequently evolved. The majority of new glycosylation motifs emerged from the top sequence predictions in the preceding years. Sites that were low probability initially, tended to gain mutations that increased their probability before subsequent motif gain in the population. Glycosylation sites that were lost in some seasons also tended to remain high probability sites, and were often regained in other strains and seasons several years later. In general, all key sites that were glycosylated in both H3 and H1 were high probability sites several years before being glycosylated. And few sites, once gained, became so mutated as to lose their high probability.

This suggests that virus strain that retain high probability of glycosylation around important antigenic sites are more dominant around the world. This could in theory results from strains that can give rise to progeny with those sites having increased fitness when infecting a large heterogeneous human population, since the antibody targeting will be diverse across the population. Taken together with the results in chapter 3 showing that glycosylation sites in the globular head of H1 and H3 also use less robust codons overall, an intriguing hypothesis is that evolution maximises variation at glycosylation sites in HA in influenza. It is possible that frequent gain and loss is increased, allowing viruses to constantly "check" whether different variants are suddenly more fit. There is an interesting overlap here between the host recognition of past antigens in the human population in different eras, so called "immune memory", and the glycosylation sites that are prevalent (Auladell *et al.* 2019). When targeting of a specific antigen becomes prevalent in the human population, adding a glycosylation site would increase fitness. As mutation rates have been shown to be high at antigenic sites, changes can accumulate while shielded from the immune system by the glycan (Civetta *et al.* 2016; Plotkin & Dushoff 2003). Since there appears to be a fitness cost of having too many glycosylated residues, possibly due to a decrease in the virus' ability to bind host cells, gain and loss of glycosylation sites are likely to accompany each other (Altman *et al.* 2019). When a new glycosylation site is gained, the virus can "sample" glycosylation loss at other sites until one that is not recognised by human antibodies (and thus has high fitness) is exposed. In this way, it is possible that maintaining high loss rate and high

evolvability of a set of sites, optimises the fitness balance between coat binding with host cells, recognition by host antibodies and the constant change of hosts infected. It is important to also consider that these observations could simply be consequences of the frequent gain and loss of glycosylation sites through the common mutations and that these codon choices might not be adaptive. In future work it will be key to determine if this is the case, as these dynamics and the predictability around glycosylation sites are central to vaccinations.

It will be interesting to see if the glycosylation site evolution pattern for the current seasonal strain of H1N1 will be similar to what has been predicted here. The similarity to the 1918 influenza is striking, and most of those sites in H1N1 have not been seen in the human population in >20 years, as H1N1 has not been a predominant seasonal strain for several decades (Finkelman *et al.* 2007). Many people alive today would not have antibodies targeted to those regions and thus it is easy to speculate that they will face very similar pressures to those in the past. Vaccination schemes are also known to induce strong selection on certain strains and drive evolution of new strains and potentially new glycosylation sites (Boni 2008). It would be of great interest to conduct experiments with the glycosylation site potential in mind to determine how vaccination attempts and site evolvability interact to determine the efficacy of vaccination efforts and the evolutionary trajectory of the current H1N1 strains.

Finally, in this chapter I expanded the motif evolution predictions to GSK3 phosphorylation sites. In theory, phosphorylation sites would be expected to have similar motif evolvability characteristics to glycosylation sites. They would be expected to face different selection pressures, but their motif patterns are similar in sequence space. However, we currently have a more limited understanding of the specific pressures and functional consequences of specific phosphorylation sites and other interaction sites. The ultimate outcome of a new site will depend completely on the important interactions, location and overall context of the protein. For specific phosphorylation sites I here predicted the future potential motif landscape. For phosphorylated residues in the protein NP there are no currently known instances of sites that have evolved since the sequence record began so predictions cannot be compared to subsequent evolutionary events. For putative new phosphorylation motifs there are several of the high probability predicted sites that have become a putative motif in related strains, as also seen for glycosylation. This corroborates those findings and suggests that with more data and a more extensive understanding of the functional motifs in different viral strains, motif predictions would be able to add value and a better understanding of where specific sites may evolve. This work can also act as a record for current high probability sites that can be used to design experimental approaches to study motif evolution and evolutionary outcomes in higher detail.

4.5. Materials and methods

4.5.1. Gain of putative phosphorylation sites in proteins NP

The NP alignment, phylogenetic tree and reconstructed ancestral sequences as outlined in section 3.5.1 were used in this analysis. This dataset consisted of 17564 strains including the reconstructed sequences at internal nodes in the phylogenetic tree.

To determine phosphorylation site emergence, I used a custom algorithm to scan through each sequence in the NP alignment and assign motifs based on regular expression matches. The regular expression used are summarised in Table 4.7. Primed CK sites were ignored for the purposes of this analysis as were alternative functional motif definitions.

Table 4.7. Phosphorylation motifs used to assess motif sampling based on definitions from ELM.

Motif	Regular expression
PKA motif	xRx[ST][^P]xx
GSK3 motif	xxx[ST]xxx[ST]
CK1 motif	[ED]xx [ST]xxx
CK2 motif	xxx[ST]xx[ED]

For each sequence in the alignment motifs were assigned to the aligned positions for each match. By cross-referencing the phylogenetic tree with the alignment and motif identities I then assigned a motif gain event for every instance where the ancestral strain in the phylogeny lacked the motif and it was present in a descendant as a result of a traceable mutation event (i.e. ignoring events resulting from gaps, or other non-mutation events). This was achieved using the same approach as I have outlined previously for both simulated data and tracking substitutions (see sections 2.7.5 and 3.5.1.4. I identified the nucleotide sequence as well as the peptide sequence in both the ancestor and descendant at the putative motif site. By comparing the two sequences I assigned the nucleotide and resulting codon substitution that yielded the new putative motif.

From the ancestral sequence I also assigned the motif emergence probability. Here I used the probability as determined using the mutation probabilities assigned from the phylogenetic tree, using the same 0.07 branch length to calculate the probability in each instance (see section 3.5.2.4 and 2.3.3 for details).

4.5.2. Motif emergence rate and probability in influenza A

To determine if there was a correlation between high probability motif evolution at specific sites in sequences and whether new putative motifs evolve I first determined all motif matches in all the strains in the NS1 alignment described previously. I then used the approach described in the previous section (4.5.1) where the alignment was cross referenced with the time-stamped phylogenetic tree. All

Table 4.8. Putative motif types identified as gained in NS1 strains across the phylogeny with the sequence expression used to identify matches.

Motif ID	Motif definition (regular expression)
TRG_NES_FLUM1	[ILMV]...[LV]..[ILV]
CLV_C14_Caspase3-7	[DSTE][^P][^DEWHFYC]D[GSAN]
CLV_PCSK_PC1ET2_1	KR.
MOD_CDK_SPK_2	...[ST]P[RK]
MOD_PKC_4_rare	[RK]...[ST].[RK]
MOD_CK2_Pho	...[ST]..[ST]
MOD_CDK_SPxxK_3	...[ST]P..[RK]
LIG_Clathr_ClatBox_1	L[IVLMF].[IVLMF][DE]
CLV_PCSK_PC7_1	R...[KR]R.
MOD_PK_1	[RK]..S[VI]..
MOD_NEK2_1	[FLM][^P][^P][ST][^DEP][^DE]
MOD_NEK2_2	[WYPCAG][^P][^P][ST][IFCVML][KRHYF]
MOD_SUMO_phospho	K.[EDST]
TRG_ENDOCYTIC_2	Y..[LMVIF]
LIG_SH3_4	KP..[QK]...
LIG_Actin_RPEL_3	[IL]..[^P][^P][^P][^P]R.....[IL]..[^P][^P][ILV][ILM]
MOD_PLK	.[DE].[ST][ILFWMVA]..
MOD_PKC_1	[RKH][RHK].[ST].[KRHN][KRLA]
LIG_SH2_STAT5	Y[VLTFIC]..
LIG_FHA_2	..(T)..[DE].
LIG_FHA_1	..(T)..[ILV].
LIG_SH2_SRC	Y[QDEVAIL][DENPYHI][IPVGAHS]
DEG_APCC_TPR_1_B	.[ILM]R
CLV_PCSK_SKI1_1	[RK].[AILMFV][LTKF].
MOD_GSK3_1	...[ST]...[ST]
LIG_PALB2_WD40_1	....WF..L
MOD_ProDKin_1	...[ST]P..
MOD_PKC_2_s	[RK]..[ST].[KR].
CLV_Caspase3-7_Flu	[DSTE].[^DEWHFC]D.
MOD_CK1_ED	[ED]..[ST]...
MOD_CK1_Pho	[ST]..[ST]...
MOD_PKC_3_s	.[RK].[ST].[RK]
MOD_PKA_2	.R.[ST][^P]..
LIG_SH3_1	[RKY]..P..P
LIG_SH3_3	...[PV]..P
MOD_PKB_1	R.R..[ST][^P]..
CLV_PCSK_FUR_1	R.[RK]R.
MOD_CK2_ED	...[ST]..[ED]
MOD_PIKK_1	...[ST]Q..
DEG_APCC_DBOX_1	.R..L..[LIVM].
MOD_SUMO_phospho_rev	[EDST].[K]
MOD_N-GLC_1	.N[^P][ST]..
TRG_LysEnd_APsAcLL_1	[DERQ]...L[LVI]
MOD_PKA_1	[RK][RK].[ST][^P]..
LIG_SH3_2	P..P.[KR]

instances of motif gain were determined by identifying traceable events where the ancestral strain lacked a motif sequence, and the descendant gained it through a mutation. Prior to the gain event the probability was calculated based on the codon usage as outlined previously. For all the motifs gained in influenza NS1, the number of gains were plotted against the average gain probability at the specific sites. The identified sites are listed in Table 4.8. The definitions are based on the ELM database, and

some additional motif variants that are known functional variant sequences in influenza have been included (however the correlations and conclusions are robust to these alternative definitions being ignored).

4.5.3. Determining glycosylation and phosphorylation probability over influenza sequences

I selected representative sequences from the relevant strains from relevant years. The strains used for the analysis are listed in Table 4.9.

I identified the evolvability of the glycosylation motif N[^P][ST] at each site across the sequence by calculating the probability using the method outlined in section 3.5.2.4 derived from the phylogenetic substitution bias estimates. A branch length of 0.07 was used to allow relative comparison across an arbitrary branch length for the different sequences. Sequence positions were adjusted to represent the mature peptide by subtracting 15 for H3N2 or 16 for H1N1. The mature peptide numbers were used to correlate residues with high surface accessible area.

Table 4.9. Strains used in glycosylation motif evolution analysis.

HA type + Year	Strain used
H3 1973	A/Hong_Kong/11/1973
H3 1981	A/Bilthoven/4791/1981
H3 1995	A/Hong_Kong/3/1995
H3 1997	A/Auckland/10/1997
H3 2006	A/Malaysia/33817/2006
H3 2019	A/California/08/2019
H1 1918	A/South Carolina/1/1918
H1 1933	A/WS/1933
H1 1940	A/Hickox/1940
H1 1953	A/Netherlands/001R1/1953
H1 2014	A/Wisconsin/V140586/2014
H1 2019	A/South Dakota/30/2019

Table 4.10. PDB structures used for accessible surface are determination.

Strain-Year	PDB
H1N1-1918	3GBN
H1N1-2009	3LZG
H3N2-1968	1ha0

Surface accessible area was determined using precalculated surface accessible areas from the PDB files (Table 4.10). The EBI service PISA was used (Krissinel & Henrick 2007). To calculate relative accessible surface areas , the maximum solvent accessibility values for each residue were assigned based on Tien *et al*, (2013). Relative accessibility was calculated as the ratio of solvent accessible surface area of the residue as calculated in PISA, and the maximum accessible area. For the relative accessibility cutoff, commonly used ratios of 0.15-0.20 was adhered to (Miller *et al.* 1987). I used a stringent relative solvent accessibility ratio of 0.19. I also used a probability cutoff to filter to only include high and medium probability motif evolvability sites. This probability cutoff was 0.015, which was assigned empirically and reflects the 0.07 branch length used for calculations.

To assess statistical significance, I used the filtered sites and assigned them into groups: Predicted and Evolved. Predicted sites were sites that were over a more stringent threshold of 0.025, and sites between 0.005 and 0.025 were considered not predicted. All sites under 0.005 were excluded completely which includes the vast majority of total sites (among these sites probabilities are too low to see any motif sampling or evolution). Sites were considered evolved if they have been identified as key glycosylated sites in literature, or if they were identified as having been sampled through a traceable mutation in the strain record. To test the significance of this contingency table I used chi-square.

For assessing the GSK3 evolvability landscape over NP, the strain A/Moscow/WRAIR4314T/2011 was used. For surface accessibility determination the PDB file 4DYA was used. The approach here was the same as for glycosylation evolvability determination outlined above.

Chapter 5

General discussion

5.1. Overview

In this thesis I have investigated the mechanisms that govern the evolution of motifs in viral proteins. Motifs play a central role in regulating protein function, signalling, localisation and activity levels among other central roles in cells. We still only have a limited understanding of the scope and complexity of how motifs allow proteins in viruses to achieve the levels of regulation and function that they do. Within this scope the rapid evolution of viruses and in particular the possibility of rapid changes to motifs is a property that is central to the evolution of increasingly complex systems. As I have shown over the chapters of this thesis, having accurate models of the underlying mechanisms that drive the evolution of certain molecular features can help us better understand the course of evolution. This more fundamental approach has been made possible by our increasing amounts of sequence data, alongside recent insights into nucleotide specific mutation rates in some of these viruses. Combining the nucleotide level details of evolutionary mechanisms with the perspective gained from population genetics has been a big part of gaining an increasingly detailed understanding of the evolution of molecular features on the protein level.

This thesis has hopefully provided a new perspective on evolution at the intersection of many of these fields and a framework that can be used to study the history of evolution in viruses. The aim is to expand the lessons on motif evolution from this work to other organisms, contexts and systems where motifs play an equally important role in regulation. In this final chapter I will discuss some of the current gaps in our knowledge surrounding motif evolution and some limitation in the approach taken here. In particular, I will highlight the importance of expanding our knowledge of how motifs function and the associated fitness effects of the sequence variants within a motif. I will also discuss the various other applications where this codon-centric motif evolutionary dynamics approach could be useful to study evolution, such as in heterogeneous cancer tissues, eukaryotic pathogens and within the evolution of highly expressed eukaryotic genes. I will also speculate on how the findings in this chapter can be of relevance in various experimental approaches where inadvertent mutations and mutational outcomes can impact the experiments. These include synthetic evolution experiments and other systems

that rely on very highly expressed genes. Finally, I will outline my perspective on the future of the motif evolution field and the questions that I think are the most relevant to address in the near future.

5.2. The importance of accurate motif consensus definitions

Throughout the work in this thesis the concept of the codon space for a motif has been a central point in the calculations and the results when studying the probabilities and mutational outcomes. The codons accessible through various point mutations will greatly impact the probabilities of certain events. The assumption is that the motif space constitutes a set of codons of equal functional relevance and any substitutions can be considered silent within that space. In reality this varies on a motif to motif basis and for the majority of motifs we do not have a detailed enough understanding of the subtle variations function that might result from the use of different residues (see Saksela & Permi (2012) for discussion on SH3 consensus variation and specificity). It can be the case that one amino acid is preferred and another one is only tolerated but with lower binding affinity. However, in most cases they will both be part of the motif consensus definition. Ultimately, a motif is simply a linear sequence of amino acids that are recognised by a binding pocket in a domain. The motif consensus is a human derived simplification that describes the motif observations. However, the delineations made in defining these sites can be a lot more diffuse in nature. It is likely that in many cases similar amino acids will have similar properties when binding, however additional interactions outside the binding site can also stabilise a less optimal motif sequence thus allowing a wider range of amino acids to be part of the motif consensus (Stein & Aloy 2010). All of these factors will impact an analysis of the evolvability of motifs since the evolutionary trajectory will be based on an optimisation of fitness based on all of the relevant contributing factors.

In this study I specifically selected a subset of motif classes where motif definitions are well established such as phosphorylation Ser/Thr sites, glycosylation sites and localisation sites. In these sites the mode of binding and the residue compatibility is known, and the functionality of the allowed residues is also relatively unbiased. For some other motifs I used the most well established consensus sequence based on the literature and the ELM databases (Gouw *et al.* 2017). But to be able to look at the evolvability of a larger range of motifs we need to expand our knowledge of the details of motif interactions and consensus sites so the different contribution of binding between the allowed residues can be accounted for.

A relevant example for consensus site ambiguity is for the recognition sites of casein kinase I and II. The canonical consensus is for an acidic residue offset by a gap of two residues to the phosphorylated S/T ([ST]xx[ED] or [ED]xx[ST] respectively). However, many known CK sites are also phosphory-

lated when they have another phosphorylated residue in the E/D position, since the phosphate can replace the negative charge of the side chains (Hrubey & Roach 1990; Marin *et al.* 2003). These functional considerations are important for the fitness impact in the case of mutations. Having a better understanding of these subtleties would greatly improve our ability to apply these motif evolvability methods to other interesting motif sites, and to better elucidate more high complexity sites such as sequentially phosphorylated sites that rely on previous phosphoresidues to modify further motif sites (Zhou *et al.* 2017). This kind of layered motif information in sequences adds significant complexity in the decision making and regulation a cell can achieve.

The current best methods for investigating the binding strength of motif variant and establishing a highly accurate set of consensus residues include high throughput phage display and other similar methods such as *in vivo* peptide screens that allow exhaustive sequence variants to be probed against many relevant binding partners (Davey *et al.* 2017; Ravarani *et al.* 2018). In addition, structures of the binding interaction and the domain can provide us with important information about the binding pocket and features of the surface surrounding the binding pocket such as hydrophobicity and charge that can affect the interaction through adjacent residues. Overall, we need several different high throughput methods and more focus on motif biology to get to a point where we can get an even better understanding of the evolution of these sites.

5.3. Complex logic and decision making

In this thesis I have mentioned the importance of motifs in complex cellular decision making, signalling and cellular logic. This is achieved through the interaction and regulation of several conditional motifs in protein sequence (Neduva & Russell 2005; Van Roey *et al.* 2012). This complexity through information processing is essential for eukaryotic life and is a big part of what enables complex organisms to carry out their functions.

This complexity in decision making and regulation is also important in the evolution of viral function. Viruses rely on short compact genomes to efficiently carry out replication and transmission. Motifs allow viruses to evolve complex regulation and interactions through rapid sampling of point mutations (Thomas *et al.* 2011). The high mutation rate of viruses generates great versatility in sequence space (Russell *et al.* 2018). This versatility and sampling allows viruses to evolve new functionality and adapt to changing hosts, host conditions and antiviral defences by changing protein structural features, regulatory mechanisms and interactions (Sun *et al.* 2011; Tokuriki *et al.* 2009).

Regulation and decision making in cells are important features that can differ greatly between organisms despite close relatedness and high genome similarity. It is likely that small numbers of point mutations in highly related proteins between closely related organisms are able to significantly change the

regulation and behaviour of those proteins and the ultimate function within the organism. These properties have been observed in closely related bacterial species. In a notable example, it was found that two highly related bacteria from different ecological contexts had large shifts in their PTM patterns, suggesting motif evolution can aid evolutionary adaptation in this manner (Li *et al.* 2014b). As also observed in chapter 4, in the sequence probability for evolving new phosphorylation site, a shift in a kinase motif around an already existing phosphorylated serine or threonine could alter protein function. This shift could easily be overlooked in functional characterisation of homologous proteins. This has important consequences for how we use model systems and homologous proteins to infer function through relatedness.

By having a better understanding of the ways in which motifs evolve we can expand our knowledge of how related proteins evolve new functionality through localisation, degradation, interactions and modification changes. By using motif evolution, we can gain a better understanding of the expected overlap in function and regulation from related proteins than can be gained from sequence similarity alone.

5.4. Motif evolvability and vaccines

An important implication for the concept of motif evolvability and potential sequence landscapes for circulating viral strains is within vaccine development. Seasonal influenza vaccinations build on evolutionary modelling of circulating strains. Each year the strains that are predicted to become the globally predominant ones are included when creating the vaccine (Agor & Özaltın 2018). These most commonly include some H1N1, H3N2 and influenza b and c strains. However, the efficacy of the seasonal vaccine is variable, some years other strains emerge that were not included in the vaccine and sometimes mutations in the strains that were included reduces the benefit of the vaccination (Belongia & McLean 2019; Lewnard & Cobey 2018). As I have discussed in this thesis there is an important relationship between common antigenic sites on the surface of HA that are targeted by vaccines and the surface glycans at key glycosylation sites (Hütter *et al.* 2019; Tate *et al.* 2014). It has been shown through several lines of experiments and computational work that antigenic sites overlap with sites that can become glycosylated and that antigenic sites and sites protected by glycans often show relatively high amounts of sequence variation and mutations compared to other HA sites (Plotkin & Dushoff 2003; Thyagarajan & Bloom 2014). However, to date no work to my knowledge has considered the implications of glycosylation site evolvability on vaccine efficacy in the short term. A recent paper did discuss the impact of antigenic mutational escape which can readily impact the antibody affinity and negatively affect vaccination efforts, and observed that some of the highest impact mutations involved gain of glycosylation sites (Doud *et al.* 2018).

An intriguing possibility for influenza evasion of the immune system is that glycans emerge at targeted antigenic sites from high probability, "primed" motifs through the fitness pressure imposed by the immune system. The antigenic sites, thus protected by glycans, can accumulate mutations and changes, separated from any continuous pressure from antibody recognition. Subsequent cycles of loss and gain of glycosylation sites can expose a mutated antigenic region which no longer will be recognised by the antibodies that targeted the prior site. In this way a small set of sites, in turn becoming glycosylated and losing glycosylation sites can evade the immune system continuously, which is what influenza has managed to do successfully in the human population despite our best efforts. This relies on having high probability glycosylation sites around the key antigenic sites so that they can be sampled frequently in the population, and can sweep across the population quickly when selection pressures change. This is consistent with the observations in my preliminary analysis of glycosylation site evolvability in historical and current strains of influenza. Glycosylation sites that are primed to evolve subsequently emerge in the population. But importantly, also glycosylation sites that were previously functional – and have since been lost – consistently remain high probability sites in descendant strains. In addition to this, several high probability sites often exist in close vicinity in sequence space suggesting that the combined probability of sampling will be higher around specific antigenic regions.

Considering vaccine development within this framework of gain and loss cycles of high probability glycosylation sites, I would hypothesise that strains targeted in vaccines would frequently experience glycosylation shifts reducing vaccine efficacy. By using the glycosylation probability map of the circulating strains, a more informed and hopefully improved selection of strains might be possible when designing the vaccine. A range of strains with different glycosylation potential or specifically choosing strains that lack high probability glycosylation sites at key regions could all be tried to improve vaccination efforts. Ultimately, directly choosing target antigenic regions of the vaccination would be the ideal way to target sites with a lower likelihood to evolve protective glycans.

Within this pursuit it would also be key to consider how vaccination impacts the glycosylation state of descendant strains generally. The imposed selection pressure generated by vaccination efforts affects downstream glycosylation sites and the strains that become globally widespread in later seasons. This will have widespread implications not only for surface proteins but also the rest of the passenger mutations that are selected alongside. It will be of future interest to use the motif evolvability framework to both better inform current vaccination efforts and to also be more informed of the evolutionary and functional consequences of the selection imposed through vaccinations on the regulation and activity of descendant viral strains.

In the current literature there are already some indications that glycosylation site evolution during vaccine development and the selection pressure imposed through targeted vaccinations do impact the evo-

lutionary trajectory of strains (Lam *et al.* 2017; Zost *et al.* 2017). However, these questions and approaches are still in early stages and in-depth exploration of sequence variation and interactions with vaccines have not yet been performed extensively. Our understanding of many viral dynamics is limited to a few prevalent lab strains, but large scale, high throughput investigations to characterise the subtle differences between historical and circulating strains have not been carried out extensively.

In addition, there is a marked difference between pandemic strains and seasonal stains. Pandemic strains generally come from animal adapted viruses, importantly birds and swine as they are important live-stock and large viral reservoirs. Antigenic sites and glycosylation frequency and potential act very differently in animals compared to humans as is evident by the fact that most pandemic strains have very few surface glycans, usually 1-2 where seasonal strains have 5-8 generally. Part of the reason for this is thought to be the shorter life span in swine populations, which prevents "immune-memory" from inducing selective pressure for surface change (Luoh *et al.* 1992; Sheerar *et al.* 1989).

Taken together, evolution optimises viral sequences in response to a whole range of factors including antigenic sites, glycosylation sites, immune memory and vaccination efforts – which all shape the evolutionary trajectory. A multi-scale model that incorporates codon-level evolutionary changes within and outside of motif contexts with several other factors could present a better idea of viral infection, adaptation and possible outbreaks (Rüdiger *et al.* 2019).

Within viral adaptation and vaccination, the viral quasispecies will also be important to understand and model. The variable sampling and viral heterogeneity that make up the quasispecies is important for the evolutionary trajectory. And as explored here, the quasispecies population is likely to be a direct manifestation of the mutational and sequence biases present in the founding population. Genotypes resulting from high probability mutational outcomes will make up a larger portion of the quasispecies population. This heterogeneity might directly affect the structure and properties of individual virions and impact immune evasion and vaccine adaptation. These questions will be of interest in future research.

5.5. Motif evolution in eukaryotes and human disease

Motifs evolved predominantly as modular systems in eukaryotes and have expanded the complexity and regulatory capacity of complex eukaryotic organisms (Bradley & Beltrao 2019; Kim *et al.* 2014; Tompa *et al.* 2014). Human proteins rely on regulation and function through phosphorylation and binding interactions and many other motifs, and it has been well established that changes to motifs can lead to disease such as cancers and aggregation based diseases (Sambataro & Pennuto 2017; Van Roey & Davey 2015; Wang *et al.* 2015).

Historically it was initially thought that since mutation rates are low, natural selection will swamp any effect seen in mutational biases during evolution (Yampolsky & Stoltzfus 2001). Since it has been shown that even when mutation rates and population sizes are small, mutational biases can influence adaptation due to a "first come, first served" effect (McCandlish & Stoltzfus 2014; Stoltzfus & McCandlish 2017). However, the authors have argued that when mutation rates are high, biased outcomes would play a smaller role as the most fit genotype will always emerge (McCandlish & Stoltzfus 2014; Stoltzfus & McCandlish 2017). But the roles of biased outcomes have since been shown to play a more varied role than that. The interplay between selection and variation is clearly a lot more dynamic in nature that initially thought. Even in RNA viruses the biased genotype space has been shown to be important (Lauring *et al.* 2012). The results in this thesis also corroborate the importance of mutational outcomes and rates in the evolutionary trajectories of RNA viruses. Taken together, it is reasonable to assume that different codon choices in motifs can play an important role in the evolutionary outcomes and fitness in eukaryotes.

I hypothesise there are two main paradigms where motif evolvability and codon choice within motifs can be important considerations for human diseases. The first mode is through codon choice and motif probability in oncogenes and cancers. Cancer tissues have more error-prone replication and higher overall replication rates (Duesberg *et al.* 2000; Hanahan & Weinberg 2011). Cancer tissues are known to be heterogeneous and to accumulate mutations resulting in dysregulation of many signalling pathways and both gain and loss of function within key proteins and pathways (Dagogo-Jack & Shaw 2017). In many ways the properties of these cancer tissues approach those of fast evolving viral quasispecies. In these environments it will be of interest to investigate the propensity of evolution of new motifs, how frequently cancer-linked mutations create or remove motifs and whether there are any links between cancer mutation frequency and primed motif sites in human populations.

The second mode of how motif evolvability can impact human disease is not on the cell replication level but on gene expression level. As has been mentioned in this thesis it is known that codon optimisation is prevalent in highly expressed genes in humans (Lavner & Kotlar 2005). There is also evidence to indicate that there is an increase in translational robustness among highly expressed genes (Drummond *et al.* 2005). The transcription error rate is approximately the same as the error rate for viral polymerases (Gout *et al.* 2017). Compared to normal human replication error rate, the transcription error rate is several orders of magnitude larger. In these genes there could be a substantial impact on dysfunctional protein load in the cell as a result of transcription errors. Another effect that could be synergistic with errors during transcription is mistranslation, where previous work suggests that there can be an overlap between codons that get mistranslated and the codon space for high probability mutations, in part because of the structure of the genetic code and codon space (Kramer & Farabaugh 2007).

The theories and methodology developed in this thesis for the analysis of mutation rate and codon space on functional outcomes is also readily applicable to transcription and translation errors. Taken together these factors could induce an increased faulty protein load within the cell with unexpected regulation and interactions. This could in theory lead to an increased aggregation risk of many proteins, in particular disordered ones, and overall reduction in fitness and an increased risk of cell senescence or apoptosis which is linked with degenerative diseases. A recent study determined the overall transcription and translation error rates to be as high as $5x10^{-5}$ and 10^{-3} respectively. Whether the error rates are high enough to impose a significant amount of faulty proteins in specific contexts is not yet known, but it could be important in contexts of high protein expression and stress conditions. Slow accumulation over time of many faulty proteins can overwhelm the quality control machinery which can be important in emergence of neurodegenerative disease (Ciechanover & Kwon 2017). Motif based interactions and misregulation is likely to exacerbate this problem, and codon choice can play a role here e.g. through increasing loss in degrons and changing the regulation of intrinsically disordered proteins (Babu *et al.* 2011). Expanding the ideas and theories of motif evolvability to these various human contexts would be very interesting for future projects.

5.6. Motif evolution in other pathogens

Another context which would be interesting to explore from the motif evolution framework is other human viruses and pathogens. In this thesis I focused on RNA viruses as they are good model systems for evolution over short time spans. Expanding this kind of analysis to other viruses such as larger DNA viruses and retroviruses would be interesting in order to delineate the timescales, population sizes and mutation rates that impact the fitness and evolutionary properties of motifs in a more detailed manner. In addition, it would be of interest to look at other important human pathogens such as *P. falciparum* (malaria) and other eukaryotic parasites that are known to evolve host-like motifs (Chemes *et al.* 2015). Bacteria that use motifs could also be of interest, however bacteria tend to use fewer motifs in general and operate quite differently from viruses mechanistically so I would expect motif evolution to have less of a fitness importance in those systems. However, there are indications that codon choice from a directly functional perspective are important there too, as shown with the examples of codon selection at stop codons in *E. coli* that have been reported (Korkmaz *et al.* 2014). In addition, both eukaryotic pathogens and bacteria often secrete proteins with motifs (Via *et al.* 2015). Overall, I would expect the underlying mutational biases given codon usage and mutation rate to behave in similar ways in most contexts, however the fitness impact of codon usage within different systems will likely be correlated with the population size, replication rate and mutation rate such that in slowly replicating organisms the pressure on optimising codon use for mutational impact might not be enough to detect any codon bias in some contexts.

5.7. Evolvability in experimental design and synthetic evolution

Another interesting implication of evolvability of functional features in proteins is in in the sphere of experimental design and evolutionary experiments. The most commonly used expression systems and techniques including PCR and culture growth have high and variable error rates and thus the potential of introducing biased changes into sequences that are being studied, replicated, or otherwise probed (Zhao *et al.* 2014). It is likely that the mutational landscape given sequence choice and nucleotide specific mutation rates can vary significantly in many steps of this process (Romero & Arnold 2009). However, in many contexts and experiments little thought is given to sequence design at the initial steps of experiments, and silent substitutions are often assumed to have no impact on subsequent analyses. In systems of high expression and relatively high mutation rates, considering the impact of potential high frequency mutations and their functional and structural implications could prove beneficial. This includes motifs, and importantly chemical modifications but could also impact other types of interactions and structural features, such as RNA secondary structure (Kudla *et al.* 2009). Important implications could be in structural determination when high expression and high homogeneity is desirable (van der Laan *et al.* 1989). In expression systems probing interactomes, there could also be an impact of high probability substitutions altering specific interactions through mutation.

The mutation probability landscape, whilst being a source for noise and variability in these contexts, could also be leveraged to improve experimental design in some cases. Within the fields of directed evolution and creating proteins with different potential functionality, having a prior understanding of the outcomes of high frequency mutations and their functional consequences could guide targeted efforts to sample variable regions of sequence space. Sequences could even be designed with variable robustness and mutational landscapes, driving exploration of sequence space in preferred regions. I think an interesting prospect within this scope is to explore motif evolution in experimental systems, as we already have a good understanding of the limitations and functions of many residues in motif space. Combined with the results from this thesis, there is also a framework for expected mutational behaviour in many contexts. Exploring how complex interactions of several motifs and motif classes evolve would be of great interest. These kinds of interactions form the equivalents of logic gates within cells which I have briefly mentioned previously (Singh 2014). Exploring the extent to which cells and sequences can evolve these logic operations and how we can design them for different cellular uses would be an exciting future direction.

5.8. Predicting evolution

A key question posed in this thesis is whether motif evolution can be predicted. As already addressed, evolution is inherently random and it is impossible to exactly predict the occurrence of individual point mutations in individual sequences. However, it is also well established that on the phenotypic

level, within the scope of defined selection pressures, evolutionary outcomes can be remarkably predictable, giving rise to convergent evolution for a wide array of proteins (Chemes *et al.* 2015; Larter *et al.* 2018). Ultimately, an important determinant of the predictability of functional phenotypic evolution is the bias in the diversity of variation and the frequency of variants in populations (Doud *et al.* 2018; Stoltzfus & McCandlish 2017). In this thesis I have tried to evaluate these properties for motif evolution. I hypothesise that the organisation and sequence properties of motifs make them particularly susceptible to biases in sampling the mutational landscapes through the mechanisms established here. That is because motifs are primary sequence features in protein space made up of few amino acids that are functionally defined collectively. In this thesis I have developed a framework for thinking about evolution of these features that incorporate both sequence level information as well as basic aspects of population dynamics.

Taken together, my perspective is that some aspects of motif evolution are predictable on the population level and there are several pieces of information and evolutionary properties that make this the case. Firstly, that nucleotide mutation frequencies are biased and some occur more often than others leading to differences of several orders of magnitude between some mutational outcomes. Secondly, that motifs are a functional unit that will affect organismal fitness together and importantly that we already know what is allowed in a residue position to maintain function and what results in function loss. Thirdly, the understanding that population level effects are key in driving evolutionary outcomes. The evolution of specific changes in individual sequences appears to be less important for evolutionary outcomes than high frequency changes that are more likely to result in overall population evolution and fixation. The frequency of events and the impact on the population level needs to be accounted for to understand evolutionary trajectories. Within most facets of evolutionary predictability, bulk effects are likely to tend towards the probabilities and frequencies of their respective event outcomes, resulting in predictability. In effect, nature optimises the statistical likelihood of outcomes in a way that maximises fitness.

Since motif-like sequences that are not functional also do evolve frequently however, I think it is necessary to combine population level expectations of motif frequencies with other important contributors to motif functionality. This includes an understanding of the protein context as well as the structure of the protein – importantly which regions are buried and exposed and which are disordered. By combining the genotype fitness landscape and the genotype space that sequences are able to sample (reach) more easily, evolutionary trajectories can be rationalised as the intersection of genotype availability and genotype fitness. These two properties can be visualised as separate genotype landscapes, and the most likely evolutionary outcome is where peaks in both landscapes overlap (Figure 5.1).

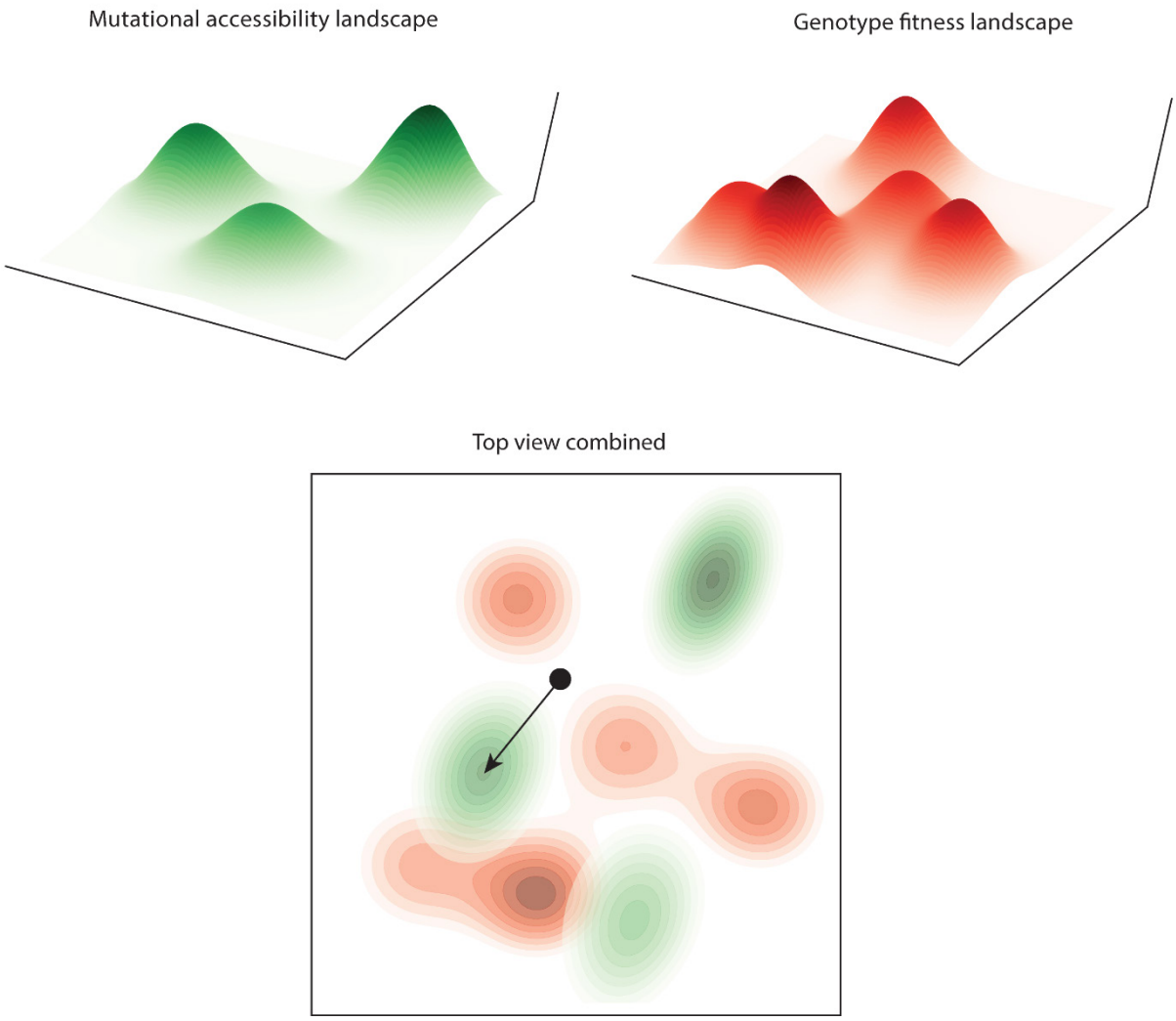


Figure 5.1. Defining the mutational trajectory and accessibility landscape in sequence evolution. The combined effect of probable mutational outcomes in genotype space with the fitness effects of those outcomes will together shape evolution. Both mutational outcomes and fitness landscapes are multidimensional, constantly shifting landscapes that depend on sequences, interactions and surroundings. Overall, mutational outcomes depend on codon spaces, mutation rates and the prior sequence state as discussed in this thesis. These properties are themselves evolvable which adds another dimension to the complexity of the optimisation problem faced by evolution.

Using these relatively basic information sets and improving our understanding of contextual structure-function relationships and variation will undoubtedly improve our ability to predict motif evolution. The early findings in this thesis for glycosylation sites importantly show that there is a good potential for predicting functional motifs in contexts where mutational accessibility and genotype fitness are relatively easy to predict. Simply using the calculated probability alongside residue surface exposure is fairly limited information but results in higher accuracy predictions for the future of glycosylation site evolution in influenza. One can get to a similar point for other motifs, importantly other modifications sites. And with improving protein-protein interaction datasets and functional contexts for e.g. influenza proteins, the effects of sampling interaction motifs and ligand motifs could also be more accurately predicted. However, with the current datasets this is outside the scope of predictability as we do not know most of the important current functional interactions of e.g. NS1 in influenza.

There are some obvious limitations to what can be predicted for motifs using this method alone. Since it deals with neutral frequencies and population level probabilities it will not accurately predict whether an individual strain will or will not gain a specific motif. Getting prediction to this level will require a deeper understanding of the selection pressures for other features for the same protein sequence as well as the environmental pressures affecting a particular strain. The information and ability to calculate probabilities that I have presented here would greatly benefit from being combined with other type of evolutionary predictors. Of future interest would be to combine the properties that shape evolution and use machine learning to try to predict outcomes. The probabilities, historical record data and sequence space I use here would be potential sources for learning for a machine learning algorithm. This could be an important future direction to consider and could lead to developments for influenza and viral evolution prediction.

5.9. The wider impact and future directions of motif evolution research

Overall it has been immensely interesting to delve into the field of motif biology and motif evolution and to try to gain a better understanding of some of the fundamental ways in which information processing and regulation evolves through motifs. Motif biology seems to have flown under the radar in wider biology to some extent. There has been some focus on post-translation modifications and some specific well-known proteins such as histones, but the vast majority of the million motifs expected to operate in the human proteome are largely unexplored (Tompa *et al.* 2014). Especially the unique properties of motifs in evolution, which were only recently fully appreciated through the findings of extensive motif-mimicry in viruses and other pathogens (Chemes *et al.* 2015; Davey *et al.* 2012, 2015; Hagai *et al.* 2014).

One of the most important goals for motif research at the moment is to develop better methods and expand our targeted effort to carefully characterise motif definitions (Davey *et al.* 2017; Ravarani *et al.* 2018). An accurate understanding of motif residue determinants and the subtle interactions between key residue positions and adjacent residues that can modulate binding will be essential to expand both motif evolutionary understanding as well as our ability to find existing motifs in the human proteome.

It would also be of value to see an increasing focus in determining accurate mutation rates for viruses and other organisms. In this thesis I have relied on a recent set of publications from Pauly *et al*, (2017), where influenza mutation rates where carefully and accurately measured. The evolutionary analyses greatly benefited from this data. To date, very few viruses and organisms have this kind of data available but I imagine that having similar datasets would prove very useful for many disparate

fields in biology. In addition to having accurate per nucleotide mutation rates, improving our understanding of other factors (e.g. repair mechanisms, environments that introduce mutations and specific sequence contexts) that impact nucleotide mutation biases would be useful, and is already starting to prove useful in cancer tissues (Petljak *et al.* 2019). Sequence features such as dinucleotides other nucleotide contexts are known to impact mutation frequencies (Aggarwala & Voight 2016). Having a clear mechanistic understanding in different organisms and contexts of the variables that play a role in determining mutation rates would ultimately improve any method that tries to predict evolutionary outcomes.

Ultimately, I would like to expand all of the concepts explored for motif evolution in this thesis into other areas of evolutionary characterisation. Whilst in this thesis I have focused exclusively on motifs, the sequence properties of amino acids in motifs are fundamentally the same as those for other structural features. An interaction surface could theoretically be defined as a multi-dimensional space of conditional amino acid allowances that define a functional codon space. Our current ability to understand the sequence limitations and allowances that govern the features and interactions in those different contexts is unfortunately very limited. In a motif, the contributing residues are clearly defined. If we had a good understanding of these same properties for other protein features, the robustness and evolvability approach could be applied in non-motif instances. A good place to start could be in other protein features that rely on a small number of limited residues. Considering an enzyme active site; if we can define a set of theoretical residues that are important for the function and a set of alternative functional substitutions, robustness at the active site would hypothetically have similar properties to a motif. Exploring the range of protein features where some of the ideas from the motif evolution approach can be used to elucidate evolutionary features would be a very interesting future direction.

For motif evolution itself I would like to see these ideas and predictions tested in experimental systems. Although we designed aspects of long-term motif evolution experiments during the first year of my PhD, we decided to focus on the theoretical aspects initially, as this was critical to design better experiments. Exploring the fitness impact of stop codon choice, and the functional innovation of primed phosphorylation sites *in vivo* would be a great next step in delineating the real-world impact of motif evolvability. An increased focus on the diversity in motifs between closely related viruses and in other organisms would also be of interest both as experimental problems and as future computational questions to address. Experimental influenza work uses a very small number of common lab strains, but previous work has shown that the diversity between strains for e.g. glycosylation and phosphorylation is very high even in closely related strains (Petri *et al.* 1982; Russell *et al.* 2018). Probing the functional diversity between these strains from a motif and PTM perspective could have some important implications for our ability to infer complex systems functionality from studying related organisms.

There is still a large landscape of unexplored problems in motif biology, and there are many fundamental and exciting questions to address in the near future. This thesis has been a first look into some of the detailed dynamics of motif evolution in a fast-evolving RNA virus. These observations can hopefully elucidate many general principles about motif evolution across organisms. I have been able to find some exciting new ways in which selection can fine tune the mutational properties of important features and this carries some interesting implications for our understanding of the evolution of complexity in general. I am greatly looking forward to learning more about motifs and their fundamental role in shaping living systems.

Bibliography

Aderem, A. (2005). Systems Biology: Its Practice and Challenges. *Cell*, **121**(4), 511–513.

Aebi, M. (2013). N-linked protein glycosylation in the ER. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, **1833**(11), 2430–2437.

Aggarwala, V., & Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, **48**(4), 349–355.

Agor, J. K., & Özaltın, O. Y. (2018). Models for predicting the evolution of influenza to inform vaccine strain selection. *Human Vaccines & Immunotherapeutics*, **14**(3), 678–683.

Ahn, H.-J., Kim, S., & Nam, H.-W. (2007). Nucleolar translocalization of GRA10 of Toxoplasma gondii transfectionally expressed in HeLa cells. *Korean Journal of Parasitology*, **45**(3), 165–174.

Akhouri, R. R., Sharma, A., Malhotra, P., & Sharma, A. (2008). Role of Plasmodium falciparum thrombospondin-related anonymous protein in host-cell interactions. *Malaria Journal*, **7**(1), 63.

Allen, J. D., & Ross, T. M. (2018). H3N2 influenza viruses in humans: Viral mechanisms, evolution, and evaluation. *Human Vaccines & Immunotherapeutics*, **14**(8), 1840–1847.

Altman, M. O., Angel, M., Košík, I., … Yewdell, J. W. (2019). Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. *MBio*, **10**(2), e00204-19.

Alto, N. M., & Orth, K. (2012). Subversion of cell signaling by pathogens. *Cold Spring Harbor Perspectives in Biology*, **4**(9), a006114–a006114.

Alymova, I. V, York, I. A., Air, G. M., … McCullers, J. A. (2016). Glycosylation changes in the globular head of H3N2 influenza hemagglutinin modulate receptor binding without affecting virus virulence. *Scientific Reports*, **6**, 36216.

Andino, R., & Domingo, E. (2015). Viral quasispecies. *Virology*, **479**–**480**, 46–51.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, **181**(4096), 223 LP – 230.

Arold, S., Franken, P., Strub, M.-P., … Dumas, C. (1997). The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure*, **5**(10), 1361–1372.

Arrese, M., & Portela, A. (1996). Serine 3 is critical for phosphorylation at the N-terminal end of the nucleoprotein of influenza virus A/Victoria/3/75. *Journal of Virology*, **70**(6), 3385 LP – 3391.

Auladell, M., Jia, X., Hensen, L., … Kedzierska, K. (2019). Recalling the Future: Immunological Memory Toward Unpredictable Influenza Viruses . *Frontiers in Immunology* , p. 1400.

Avalos, R. T., Yu, Z., & Nayak, D. P. (1997). Association of influenza virus NP and M1 proteins with cellular cytoskeletal elements in influenza virus-infected cells. *Journal of Virology*, **71**(4), 2947 LP – 2958.

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions*, **44**(5), 1185 LP – 1200.

Babu, M. M., van der Lee, R., de Groot, N. S., & Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology*, **21**(3), 432–440.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, **21**(3), 381–395.

Barford, D. (2011). Structure, function and mechanism of the anaphase promoting complex (APC/C). *Quarterly Reviews of Biophysics*, **44**(2), 153–190.

Baudin, F., Bach, C., Cusack, S., & Ruigrok, R. W. (1994). Structure of influenza virus RNP. I. Influenza virus nucleoprotein melts secondary structure in panhandle RNA and exposes the bases to the solvent. *The EMBO Journal*, **13**(13), 3158–3165.

Beale, R., Wise, H., Stuart, A., Ravenhill, B. J., Digard, P., & Randow, F. (2014). A LC3-Interacting Motif in the Influenza A Virus M2 Protein Is Required to Subvert Autophagy and Maintain Virion Stability. *Cell Host & Microbe*, **15**(2), 239–247.

Bedi, S., & Ono, A. (2019). Friend or Foe: The Role of the Cytoskeleton in Influenza A Virus Assembly. *Viruses*, **11**(1), 46.

Belalov, I. S., & Lukashev, A. N. (2013). Causes and Implications of Codon Usage Bias in RNA Viruses. *PLOS ONE*, **8**(2), e56642.

Belinky, F., Babenko, V. N., Rogozin, I. B., & Koonin, E. V. (2018). Purifying and positive selection in the evolution of stop codons. *Scientific Reports*, **8**(1), 9260.

Belongia, E. A., & McLean, H. Q. (2019). Influenza Vaccine Effectiveness: Defining the H3N2 Problem. *Clinical Infectious Diseases*. doi:10.1093/cid/ciz411

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., & Tawfik, D. S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, **444**(7121), 929–932.

Bertram, S., Glowacka, I., Steffen, I., Kühl, A., & Pöhlmann, S. (2010). Novel insights into proteolytic cleavage of influenza virus hemagglutinin. *Reviews in Medical Virology*, **20**(5), 298–310.

Bhattacharyya, R. P., Reményi, A., Yeh, B. J., & Lim, W. A. (2006). Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits. *Annual Review of Biochemistry*, **75**(1), 655–680.

Blount, Z. D., Borland, C. Z., & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli *Proceedings of the National Academy of Sciences*, **105**(23), 7899 LP – 7906.

Bockaert, J., Marin, P., Dumuis, A., & Fagni, L. (2003). The 'magic tail' of G protein-coupled receptors: an anchorage for functional protein networks. *FEBS Letters*, **546**(1), 65–72.

Boeynaems, S., Alberti, S., Fawzi, N. L., … Fuxreiter, M. (2018). Protein Phase Separation: A New Phase in Cell Biology. *Trends in Cell Biology*, **28**(6), 420–435.

Bohne-Lang, A., & von der Lieth, C.-W. (2005). GlyProt: in silico glycosylation of proteins. *Nucleic Acids Research*, **33**(Web Server issue), W214–W219.

Boivin, S., Cusack, S., Ruigrok, R. W. H., & Hart, D. J. (2010). Influenza A virus polymerase: structural insights into replication and host adaptation mechanisms. *The Journal of Biological Chemistry*, **285**(37), 28411–28417.

Boni, M. F. (2008). Vaccination and antigenic drift in influenza. *Vaccine*, **26**, C8–C14.

Bottermann, M., & James, L. C. (2018). Chapter Thirteen - Intracellular Antiviral Immunity. In M. Kielian, T. C. Mettenleiter, & M. J. B. T.-A. in V. R. Roossinck, eds., , Vol. 100, Academic Press, pp. 309–354.

Bouvier, N. M., & Palese, P. (2008). The biology of influenza viruses. *Vaccine*, **26 Suppl 4**(Suppl 4), D49–D53.

Bradley, D., & Beltrao, P. (2019). Evolution of protein kinase substrate recognition at the active site. *PLOS Biology*, **17**(6), e3000341.

Brandis, G., & Hughes, D. (2016). The Selective Advantage of Synonymous Codon Usage Bias in Salmonella. *PLOS Genetics*, **12**(3), e1005926.

Brister, J. R., Ako-adjei, D., Bao, Y., & Blinkova, O. (2014). NCBI Viral Genomes Resource. *Nucleic Acids Research*, **43**(D1), D571–D577.

Brooks, C. L., & Gu, W. (2011). p53 regulation by ubiquitin. *FEBS Letters*, **585**(18), 2803–2809.

Brown, C. J., Johnson, A. K., & Daughdrill, G. W. (2010). Comparing models of evolution for ordered and disordered proteins. *Molecular Biology and Evolution*, **27**(3), 609–621.

Bruggeman, L. A. (2007). Viral Subversion Mechanisms in Chronic Kidney Disease Pathogenesis. *Clinical Journal of the American Society of Nephrology*, **2**(Supplement 1), S13 LP-S19.

Buljan, M., Chalancon, G., Dunker, A. K., … Babu, M. M. (2013). Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Current Opinion in Structural Biology*, **23**(3), 443–450.

Burch, C. L., & Chao, L. (2000). Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature*, **406**(6796), 625–628.

Bustamante, C. D. (2005). Population Genetics of Molecular Evolution. *Statistical Methods in Molecular Evolution*, 63–99.

Cady, S. D., Luo, W., Hu, F., & Hong, M. (2009). Structure and function of the influenza A M2 proton channel. *Biochemistry*, **48**(31), 7356–7364.

Canman, C. E., Lim, D.-S., Cimprich, K. A., … Siliciano, J. D. (1998). Activation of the ATM Kinase by Ionizing Radiation and Phosphorylation of p53. *Science*, **281**(5383), 1677 LP – 1679.

Cao, S., Liu, X., Yu, M., … Liu, W. (2012). A nuclear export signal in the matrix protein of Influenza A virus is required for efficient virus replication. *Journal of Virology*, **86**(9), 4883–4891.

Carlini, D. B., Chen, Y., & Stephan, W. (2001). Bias , mRNA Secondary Structure and Gene Expression in the Drosophilid Alcohol Dehydrogenase Genes Adh and Adhr. *Genetics*, **159**, 623–633.

Carrillo, B., Choi, J.-M., Bornholdt, Z. A., Sankaran, B., Rice, A. P., & Prasad, B. V. V. (2014). The influenza A virus protein NS1 displays structural polymorphism. *Journal of Virology*, **88**(8), 4113–4122.

Chemes, L. B., de Prat-Gay, G., & Sánchez, I. E. (2015). Convergent evolution and mimicry of protein linear motifs in host–pathogen interactions. *Current Opinion in Structural Biology*, **32**, 91–101.

Chen, H., Alvarez, J. J. S., Ng, S. H., Nielsen, R., & Zhai, W. (2018a). Passage Adaptation Correlates With the Reduced Efficacy of the Influenza Vaccine. *Clinical Infectious Diseases*, **69**(7), 1198–1204.

Chen, W., Calvo, P. A., Malide, D., … Yewdell, J. W. (2001). A novel influenza A virus mitochondrial protein that induces cell death. *Nature Medicine*, **7**(12), 1306–1312.

Chen, Y. Q., Wohlbold, T. J., Zheng, N. Y., … Wilson, P. C. (2018b). Influenza Infection in Humans Induces Broadly Cross-Reactive and Protective Neuraminidase-Reactive Antibodies. *Cell*, **173**(2), 417-429.e10.

Cheng, Q., Cross, B., Li, B., Chen, L., Li, Z., & Chen, J. (2011). Regulation of MDM2 E3 ligase activity by phosphorylation after DNA damage. *Molecular and Cellular Biology*, **31**(24), 4951–4963.

Ciechanover, A., & Kwon, Y. T. (2017). Protein Quality Control by Molecular Chaperones in Neurodegeneration . *Frontiers in Neuroscience* , p. 185.

Civetta, A., Ostapchuk, D. C. M., & Nwali, B. (2016). Genome Hotspots for Nucleotide Substitutions and the Evolution of Influenza A (H1N1) Human Strains. *Genome Biology and Evolution*, **8**(4), 986–993.

Cock, P. J. A., Antao, T., Chang, J. T., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Cruz, E., Cain, J., Crossett, B., & Kayser, V. (2018). Site-specific glycosylation profile of influenza A (H1N1) hemagglutinin through tandem mass spectrometry. *Human Vaccines & Immunotherapeutics*, **14**(3), 508–517.

Csibra, E., Brierley, I., & Irigoyen, N. (2014). Modulation of Stop Codon Read-Through Efficiency and Its Effect on the Replication of Murine Leukemia Virus. *Journal of Virology*, **88**(18), 10364 LP – 10376.

Cuchillo, C. M., Nogués, M. V., & Raines, R. T. (2011). Bovine Pancreatic Ribonuclease: Fifty Years of the First Enzymatic Reaction Mechanism. *Biochemistry*, **50**(37), 7835–7841.

Dabrowski, M., Bukowy-Bieryllo, Z., & Zietkiewicz, E. (2015). Translational readthrough potential of natural termination codons in eucaryotes--The impact of RNA sequence. *RNA Biology*, **12**(9), 950–958.

Dagogo-Jack, I., & Shaw, A. T. (2017). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, **15**, 81.

Daniels, R., Kurowski, B., Johnson, A. E., & Hebert, D. N. (2003). N-Linked Glycans Direct the Cotranslational Folding Pathway of Influenza Hemagglutinin. *Molecular Cell*, **11**(1), 79–90.

Database resources of the National Center for Biotechnology Information. (2017). *Nucleic Acids Research*, **46**(D1), D8–D13.

Davey, N. E., Cyert, M. S., & Moses, A. M. (2015). Short linear motifs -- ex nihilo evolution of protein regulation. *Cell Communication and Signaling*, **13**(1), 43.

Davey, N. E., Seo, M.-H., Yadav, V. K., … Ivarsson, Y. (2017). Discovery of short linear motif-mediated interactions through phage display of intrinsically disordered regions of the human proteome. *The FEBS Journal*, **284**(3), 485–498.

Davey, N. E., Travé, G., & Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends in Biochemical Sciences*, **36**(3), 159–169.

Davey, N. E., Van Roey, K., Weatheritt, R. J., … Gibson, T. J. (2012). Attributes of short linear motifs. *Molecular BioSystems*, **8**(1), 268.

Day, E. K., Sosale, N. G., & Lazzara, M. J. (2016). Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. *Current Opinion in Biotechnology*, **40**, 185–192.

Denamur, E., & Matic, I. (2006). Evolution of mutation rates in bacteria. *Molecular Microbiology*, **60**(4), 820–827.

Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L., & Rexach, M. (2003). Disorder in the nuclear pore complex: The FG repeat regions of nucleoporins are natively unfolded. *Proceedings of the National Academy of Sciences*, **100**(5), 2450 LP – 2455.

Diehl, N., & Schaal, H. (2013). Make yourself at home: viral hijacking of the PI3K/Akt signaling pathway. *Viruses*, **5**(12), 3192–3212.

Domingo, E., Martínez-Salas, E., Sobrino, F., … Holland, J. (1985). The quasispecies (extremely

heterogeneous) nature of viral RNA genome populations: biological relevance — a review. *Gene*, **40**(1), 1–8.

Domingo, E., Sheldon, J., & Perales, C. (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews*, **76**(2), 159 LP – 216.

Dosztanyi, Z., Chen, J., & Dunker, A. (2006). Disorder and Sequence Repeats in Hub Proteins and Their Implication for Network Evolution. *Journal of Proteome …*, **5**(11), 2985–2995.

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**(16), 3433–3434.

Doud, M. B., Lee, J. M., & Bloom, J. D. (2018). How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Nature Communications*, **9**(1), 1386.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(40), 14338–14343.

Duan, G., & Walther, D. (2015). The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLOS Computational Biology*, **11**(2), e1004049.

Duesberg, P., Stindl, R., & Hehlmann, R. (2000). Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(26), 14295–14300.

Duff, K. C., & Ashley, R. H. (1992). The transmembrane domain of influenza A M2 protein forms amantadine-sensitive proton channels in planar lipid bilayers. *Virology*, **190**(1), 485–489.

Dunker, A. K., Lawson, J. D., Brown, C. J., … Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, **19**(1), 26–59.

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, **6**. doi:10.1038/nrm1589

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.

Efeyan, A., & Serrano, M. (2007). p53: Guardian of the Genome and Policeman of the Oncogenes. *Cell Cycle*, **6**(9), 1006–1010.

Eisfeld, A. J., Neumann, G., & Kawaoka, Y. (2014). At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology*, **13**, 28.

Fan, H., Walker, A. P., Carrique, L., … Fodor, E. (2019). Structures of influenza A virus RNA polymerase offer insight into viral genome replication. *Nature*, **573**(7773), 287–290.

Fan, R. L. Y., Valkenburg, S. A., Wong, C. K. S., … Poon, L. L. M. (2015). Generation of Live Attenuated Influenza Virus by Using Codon Usage Bias. *Journal of Virology*, **89**(21), 10762 LP – 10773.

Fares, M. A. (2015). The origins of mutational robustness. *Trends in Genetics*, **31**(7), 373–381.

Félix, M.-A., & Wagner, A. (2006). Robustness and evolution: concepts, insights and challenges from a developmental model system. *Heredity*, **100**, 132.

Findly, D., Herries, D. G., Mathias, A. P., Rabin, B. R., & Ross, C. A. (1961). The Active Site and Mechanism of Action of Bovine Pancreatic Ribonuclease. *Nature*, **190**(4778), 781–784.

Finkelman, B. S., Viboud, C., Koelle, K., Ferrari, M. J., Bharti, N., & Grenfell, B. T. (2007). Global

patterns in seasonal activity of influenza A/H3N2, A/H1N1, and B from 1997 to 2005: viral coexistence and latitudinal gradients. *PloS One*, **2**(12), e1296–e1296.

Fischer, N. W., Prodeus, A., Malkin, D., & Gariépy, J. (2016). p53 oligomerization status modulates cell fate decisions between growth, arrest and apoptosis. *Cell Cycle*, **15**(23), 3210–3219.

Fischer, U., Jänicke, R. U., & Schulze-Osthoff, K. (2003). Many cuts to ruin: a comprehensive update of caspase substrates. *Cell Death And Differentiation*, **10**, 76.

Fisher, C. K., & Stultz, C. M. (2011). Constructing ensembles for intrinsically disordered proteins. *Current Opinion in Structural Biology*, **21**(3), 426–431.

Flock, T., Weatheritt, R. J., Latysheva, N. S., & Babu, M. M. (2014). Controlling entropy to tune the functions of intrinsically disordered regions. *Current Opinion in Structural Biology*, **26**, 62–72.

Floquet, C., Hatin, I., Rousset, J.-P., & Bidou, L. (2012). Statistical Analysis of Readthrough Levels for Nonsense Mutations in Mammalian Cells Reveals a Major Determinant of Response to Gentamicin. *PLOS Genetics*, **8**(3), e1002608.

Fogelmark, K., Peterson, C., & Troein, C. (2016). Selection shapes transcriptional logic and regulatory specialization in genetic networks. *PLoS ONE*, **11**(2), 1–19.

Fontes, M. R. M., Teh, T., Toth, G., … Kobe, B. (2003). Role of flanking sequences and phosphorylation in the recognition of the simian-virus-40 large T-antigen nuclear localization sequences by importin-α. *Biochemical Journal*, **375**(2), 339 LP – 349.

Fortuna, M. A., Zaman, L., Wagner, A., & Bascompte, J. (2017). Non-adaptive origins of evolutionary innovations increase network complexity in interacting digital organisms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**(1735), 20160431.

Foster, P. L. (2006). Methods for determining spontaneous mutation rates. *Methods in Enzymology*, **409**, 195–213.

Frensing, T., Kupke, S. Y., Bachmann, M., Fritzsche, S., Gallo-Ramirez, L. E., & Reichl, U. (2016). Influenza virus intracellular replication dynamics, release kinetics, and particle morphology during propagation in MDCK cells. *Applied Microbiology and Biotechnology*, **100**(16), 7181–7192.

Fung, H. Y. J., Birol, M., & Rhoades, E. (2018). IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Current Opinion in Structural Biology*, **49**, 36–43.

Furth, N., Gertman, O., Shiber, A., … Ravid, T. (2011). Exposure of bipartite hydrophobic signal triggers nuclear quality control of Ndc10 at the endoplasmic reticulum/nuclear envelope. *Molecular Biology of the Cell*, **22**(24), 4726–4739.

Gajadhar, A. S., & White, F. M. (2014). System level dynamics of post-translational modifications. *Current Opinion in Biotechnology*, **28**, 83–87.

Gallwitz, M., Enoksson, M., Thorpe, M., & Hellman, L. (2012). The Extended Cleavage Specificity of Human Thrombin. *PLOS ONE*, **7**(2), e31756.

Gamblin, S. J., & Skehel, J. J. (2010). Influenza hemagglutinin and neuraminidase membrane glycoproteins. *The Journal of Biological Chemistry*, **285**(37), 28403–28409.

Ganai, R. A., & Johansson, E. (2016). DNA Replication-A Matter of Fidelity. *Molecular Cell*, **62**(5), 745–755.

Gibson, T. J. T., Dinkel, H., Van Roey, K., … Uhlen, M. (2015). Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Communication and Signaling : CCS*, **13**(1), 42.

Gomez-Navarro, N., & Miller, E. (2016). Protein sorting at the ER–Golgi interface. *The Journal of Cell Biology*, **215**(6), 769 LP – 778.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, **551**(7678), 45–50.

Goonesekere, N. C. W., & Lee, B. (2008). Context-specific amino acid substitution matrices and their use in the detection of protein homologs. *Proteins: Structure, Function, and Bioinformatics*, **71**(2), 910–919.

Gout, J.-F., Li, W., Fritsch, C., … Vermulst, M. (2017). The landscape of transcription errors in eukaryotic cells. *Science Advances*, **3**(10), e1701484–e1701484.

Gouw, M., Michael, S., Sámano-Sánchez, H., … Gibson, T. J. (2017). The eukaryotic linear motif resource – 2018 update. *Nucleic Acids Research*, **46**(D1), D428–D434.

Grady, E. F., Bohm, S. K., & Bunnett, N. W. (1997). Turning off the signal: mechanisms that attenuate signaling by G protein-coupled receptors. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, **273**(3), G586–G601.

Green, D. R., & Kroemer, G. (2009). Cytoplasmic functions of the tumour suppressor p53. *Nature*, **458**, 1127.

Greenspan, D., Palese, P., & Krystal, M. (1988). Two nuclear location signals in the influenza virus NS1 nonstructural protein. *Journal of Virology*, **62**(8), 3020 LP – 3026.

Grossegesse, M., Doellinger, J., Fritsch, A., … Nitsche, A. (2018). Global ubiquitination analysis reveals extensive modification and proteasomal degradation of cowpox virus proteins, but preservation of viral cores. *Scientific Reports*, **8**(1), 1807.

Hafner, A., Bulyk, M. L., Jambhekar, A., & Lahav, G. (2019). The multiple mechanisms that regulate p53 activity and cell fate. *Nature Reviews Molecular Cell Biology*, **20**(4), 199–210.

Hagai, T., Azia, A., Babu, M. M., & Andino, R. (2014). Use of Host-like Peptide Motifs in Viral Proteins Is a Prevalent Strategy in Host-Virus Interactions. *Cell Reports*, **7**(5), 1729–1739.

Hale, B. G., Randall, R. E., Ortı, J., Jackson, D., & Jackson, D. (2008). The multifunctional NS1 protein of influenza A viruses, 2359–2376.

Han, H., Cui, Z. Q., Wang, W., … Zhang, X. E. (2010). New regulatory mechanisms for the intracellular localization and trafficking of influenza A virus NS1 protein revealed by comparative analysis of A/PR/8/34 and A/Sydney/5/97. *Journal of General Virology*, **91**(12), 2907–2917.

Han, Q., Chang, C., Li, L., … Xu, K. (2014). Sumoylation of Influenza A Virus Nucleoprotein Is Essential for Intracellular Trafficking and Virus Growth. *Journal of Virology*, **88**(16), 9379 LP – 9390.

Hanahan, D., & Weinberg, R. A. a. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, **144**(5), 646–674.

Harmon, T. S., Holehouse, A. S., Rosen, M. K., & Pappu, R. V. (2017). Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *ELife*, **6**, e30294.

Harris, B. Z., & Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *Journal of Cell Science*, **114**(18), 3219 LP – 3231.

Haynes, C., Oldfield, C. J., Ji, F., … Iakoucheva, L. M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Computational Biology*, **2**(8), 0890–0901.

Heikkinen, L. S., Kazlauskas, A., Melén, K., … Saksela, K. (2008). Avian and 1918 Spanish Influenza A Virus NS1 Proteins Bind to Crk/CrkL Src Homology 3 Domains to Activate Host Cell Signaling. *Journal of Biological Chemistry* , **283**(9), 5719–5727.

Heldt, F. S., Frensing, T., & Reichl, U. (2012). Modeling the intracellular dynamics of influenza virus replication to understand the control of viral RNA synthesis. *Journal of Virology*, **86**(15), 7806–7817.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(22), 10915–10919.

Hintze, A., & Adami, C. (2008). Evolution of Complex Modular Biological Networks. *PLOS Computational Biology*, **4**(2), e23.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., & VandePol, S. (1982). Rapid evolution of RNA genomes. *Science*, **215**(4540), 1577 LP – 1585.

Holsinger, L. J., Shaughnessy, M. A., Micko, A., Pinto, L. H., & Lamb, R. A. (1995). Analysis of the posttranslational modifications of the influenza virus M2 protein. *Journal of Virology*, **69**(2), 1219–25.

Holt, L. J., Tuch, B. B., Villen, J., Johnson, A. D., Gygi, S. P., & Morgan, D. O. (2009). Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**. doi:10.1126/science.1172867

Honnappa, S., Gouveia, S. M., Weisbrich, A., … Steinmetz, M. O. (2009). An EB1-Binding Motif Acts as a Microtubule Tip Localization Signal. *Cell*, **138**(2), 366–376.

Hrubey, T. W., & Roach, P. J. (1990). Phosphoserine in peptide substrates can specify casein kinase II action. *Biochemical and Biophysical Research Communications*, **172**(1), 190–196.

Hsiang, T.-Y., Zhou, L., & Krug, R. M. (2012). Roles of the Phosphorylation of Specific Serines and Threonines in the NS1 Protein of Human Influenza A Viruses. *Journal of Virology*, **86**(19), 10370 LP – 10376.

Huang, S., Chen, J., Chen, Q., … Chen, Z. (2013). A Second CRM1-Dependent Nuclear Export Signal in the Influenza A Virus NS2 Protein Contributes to the Nuclear Export of Viral Ribonucleoproteins. *Journal of Virology*, **87**(2), 767 LP – 778.

Hutchinson, E. C., Charles, P. D., Hester, S. S., … Fodor, E. (2014). Conserved and host-specific features of influenza virion architecture. *Nature Communications*, **5**, 4816.

Hutchinson, E. C., Denham, E. M., Thomas, B., … Fodor, E. (2012). Mapping the Phosphoproteome of Influenza A and B Viruses by Mass Spectrometry. *PLOS Pathogens*, **8**(11), e1002993.

Hütter, J., Rödig, J. V, Höper, D., Seeberger, P. H., & Hu, J. (2019). Toward Animal Cell Culture − Based Influenza Vaccine Design: Viral Hemagglutinin N- Glycosylation Markedly Impacts Immunogenicity. doi:10.4049/jimmunol.1201060

Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., & Dunker, A. K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology*, **323**(3), 573–584.

Igarashi, M., Ito, K., Kida, H., & Takada, A. (2008). Genetically destined potentials for N-linked glycosylation of influenza virus hemagglutinin. *Virology*, **376**. doi:10.1016/j.virol.2008.03.036

Iwatsuki-Horimoto, K., Horimoto, T., Fujii, Y., & Kawaoka, Y. (2004). Generation of Influenza A Virus NS2 (NEP) Mutants with an Altered Nuclear Export Signal Sequence. *Journal of Virology*, **78**(18), 10149 LP – 10155.

Jeff, C., Jean-Baptiste, M., & Hod, L. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society B: Biological Sciences*, **280**(1755), 20122863.

Kathum, O. A., Schräder, T., Anhlan, D., … Ludwig, S. (2016). Phosphorylation of influenza A virus NS1 protein at threonine 49 suppresses its interferon antagonistic activity. *Cellular Microbiology*, **18**(6), 784–791.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**(4), 772–780.

Kawakami, E., Watanabe, T., Fujii, K., … Kawaoka, Y. (2011). Strand-specific real-time RT-PCR for distinguishing influenza vRNA, cRNA, and mRNA. *Journal of Virological Methods*, **173**(1), 1–6.

Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerging Infectious Diseases*, **12**(1), 9–14.

Kim, I., Lee, H., Han, S. K., & Kim, S. (2014). Linear Motif-Mediated Interactions Have Contributed to the Evolution of Modularity in Complex Protein Interaction Networks. *PLOS Computational Biology*, **10**(10), e1003881.

Kim, J., Kim, I., Yang, J. J.-S., … Geiger, T. (2012). Rewiring of PDZ Domain-Ligand Interaction Network Contributed to Eukaryotic Evolution. *PLOS Genetics*, **8**(2), e1002510.

Kim, P., Jang, Y. H., Kwon, S. Bin, Lee, C. M., Han, G., & Seong, B. L. (2018). Glycosylation of Hemagglutinin and Neuraminidase of Influenza A Virus as Signature for Ecological Spillover and Adaptation among Influenza Reservoirs. *Viruses*, **10**(4), 183.

Kimura, M., & Ota, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **71**(7), 2848–2852.

Klemm, C., Boergeling, Y., Ludwig, S., & Ehrhardt, C. (2018). Immunomodulatory Nonstructural Proteins of Influenza A Viruses. *Trends in Microbiology*, **26**(7), 624–636.

Kodym, A., & Afza, R. (2003). Physical and chemical mutagenesis. *Methods in Molecular Biology (Clifton, N.J.)*, **236**(February 2003), 189–204.

Komar, A. A. (2016). The Yin and Yang of codon usage. *Human Molecular Genetics*, **25**(R2), R77–R85.

Komarnitsky, P., Cho, E., & Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription, (617), 2452–2460.

König, R., Stertz, S., Zhou, Y., … Chanda, S. K. (2009). Human host factors required for influenza virus replication. *Nature*, **463**, 813.

Koonin, E. V, & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life*, **61**(2), 99–111.

Korkmaz, G., Holm, M., Wiens, T., & Sanyal, S. (2014). Comprehensive Analysis of Stop Codon Usage in Bacteria and Its Correlation with Release Factor Abundance. *Journal of Biological Chemistry* , **289**(44), 30334–30342.

Kosugi, S., Hasebe, M., Matsumura, N., … Yanagawa, H. (2009). Six Classes of Nuclear Localization Signals Specific to Different Binding Grooves of Importin α. *Journal of Biological Chemistry* , **284**(1), 478–485.

Kramer, E. B., & Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA (New York, N.Y.)*, **13**(1), 87–96.

Krissinel, E., & Henrick, K. (2007). Inference of Macromolecular Assemblies from Crystalline State.

Journal of Molecular Biology, **372**(3), 774–797.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*, **305**(3), 567–580.

Krug, R. M. (2015). Functions of the influenza A virus NS1 protein in antiviral defense. *Current Opinion in Virology*, **12**, 1–6.

Kucharavy, A., Rubinstein, B., Zhu, J., & Li, R. (2018). Robustness and evolvability of heterogeneous cell populations. *Molecular Biology of the Cell*, **29**(11), 1400–1409.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli *Science*, **324**(5924), 255 LP – 258.

Kumar, A., Manatschal, C., Rai, A., … Steinmetz, M. O. (2017). Short Linear Sequence Motif LxxPTPh Targets Diverse Proteins to Growing Microtubule Ends. *Structure*, **25**(6), 924-932.e4.

Kumar, N., Bera, B. C., Greenbaum, B. D., … Virmani, N. (2016). Revelation of Influencing Factors in Overall Codon Usage Bias of Equine Influenza Viruses. *PLOS ONE*, **11**(4), e0154376.

Kummer, S., Flöttmann, M., Schwanhäusser, B., … Herrmann, A. (2014). Alteration of Protein Levels during Influenza Virus H1N1 Infection in Host Cells: A Proteomic Survey of Host and Virus Reveals Differential Dynamics. *PLOS ONE*, **9**(4), e94257.

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., & Rokas, A. (2019). Variation and selection on codon usage bias across an entire subphylum. *PLOS Genetics*, **15**(7), e1008304.

Lam, H. C., Bi, X., Sreevatsan, S., & Boley, D. (2017). Evolution and Vaccination of Influenza Virus. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, **24**(8), 787–798.

Lamb, R. A., & Lai, C.-J. (1980). Sequence of interrupted and uninterrupted mRNAs and cloned DNA coding for the two overlapping nonstructural proteins of influenza virus. *Cell*, **21**(2), 475–485.

Lande, R. (1976). Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution*, **30**(2), 314.

Landry, C. R., Levy, E. D., Abd Rabbo, D., Tarassov, K., & Michnick, S. W. (2013). Extracting Insight from Noisy Cellular Networks. *Cell*, **155**(5), 983–989.

Langan, R. A., Boyken, S. E., Ng, A. H., … Baker, D. (2019). De novo design of bioactive protein switches. *Nature*, **572**(7768), 205–210.

Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E., & Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin alpha. *The Journal of Biological Chemistry*, **282**(8), 5101–5105.

Larkin, M. A., Blackshields, G., Brown, N. P., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–2948.

Larter, M., Dunbar-Wallis, A., Berardi, A. E., & Smith, S. D. (2018). Convergent Evolution at the Pathway Level: Predictable Regulatory Changes during Flower Color Transitions. *Molecular Biology and Evolution*, **35**(9), 2159–2169.

Lässig, M., Mustonen, V., & Walczak, A. M. (2017). Predicting evolution. *Nature Ecology & Evolution*, **1**(3), 77.

Latysheva, N. S., Flock, T., Weatheritt, R. J., Chavali, S., & Babu, M. M. (2015). How do disordered regions achieve comparable functions to structured domains? *Protein Science : A Publication of the Protein Society*, **24**(6), 909–922.

Latysheva, N. S., Oates, M. E., Maddox, L., … Babu, M. M. (2016). Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer. *Molecular Cell*, **63**(4), 579–592.

Lauring, A. S., Acevedo, A., Cooper, S. B., & Andino, R. (2012). Codon usage determines the mutational robustness, evolutionary capacity, and virulence of an RNA virus. *Cell Host and Microbe*, **12**(5), 623–632.

Lavner, Y., & Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, **345**(1), 127–138.

Le Nouën, C., Collins, P. L., & Buchholz, U. J. (2019). Attenuation of Human Respiratory Viruses by Synonymous Genome Recoding. *Frontiers in Immunology*, **10**, 1250.

Levy, E. D., Landry, C. R., & Michnick, S. W. (2009). How perfect can protein interactomes be? *Science Signaling*, **2**. Retrieved from http://dx.doi.org/10.1126/scisignal.260pe11

Lewnard, J. A., & Cobey, S. (2018). Immune History and Influenza Vaccine Effectiveness, 1–14.

Li, D., Fedeles, B. I., Singh, V., … Essigmann, J. M. (2014a). Tautomerism provides a molecular explanation for the mutagenic properties of the anti-HIV nucleoside 5-aza-5,6-dihydro-2′-deoxycytidine. *Proceedings of the National Academy of Sciences*, **111**(32), E3252 LP-E3259.

Li, J., Yu, M., Zheng, W., & Liu, W. (2015). Nucleocytoplasmic shuttling of influenza A virus proteins. *Viruses*, **7**(5), 2668–2682.

Li, X.-H., Chavali, P. L., Pancsa, R., Chavali, S., & Babu, M. M. (2018a). Function and Regulation of Phase-Separated Biological Condensates. *Biochemistry*, **57**(17), 2452–2461.

Li, Y., Sahni, N., Pancsa, R., … Yi, S. (2017). Revealing the Determinants of Widespread Alternative Splicing Perturbation in Cancer. *Cell Reports*, **21**(3), 798–812.

Li, Y., Sun, L., Zheng, W., … Luo, T. R. (2018b). Phosphorylation and dephosphorylation of threonine 188 in nucleoprotein is crucial for the replication of influenza A virus. *Virology*, **520**, 30–38.

Li, Y., Yamakita, Y., & Krug, R. M. (1998). Regulation of a nuclear export signal by an adjacent inhibitory sequence: The effector domain of the influenza virus NS1 protein. *Proceedings of the National Academy of Sciences*, **95**(9), 4864 LP – 4869.

Li, Z., Wang, Y., Yao, Q., … Pan, C. (2014b). Diverse and divergent protein post-translational modifications in two growth stages of a natural microbial community. *Nature Communications*, **5**, 4405.

Lin, D., Lan, J., & Zhang, Z. (2007). Structure and Function of the NS1 Protein of Influenza A Virus N-terminus of the NS1A Protein, **39**(3), 155–162.

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., & Gerlinger, M. (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, **2**(1), 49–63.

Liu, W.-C., Jan, J.-T., Huang, Y.-J., Chen, T.-H., & Wu, S.-C. (2016). Unmasking Stem-Specific Neutralizing Epitopes by Abolishing N-Linked Glycosylation Sites of Influenza Virus Hemagglutinin Proteins for Vaccine Design. *Journal of Virology*, **90**(19), 8496 LP – 8508.

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**(6648), 251–260.

Luoh, S., Mcgregor, M. W., & Hinshaw, V. S. (1992). Hemagglutinin Mutations Related to Antigenic Variation in Hi Swine Influenza Viruses, **66**(2), 1066–1073.

Luria, S. E., & Delbrück, M. (1943). MUTATIONS OF BACTERIA FROM VIRUS SENSITIVITY TO VIRUS RESISTANCE. *Genetics*, **28**(6), 491 LP – 511.

Lyons, D. M., & Lauring, A. S. (2018). Mutation and Epistasis in Influenza Virus Evolution. *Viruses*, **10**(8), 407.

Marin, O., Bustos, V. H., Cesaro, L., … Allende, J. E. (2003). A noncanonical sequence phosphorylated by casein kinase 1 in β-catenin may play a role in casein kinase 1 targeting of important signaling proteins. *Proceedings of the National Academy of Sciences*, **100**(18), 10193 LP – 10200.

Martinez, E., Schroeder, G. N., Berger, C. N., … Frankel, G. (2010). Binding to Na+/H+ exchanger regulatory factor 2 (NHERF2) affects trafficking and function of the enteropathogenic Escherichia coli type III secretion system effectors Map, EspI and NleH. *Cellular Microbiology*, **12**(12), 1718–1731.

Maynard Smith, J. (1974). The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, **47**(1), 209–221.

McAuley, J. L., Gilbertson, B. P., Trifkovic, S., Brown, L. E., McKimm-Breschkin, J. L., & McAuley, J. L. (2019). Influenza Virus Neuraminidase Structure and Functions. *Frontiers in Microbiology*, **10**(January), 39.

McCandlish, D. M., & Stoltzfus, A. (2014). Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology*, **89**(3), 225–252.

Mehle, A., Dugan, V. G., Taubenberger, J. K., & Doudna, J. A. (2012). Reassortment and Mutation of the Avian Influenza Virus Polymerase PA Subunit Overcome Species Barriers. *Journal of Virology*, **86**(3), 1750 LP – 1757.

Meineke, R., & Rimmelzwaan, G. F. (2019). Influenza Virus Infections and Cellular Kinases. doi:10.3390/v11020171

Melén, K., Kinnunen, L., Fagerlund, R., … Julkunen, I. (2007). Nuclear and Nucleolar Targeting of Influenza A Virus NS1 Protein: Striking Differences between Different Virus Subtypes. *Journal of Virology*, **81**(11), 5995 LP – 6006.

Melén, K., Tynell, J., Fagerlund, R., Roussel, P., Hernandez-Verdun, D., & Julkunen, I. (2012). Influenza A H3N2 subtype virus NS1 protein targets into the nucleus and binds primarily via its C-terminal NLS2/NoLS to nucleolin and fibrillarin. *Virology Journal*, **9**(1), 167.

Miller, S., Janin, J., Lesk, A. M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of Molecular Biology*, **196**(3), 641–656.

Min, J.-Y., Li, S., Sen, G. C., & Krug, R. M. (2007). A site on the influenza A virus NS1 protein mediates both inhibition of PKR activation and temporal regulation of viral RNA synthesis. *Virology*, **363**(1), 236–243.

Mittag, T., Marsh, J., Grishaev, A., … Forman-Kay, J. D. (2010). Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure*, **18**(4), 494–506.

Mondal, A., Dawson, A. R., Potts, G. K., … Mehle, A. (2017). Influenza virus recruits host protein kinase C to control assembly and activity of its replication machinery. *ELife*, **6**, e26910.

Mondal, A., Potts, G. K., Dawson, A. R., Coon, J. J., & Mehle, A. (2015). Phosphorylation at the Homotypic Interface Regulates Nucleoprotein Oligomerization and Assembly of the Influenza Virus Replication Machinery. *PLOS Pathogens*, **11**(4), e1004826.

Montville, R., Froissart, R., Remold, S. K., Tenaillon, O., & Turner, P. E. (2005). Evolution of Mutational Robustness in an RNA Virus. *PLOS Biology*, **3**(11), e381.

Moratorio, G., Henningsson, R., Barbezange, C., … Vignuzzi, M. (2017). Attenuation of RNA viruses by redirecting their evolution in sequence space. *Nature Microbiology*, **2**(June). doi:10.1038/nmicrobiol.2017.88

Moses, A. M., Liku, M. E., Li, J. J., & Durbin, R. (2007). Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. *Proc Natl Acad Sci U S A*, **104**. doi:10.1073/pnas.0700997104

Moya, A., Elena, S. F., Bracho, A., Miralles, R., & Barrio, E. (2000). The evolution of RNA viruses: A population genetics view. *Proceedings of the National Academy of Sciences*, **97**(13), 6967 LP – 6973.

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, **156**(1), 297 LP – 304.

Neduva, V., & Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Letters*, **579**(15), 3342–3345.

Neumann, G., Castrucci, M. R., & Kawaoka, Y. (1997). Nuclear import and export of influenza virus nucleoprotein. *Journal of Virology*, **71**(12), 9690 LP – 9700.

Neurath, H., Walsh, K. A., & Winter, W. P. (1967). Evolution of Structure and Function of Proteases. *Science*, **158**(3809), 1638 LP – 1644.

Nilsson, T., Jackson, M., & Peterson, P. A. (1989). Short cytoplasmic sequences serve as retention signals for transmembrane proteins in the endoplasmic reticulum. *Cell*, **58**(4), 707–718.

Nobusawa, E., & Sato, K. (2006). Comparison of the Mutation Rates of Human Influenza A and B Viruses. *Journal of Virology*, **80**(7), 3675 LP – 3678.

Noda, T., Sagara, H., Yen, A., … Kawaoka, Y. (2006). Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature*, **439**(7075), 490–492.

Nogales, A., Martinez-Sobrido, L., Topham, J. D., & DeDiego, L. M. (2018). Modulation of Innate Immune Responses by the Influenza A NS1 and PA-X Proteins. *Viruses* . doi:10.3390/v10120708

O'Neill, R. E., Talon, J., & Palese, P. (1998). The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. *The EMBO Journal*, **17**(1), 288–296.

Oates, M. E., Romero, P., Ishida, T., … Gough, J. (2012). D2P2: database of disordered protein predictions. *Nucleic Acids Research*, **41**(D1), D508–D516.

Olson-Manning, C. F., Wagner, M. R., & Mitchell-Olds, T. (2012). Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews. Genetics*, **13**(12), 867–877.

Oscamou, M., McDonald, D., Bing, V. B., Huttley, G. A., Lladser, M. E., & Knight, R. (2008). Comparison of methods for estimating the nucleotide substitution matrix. *BMC Bioinformatics*, **9**. doi:10.1186/1471-2105-9-511

Pancsa, R., & Tompa, P. (2012). Structural Disorder in Eukaryotes. *PLOS ONE*, **7**(4), e34687.

Paterson, D., & Fodor, E. (2012). Emerging Roles for the Influenza A Virus Nuclear Export Protein (NEP). *PLOS Pathogens*, **8**(12), e1003019.

Patwa, Z., & Wahl, L. M. (2008). The fixation probability of beneficial mutations. *Journal of the Royal Society, Interface*, **5**(28), 1279–1289.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, **37**(4), 205 LP – 211.

Pauly, M. D., Procario, M. C., & Lauring, A. S. (2017). A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *ELife*, **6**, e26437.

Pawson, T., Olivier, P., Rozakis-Adcock, M., McGlade, J., & Henkemeyer, M. (1993). Proteins with SH2 and SH3 Domains Couple Receptor Tyrosine Kinases to Intracellular Signalling Pathways. *Philosophical Transactions: Biological Sciences*, **340**(1293), 279–285.

Peng, Q., Zhu, R., Wang, X., … Liu, X. (2019). Impact of the variations in potential glycosylation sites of the hemagglutinin of H9N2 influenza virus. *Virus Genes*, **55**(2), 182–190.

Perczel, A., Hudáky, P., & Pálfi, V. K. (2007). Dead-End Street of Protein Folding: Thermodynamic Rationale of Amyloid Fibril Formation. *Journal of the American Chemical Society*, **129**(48), 14959–14965.

Petljak, M., Alexandrov, L. B., Brammeld, J. S., Nik-zainal, S., Campbell, P. J., & Stratton, M. R. (2019). Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Article Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis, 1282–1294.

Petri, T., Patterson, S., & Dimmock, N. J. (1982). Polymorphism of the NS1 Proteins of Type A Influenza Virus. *Journal of General Virology*, **61**(2), 217–231.

Petrova, V. N., & Russell, C. A. (2017). The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, **16**, 47.

Pielak, R. M., & Chou, J. J. (2011). Influenza M2 proton channels. *Biochimica et Biophysica Acta*, **1808**(2), 522–529.

Plotkin, J. B., & Dushoff, J. (2003). Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proceedings of the National Academy of Sciences*, **100**(12), 7152 LP – 7157.

Pohl, M. O., Lanz, C., & Stertz, S. (2016). Late stages of the influenza A virus replication cycle — a tight interplay between virus and host, 2058–2072.

Pope, C. F., O'Sullivan, D. M., McHugh, T. D., & Gillespie, S. H. (2008). A Practical Guide to Measuring Mutation Rates in Antibiotic Resistance. *Antimicrobial Agents and Chemotherapy*, **52**(4), 1209 LP – 1214.

Portela, A., & Digard, P. (2002). The influenza virus nucleoprotein : a multifunctional RNA-binding protein pivotal to virus replication, 723–734.

Potter, C. W. (2001). A history of influenza. *Journal of Applied Microbiology*, **91**(4), 572–579.

Pushker, R., Mooney, C., Davey, N. E., Jacqué, J.-M., & Shields, D. C. (2013). Marked Variability in the Extent of Protein Disorder within and between Viral Families. *PLOS ONE*, **8**(4), e60724.

Rajabian, T., Gavicherla, B., Heisig, M., … Ireton, K. (2009). The bacterial virulence factor InlC perturbs apical cell junctions and promotes cell-to-cell spread of Listeria. *Nature Cell Biology*, **11**, 1212.

Ravarani, C. N. J., Erkina, T. Y., De Baets, G., Dudman, D. C., Erkine, A. M., & Babu, M. M. (2018). High-throughput discovery of functional disordered regions: investigation of transactivation domains. *Molecular Systems Biology*, **14**(5), e8190.

Ravi Chandra, B., Gowthaman, R., Raj Akhouri, R., Gupta, D., & Sharma, A. (2004). Distribution of proline-rich (PxxP) motifs in distinct proteomes: functional and therapeutic implications for malaria and tuberculosis. *Protein Engineering, Design and Selection*, **17**(2), 175–182.

Reuther, P., Giese, S., Götz, V., Riegger, D., & Schwemmle, M. (2014). Phosphorylation of highly conserved serine residues in the influenza A virus nuclear export protein NEP plays a minor role

in viral growth in human cells and mice. *Journal of Virology*, **88**(13), 7668–7673.

Richard, A., & Tulasne, D. (2012). Caspase cleavage of viral proteins, another way for viruses to make the best of apoptosis. *Cell Death & Disease*, **3**(3), e277–e277.

Romero, P. A., & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews. Molecular Cell Biology*, **10**(12), 866–876.

Rossman, J. S., & Lamb, R. A. (2011). Influenza virus assembly and budding. *Virology*, **411**(2), 229–236.

Rüdiger, D., Kupke, S. Y., Laske, T., Zmora, P., & Reichl, U. (2019). Multiscale modeling of influenza A virus replication in cell cultures predicts infection dynamics for highly different infection conditions. *PLOS Computational Biology*, **15**(2), e1006819.

Russell, A. B., Trapnell, C., & Bloom, J. D. (2018). Extreme heterogeneity of influenza virus infection in single cells. *ELife*, **7**, e32303.

Rust, H. L., & Thompson, P. R. (2011). Kinase consensus sequences: a breeding ground for crosstalk. *ACS Chemical Biology*, **6**(9), 881–892.

Sagulenko, P., Puller, V., & Neher, R. A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution*, **4**(1), vex042–vex042.

Sakaguchi, K., Saito, S., Higashimoto, Y., Roy, S., Anderson, C. W., & Appella, E. (2000). Damage-mediated Phosphorylation of Human p53 Threonine 18 through a Cascade Mediated by a Casein 1-like Kinase: EFFECT ON Mdm2 BINDING . *Journal of Biological Chemistry* , **275**(13), 9278–9283.

Saksela, K., & Permi, P. (2012). SH3 domain ligand binding: What's the consensus and where's the specificity? *FEBS Letters*, **586**(17), 2609–2614.

Sambataro, F., & Pennuto, M. (2017). Post-translational Modifications and Protein Quality Control in Motor Neuron and Polyglutamine Diseases. *Frontiers in Molecular Neuroscience*, **10**, 82.

Samji, T. (2009). Influenza A: understanding the viral life cycle. *The Yale Journal of Biology and Medicine*, **82**(4), 153–159.

Sanger, F., & Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, **49**(4), 463 LP – 481.

Sanjuán, R. (2010). Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**(1548), 1975–1982.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M., & Belshaw, R. (2010). Viral Mutation Rates. *Journal of Virology*, **84**(19), 9733 LP – 9748.

Santos, A., Pal, S., Chacón, J., … Rosas-Acosta, G. (2013). SUMOylation Affects the Interferon Blocking Activity of the Influenza A Nonstructural Protein NS1 without Affecting Its Stability or Cellular Localization. *Journal of Virology*, **87**(10), 5602 LP – 5620.

Schad, E., Tompa, P., & Hegyi, H. (2011). The relationship between proteome size, structural disorder and organism complexity. *Genome Biol*, **12**. doi:10.1186/gb-2011-12-12-r120

Schlessinger, J. (1994). SH2/SH3 signaling proteins. *Current Opinion in Genetics & Development*, **4**(1), 25–30.

Schubert, U., Antón, L. C., Gibbs, J., Norbury, C. C., Yewdell, J. W., & Bennink, J. R. (2000). Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*, **404**(6779),

770–774.

Schwarz, F., & Aebi, M. (2011). Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology*, **21**(5), 576–582.

Sen, N., Sen, A., & Mackow, E. R. (2007). Degrons at the C Terminus of the Pathogenic but Not the Nonpathogenic Hantavirus G1 Tail Direct Proteasomal Degradation. *Journal of Virology*, **81**(8), 4323 LP – 4330.

Sever, R., & Brugge, J. S. (2015). Signal transduction in cancer. *Cold Spring Harbor Perspectives in Medicine*, **5**(4), a006098.

Shan, B., Li, D.-W., Brüschweiler-Li, L., & Brüschweiler, R. (2012). Competitive Binding between Dynamic p53 Transactivation Subdomains to Human MDM2 Protein: IMPLICATIONS FOR REGULATING THE p53·MDM2/MDMX INTERACTION . *Journal of Biological Chemistry* , **287**(36), 30376–30384.

Shao, W., Li, X., Goraya, M. U., Wang, S., & Chen, J.-L. (2017). Evolution of Influenza A Virus by Mutation and Re-Assortment. *International Journal of Molecular Sciences*, **18**(8), 1650.

Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**(1544), 1203–1212.

Shaulsky, G., Goldfinger, N., Ben-Ze'ev, A., & Rotter, V. (1990). Nuclear accumulation of p53 protein is mediated by several nuclear localization signals and plays a role in tumorigenesis. *Molecular and Cellular Biology*, **10**(12), 6565–6577.

Sheerar, M. G., Easterday, B. C., & Hinshaw, V. S. (1989). Antigenic Conservation of H1N1 Swine Influenza Viruses. *Journal of General Virology*, **70**(12), 3297–3303.

Shilova, L. A., Lyushnyak, A. S., Knyazev, D. G., Fedorova, N. V, Baratova, L. A., & Batishchev, O. V. (2017). Assembly of Matrix Protein 1 of Influenza a Virus and its Role in Budding Process. *Biophysical Journal*, **112**(3), 390a.

Shin, Y.-K., Li, Y., Liu, Q., Anderson, D. H., Babiuk, L. A., & Zhou, Y. (2007a). SH3 Binding Motif 1 in Influenza A Virus NS1 Protein Is Essential for PI3K/Akt Signaling Pathway Activation. *Journal of Virology*, **81**(23), 12730 LP – 12739.

Shin, Y.-K., Liu, Q., Tikoo, S. K., Babiuk, L. A., & Zhou, Y. (2007b). Influenza A virus NS1 protein activates the phosphatidylinositol 3-kinase (PI3K)/Akt pathway by direct interaction with the p85 subunit of PI3K. *Journal of General Virology*, **88**(1), 13–18.

Shtykova, E. V, Baratova, L. A., Fedorova, N. V, … Svergun, D. I. (2013). Structural Analysis of Influenza A Virus Matrix Protein M1 and Its Self-Assemblies at Low pH. *PLOS ONE*, **8**(12), e82431.

Sierra, T., Arago, T., Sanz-ezquerro, J. J., … Nieto, A. (1998). The PA influenza virus polymerase subunit is a phosphorylated protein, 471–478.

Singh, V. (2014). Recent advances and opportunities in synthetic logic gates engineering in living cells. *Systems and Synthetic Biology*, **8**(4), 271–282.

Sironen, T., Kallio, E. R., & Vaheri, A. (2008). Communication Quasispecies dynamics and fixation of a synonymous mutation in hantavirus transmission, 1309–1313.

Smith, J. M., & Price, G. R. (1973). The Logic of Animal Conflict. *Nature*, **246**(5427), 15–18.

Spielman, S. J., & Wilke, C. O. (2015). Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. *PLOS ONE*, **10**(9), e0139047.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Stegmann, T. (2000). Membrane Fusion Mechanisms : The Influenza Hemagglutinin Paradigm and its Implications for Intracellular Fusion, (2), 598–604.

Stein, A., & Aloy, P. (2010). Novel Peptide-Mediated Interactions Derived from High-Resolution 3-Dimensional Structures. *PLOS Computational Biology*, **6**(5), e1000789.

Stoltzfus, A., & McCandlish, D. M. (2017). Mutational Biases Influence Parallel Adaptation. *Molecular Biology and Evolution*, **34**(9), 2163–2172.

Stommel, J. M., Marchenko, N. D., Jimenez, G. S., Moll, U. M., Hope, T. J., & Wahl, G. M. (1999). A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *The EMBO Journal*, **18**(6), 1660–1672.

Sukumaran, J., & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**(12), 1569–1571.

Sun, E., He, J., & Zhuang, X. (2013). Dissecting the Role of COPI Complexes in Influenza Virus Infection. *Journal of Virology*, **87**(5), 2673 LP – 2685.

Sun, S., Wang, Q., Zhao, F., Chen, W., & Li, Z. (2011). Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. *PloS One*, **6**(7), e22844–e22844.

Tarendeau, F., Boudet, J., Guilligay, D., … Hart, D. J. (2007). Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nature Structural &Amp; Molecular Biology*, **14**, 229.

Tate, M. D., Job, E. R., Deng, Y.-M., Gunalan, V., Maurer-Stroh, S., & Reading, P. C. (2014). Playing hide and seek: how glycosylation of the influenza virus hemagglutinin can modulate the immune response to infection. *Viruses*, **6**(3), 1294–1316.

Taubenberger, J. K., & Morens, D. M. (2010). Influenza: the once and future pandemic. *Public Health Reports (Washington, D.C. : 1974)*, **125 Suppl**(Suppl 3), 16–26.

Tavassoly, I., Goldfarb, J., & Iyengar, R. (2018). Systems biology primer: the basic methods and approaches. *Essays In Biochemistry*, EBC20180003.

te Velthuis, A. J. W., & Fodor, E. (2016). Influenza virus RNA polymerase: insights into the mechanisms of viral RNA synthesis. *Nature Reviews Microbiology*, **14**, 479.

Thomas, J. M., Stevens, M. P., Percy, N., & Barclay, W. S. (1998). Phosphorylation of the M2 protein of influenza A virus is not essential for virus viability. *Virology*, **252**(1), 54–64.

Thomas, M., Kranjec, C., Nagasaka, K., Matlashewski, G., & Banks, L. (2011). Analysis of the PDZ binding specificities of Influenza A Virus NS1 proteins. *Virology Journal*, **8**(1), 25.

Thyagarajan, B., & Bloom, J. D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *ELife*, **3**, e03300.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE*, **8**(11), e80635.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., & Tawfik, D. S. (2009). Do viral proteins possess unique biophysical features? *Trends in Biochemical Sciences*, **34**(2), 53–59.

Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*, **37**. doi:10.1016/j.tibs.2012.08.004

Tompa, P., Davey, N. E. E., Gibson, T. J. J., & Babu, M. M. M. (2014). A million peptide motifs for the molecular biologist. *Molecular Cell*, **55**(2), 161–169.

Tompa, P., & Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences*, **33**(1), 2–8.

Topal, M. D., & Fresco, J. R. (1976). Complementary base pairing and the origin of substitution mutations. *Nature*, **263**(5575), 285–289.

Tóth-Petróczy, Á., & Tawfik, D. S. (2014). The robustness and innovability of protein folds. *Current Opinion in Structural Biology*, **26**, 131–138.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., & Li, Z. (2002). Effect of addition of new oligosaccharide chains to the globular head of influenza A / H2N2 virus haemagglutinin on the intracellular transport and biological activities of the molecule, 1137–1146.

Turrell, L., Hutchinson, E. C., Vreede, F. T., & Fodor, E. (2015). Regulation of Influenza A Virus Nucleoprotein Oligomerization by Phosphorylation. *Journal of Virology*, **89**(2), 1452 LP – 1455.

Tynell, J., Melén, K., & Julkunen, I. (2014). Mutations within the conserved NS1 nuclear export signal lead to inhibition of influenza A virus replication. *Virology Journal*, **11**, 128.

Ubersax, J. A., & Ferrell, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, **8**. doi:10.1038/nrm2203

Uversky, V. N. (2015). Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders . *Frontiers in Aging Neuroscience* , p. 18.

Vahey, M. D., & Fletcher, D. A. (2019). Low-Fidelity Assembly of Influenza A Virus Promotes Escape from Host Cells. *Cell*, **176**(1), 281-294.e19.

van der Laan, J. M., Swarte, M. B. A., Groendijk, H., Hol, W. G. J., & Drenth, J. (1989). The influence of purification and protein heterogeneity on the crystallization of p-hydroxybenzoate hydroxylase. *European Journal of Biochemistry*, **179**(3), 715–724.

van der Lee, R., Buljan, M., Lang, B., … Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, **114**(13), 6589–6631.

Van Roey, K., & Davey, N. E. (2015). Motif co-regulation and co-operativity are common mechanisms in transcriptional, post-transcriptional and post-translational regulation. *Cell Communication and Signaling*, **13**(1), 45.

Van Roey, K., Gibson, T. J., Davey, N. E., Van Roey, K., Gibson, T. J., & Davey, N. E. (2012). Motif switches: decision-making in cell regulation. *Current Opinion in Structural Biology*, **22**(3), 378–385.

Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., & Budd, A. (2014). Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev*, **114**. doi:10.1021/cr400585q

Varghese, J. N., Laver, W. G., & Colman, P. M. (1983). Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature*, **303**(5912), 35–40.

Vaseva, A. V, & Moll, U. M. (2009). The mitochondrial p53 pathway. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, **1787**(5), 414–420.

Venkatakrishnan, A. J., Flock, T., Prado, D. E., Oates, M. E., Gough, J., & Madan Babu, M. (2014). Structured and disordered facets of the GPCR fold. *Current Opinion in Structural Biology*, **27**, 129–137.

Via, A., Gould, C. M., Gemünd, C., Gibson, T. J., & Helmer-Citterich, M. (2009). A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*, **10**, 351.

Via, A., Uyar, B., Brun, C., & Zanzoni, A. (2015). How pathogens use linear motifs to perturb host

cell networks. *Trends in Biochemical Sciences*, **40**(1), 36–48.

Vigerust, D. J., & Shepherd, V. L. (2007). Virus glycosylation: role in virulence and immune interactions. *Trends in Microbiology*, **15**(5), 211–218.

Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E., & Andino, R. (2006). Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, **439**(7074), 344–348.

Vu, T.-K. H., Hung, D. T., Wheaton, V. I., & Coughlin, S. R. (1991). Molecular cloning of a functional thrombin receptor reveals a novel proteolytic mechanism of receptor activation. *Cell*, **64**(6), 1057–1068.

Wagner, A. (2012). The role of robustness in phenotypic adaptation and innovation. *Proceedings. Biological Sciences*, **279**(1732), 1249–1258.

Wang, S., Zhao, Z., Bi, Y., Sun, L., Liu, X., & Liu, W. (2013). Tyrosine 132 Phosphorylation of Influenza A Virus M1 Protein Is Crucial for Virus Replication by Controlling the Nuclear Import of M1. *Journal of Virology*, **87**(11), 6182 LP – 6191.

Wang, Y., Cheng, H., Pan, Z., Ren, J., Liu, Z., & Xue, Y. (2015). Reconfiguring phosphorylation signaling by genetic polymorphisms affects cancer susceptibility. *Journal of Molecular Cell Biology*, **7**(3), 187–202.

Weatheritt, R. J., & Babu, M. M. (2013). The Hidden Codes That Shape Protein Evolution. *Science*, **342**(6164), 1325 LP – 1326.

Whittington, A. C., Mason, A. J., & Rokyta, D. R. (2018). A Single Mutation Unlocks Cascading Exaptations in the Origin of a Potent Pitviper Neurotoxin. *Molecular Biology and Evolution*, **35**(4), 887–898.

Wilke, C. O. (2003). Probability of fixation of an advantageous mutant in a viral quasispecies. *Genetics*, **163**(2), 467–474.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., & Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, **412**(6844), 331–333.

Wong, E. H. M., Smith, D. K., Rabadan, R., Peiris, M., & Poon, L. L. M. (2010). Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evolutionary Biology*, **10**(1), 253.

Worobey, M., Han, G.-Z., & Rambaut, A. (2014). A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, **508**, 254.

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, **293**(2), 321–331.

Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, **16**. doi:10.1038/nrm3920

Wu, C.-Y., Jeng, K.-S., & Lai, M. M.-C. (2011). The SUMOylation of matrix protein M1 modulates the assembly and morphogenesis of influenza A virus. *Journal of Virology*, **85**(13), 6618–6628.

Wu, C.-Y., Lin, C.-W., Tsai, T.-I., … Wong, C.-H. (2017a). Influenza A surface glycosylation and vaccine design. *Proceedings of the National Academy of Sciences*, **114**(2), 280 LP – 285.

Wu, N. C., Zost, S. J., Thompson, A. J., … Wilson, I. A. (2017b). A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. *PLOS Pathogens*, **13**(10), e1006682.

Wu, W. W. H., Sun, Y.-H. B., & Panté, N. (2007). Nuclear import of influenza A viral ribonucleoprotein complexes is mediated by two nuclear localization sequences on viral

nucleoprotein. *Virology Journal*, **4**(1), 49.

Xu, R., Ekiert, D. C., Krause, J. C., Hai, R., Crowe, J. E., & Wilson, I. A. (2010). Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza Virus. *Science*, **328**(5976), 357 LP – 360.

Xue, B., Mizianty, M. J., Kurgan, L., & Uversky, V. N. (2012). Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cellular and Molecular Life Sciences*, **69**(8), 1211–1259.

Yampolsky, L. Y., & Stoltzfus, A. (2001). Bias in the introduction of variation as an orienting factor in evolution. *Evolution & Development*, **3**(2), 73–83.

Yan, A., & Lennarz, W. J. (2005). Unraveling the Mechanism of Protein N-Glycosylation. *Journal of Biological Chemistry* , **280**(5), 3121–3124.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution: Journal of Molecular Evolution. *J Mol Evol*, **39**, 105–111.

Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*, Oxford University Press.

Yang, Z., & Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution*, **17**(1), 32–43.

Ylösmäki, L., Schmotz, C., Ylösmäki, E., & Saksela, K. (2015). Reorganization of the host cell Crk(L)–PI3 kinase signaling complex by the influenza A virus NS1 protein. *Virology*, **484**, 146–152.

York, I. A., Stevens, J., & Alymova, I. V. (2019). Influenza virus N-linked glycosylation and innate immunity. *Bioscience Reports*, **39**(1), BSR20171505.

Yu, M., Liu, X., Cao, S., … Liu, W. (2012). Identification and Characterization of Three Novel Nuclear Export Signals in the Influenza A Virus Nucleoprotein. *Journal of Virology*, **86**(9), 4970 LP – 4980.

Zhang, Y., Dasgupta, J., Ma, R. Z., Banks, L., Thomas, M., & Chen, X. S. (2007). Structures of a Human Papillomavirus (HPV) E6 Polypeptide Bound to MAGUK Proteins: Mechanisms of Targeting Tumor Suppressors by a High-Risk HPV Oncoprotein. *Journal of Virology*, **81**(7), 3618 LP – 3626.

Zhang, Y., & Xiong, Y. (2001). A p53 Amino-Terminal Nuclear Export Signal Inhibited by DNA Damage-Induced Phosphorylation. *Science*, **292**(5523), 1910 LP – 1915.

Zhao, J., Kardashliev, T., Joëlle Ruff, A., Bocola, M., & Schwaneberg, U. (2014). Lessons from diversity of directed evolution experiments by an analysis of 3,000 mutations. *Biotechnology and Bioengineering*, **111**(12), 2380–2389.

Zhao, M., Wang, L., & Li, S. (2017). Influenza A Virus-Host Protein Interactions Control Viral Pathogenesis. *International Journal of Molecular Sciences*, **18**(8), 1673.

Zheng, W., Cao, S., Chen, C., … Liu, W. (2017). Threonine 80 phosphorylation of non-structural protein 1 regulates the replication of influenza A virus by reducing the binding affinity with RIG-I. *Cellular Microbiology*, **19**(2). doi:10.1111/cmi.12643

Zheng, W., Li, J., Wang, S., … Liu, W. (2015). Phosphorylation Controls the Nuclear-Cytoplasmic Shuttling of Influenza A Virus Nucleoprotein. *Journal of Virology*, **89**(11), 5822 LP – 5834.

Zhirnov, O. P., & Syrtzev, V. V. (2009). Influenza virus pathogenicity is determined by caspase cleavage motifs located in the viral proteins. *Journal of Molecular and Genetic Medicine : An International Journal of Biomedical Research*, **3**(1), 124–132.

Zhou, J., Tien, A.-C., Alberta, J. A., … Stiles, C. D. (2017). A Sequentially Priming Phosphorylation

Cascade Activates the Gliomagenic Transcription Factor Olig2. *Cell Reports*, **18**(13), 3167–3177.

Zost, S. J., Parkhouse, K., Gumina, M. E., … Hensley, S. E. (2017). Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences*, **114**(47), 12578 LP – 12583.