**T.C.**
**İSTANBUL UNIVERSITY**
**INSTITUTE OF GRADUATE STUDIES IN SCIENCE AND ENGINEERING**

# M.Sc. THESIS

COMPARATIVE DIVERSITY DYNAMICS OF INFLUENZA A VIRUS SUBTYPES

**Rashid MUKAILA**

**Department of Informatics**

**Informatics Programme**

**SUPERVISOR**
**Assoc. Prof. Dr. Çiğdem EROL**

**CO-SUPERVISOR**
**Assoc. Prof. Dr. Asif KHAN**

Tarih girmek için burayı tıklatın.

**İSTANBUL**

This study was accepted on 21/12/2022 as a M. Sc. thesis in Department of Informatics, Informatics programme by the following Committee:

**Examining Committee Members**

Assoc. Prof. Dr. Çiğdem Erol (Supervisor)
İstanbul University
Faculty of Science

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul University
Faculty of Science

Assist. Prof. Yalçın ÖZKAN
Istinye University
Faculty of Economics, Administrative and Social Sciences

As required by the 9/2 and 22/2 articles of the Graduate Education Regulation which was published in the Official Gazette on 20.04.2016, this graduate thesis is reported as in accordance with criteria determined by the Institute of Graduate Sttudies in Science and Engineering by using the plagiarism software to which İstanbul University is a subscriber.

This thesis benefited from a project (ViVA) that is supported by the project numbered 118C314 of TÜBİTAK.

## FOREWORD

After thanking God Almighty for the gift of intelligence to humankind and for His privileged audience, inspirations, and sustenance accorded to me while I sailed through the waves of this thesis work, I want to express my appreciation to some personalities in the following paragraphs, for their invaluable support towards this thesis work.

I'm deeply grateful to my main supervisor Assoc. Prof. Çiğdem SELÇUKCAN EROL for her unforgettable, forbearing, and motherly patience, her interest in not only my academics but also my personal life success, her guidance, suggestions, thoughts, encouragements and her sacrifices towards me and this thesis work as we worked out the scholarly tasks as well as the administrative tasks amidst the chaos of the deadly covid-19 pandemic.

Also, I am enormously thankful to my co-supervisor Assoc. Prof. MOHAMMAD ASIF KHAN for the patience, extreme tolerance, guidance, and suggestions he bore towards this thesis work. I would like to thank him for his pieces of advice on life, strong belief in my potentials, and deliberately exposing me to a wealth of opportunities.

Moreover, I would like to thank my lecturer Dr. MURAT GEZER for his commitment and excellently introducing me to the world of python programming and machine learning.

I owe much gratitude to our head of department, Prof. Dr. Sevinç GÜLSEÇEN for her patience and nobility in timely responding to all my academic requests without regard to the time and medium of communication despite her busy schedules.

I thank my entire family, especially my mom for her encouragements and voluntary prayers towards my well-being and success in all aspect of life.

And to say the least, I'm heavily indebted with gratitude and appreciation to my wife, Nana MARIAM for her support during this thesis work, her tolerance, and her unique respect and recognition of my achievements, however small they may be.

December 2022 Rashid MUKAILA

# TABLE OF CONTENTS

**Page**

## LIST OF FIGURES

**Page**

## LIST OF TABLES

**Page**

# LIST OF SYMBOLS AND ABBREVIATIONS

| Symbol | Explanation |
|---|---|
| **n(x)** | : Total number of distinct nonamers at a position |
| **x** | : Position |
| **i** | : Nonamer (9-mer) |
| **p** | : Probability |
| **H(x)** | : Entropy at a position |
| **p(i, x)** | : Probability at position x for a nonamer i |

| Abbreviation | Explanation |
|---|---|
| **WHO** | : World Health Organization |
| **IAV** | : Influenza A Virus |
| **CDC** | : Centers for Disease Control and Prevention |
| **SA** | : Sialic acid |
| **RNP** | : Ribonucleoprotein |
| **vRNP** | : viral Ribonucleoprotein |
| **RNA** | : Ribonucleic Acid |
| **CD-HIT** | : Cluster Database at High Identity with Tolerance |
| **MSA** | : Multiple Sequence Alignment |
| **NR** | : Non-redundant (deduplicated sequences) |
| **MUSCLE** | : MUltiple Sequence Comparison by Log- Expectation |
| **DIMA** | : Diversity Motif Analyser |

# ÖZET

## YÜKSEK LİSANS TEZİ

### INFLUENZA A VİRUSU ALT TİPLERİNİN KARŞILAŞTIRMALI ÇEŞİTLİLİK DİNAMİKLER

**Rashid MUKAILA**

**İstanbul Üniversitesi**

**Fen Bilimleri Enstitüsü**

**Enformatik Anabilim Dalı**

**Danışman: Doç. Dr. Çiğdem EROL**

**II. Danışman: Doç. Dr. Asif KHAN**

İnfluenza A virüs (IAV) alt tipleri, geçmişte birden fazla pandemiye neden olmasına rağmen, insanlık için pandemik bir tehdit olmaya devam ediyor. Halihazırda insanları enfekte eden tüm influenza A virüsleri (IAV'ler), farklı konakçılardan (insan dışı) kaynaklanmıştır, ancak mutasyon ve yeniden sınıflandırma dahil olmak üzere bir dizi çeşitlilik faktörü aracılığıyla insanları enfekte etme kabiliyetine ulaşmıştır. Dünya Sağlık Örgütü (WHO), insanları enfekte eden IAV'lerin yüksek çeşitliliğine ayak uydurmak için her yıl iki kez IAV aşılarını yeniden tasarlar. Halka açık veri tabanlarında IAV'lerin sekans verilerinin geniş mevcudiyeti, çeşitliliğin daha iyi anlaşılması için gerekli olan ve böylece IAV'lere karşı aşılar gibi etkili müdahalelerin tasarlanmasını kolaylaştıran mevcut insan enfekte IAV'lerin çeşitlilik çalışmalarını mümkün kılmaktadır. Alt türler arasında ve konakçı türler (kuş ve insan) arasında karşılaştırmalı analizler ile mevcut tüm verilerle çeşitliliğe nicel bir yaklaşıma çok ihtiyaç vardır.

Bu tez, seçilen ve kategorize edilen üç (3) influenza A virüsü alt tipinin çeşitlilik dinamikleri üzerinde konaklar arası (insan ve kuş) kapsamlı bir çalışma yürütmeyi amaçlamaktadır. Üç (3) influenza A virüsü alt tipi şu şekilde kategorize edildi: Yerleşik insan (H3N2), Gelişen insan (H7N9) ve Yalnızca kuş (H4N6). Bu alt tiplerin protein dizilerinin varyasyonundaki eğilimler, Shannon entropisinin kullanımı yoluyla çeşitliliğin miktarının belirlenmesinin yanı sıra,

proteinlerin her bir hizalanmış üst üste binen nonamer (9-mer) pozisyonundaki farklı çeşitlilik motiflerinin insidanslarını (oluşmalarını) hesaplanarak karşılaştırmalı olarak incelenmiştir. Nonamerler (9-merler) immünolojik uygulamalar için seçildi ve 9-mer'in sonraki 9-mer ile 8 amino asitle örtüşmesini sağlayan bir kayar pencere (sliding window) tekniği kullanılarak parçalandı, örneğin: 1-9, 2-10, 3-11, vs. 9-merlerin insidanslarının analizi için, her bir 9-mer'in insidansına göre (yüzde- % olarak) aşağıdaki kategorilere (diversity motifs) ayrıca gruplandırılmıştır: indeks, majör, minör ve benzersiz (index, major, minor, and unique). İndeks, pozisyonundaki baskın dizidir. İndeks diziden en az bir amino asit farklı olan tüm dizilere varyant denir. Majör varyant, en yaygın varyant sekansıdır, benzersiz varyantlar yalnızca bir kez meydana gelen tekillerdir ve minör varyantlar, çok sayıda farklı tekrarlanan olup da majör varyanttan daha az insidansı olan sekanslardır.

Tüm alt tipler arasında, PB1-F2'nin kuş konakçı dizilerinde insan konakçı dizilerine göre sürekli olarak çeşitliliği daha yüksek (dengesizliği) olduğu bulundu. Başka bir deyişle, pb1-f2'nin insan konakçılarda, aynı alt tipteki kuş konakçı karşılıklarına göre nispeten daha fazla korunmuş (stabile) olduğu bulundu. Ayrıca, PB1 ve PB2 proteinlerinin tüm alt tipler ve konakçılar arasında en çok korunan proteinler olduğu bulundu. Ek olarak, en yüksek insidans/oluşma (en düşük entropi) olan 9-mer'in yanı sıra, neredeyse her 9-mer pozisyonunda bir 9-mer varyantı vardı. 9-mer pozisyonlarının entropi değerleri de tüm alt tiplerde 0.90'dan fazla yüksek bir pozitif korelasyon gösterdi. Son olarak, her bir varyant motifi (majör, minör ve benzersiz), artan toplam varyantlarla (9-mer konumunda varyant motiflerinin toplamı) benzersiz bir ilişkiye sahipti. Major varyantı, toplam varyantların %50'sinde bir tepe insidansına sahipti, bunun ötesinde sadece minor varyantı yükselmeye devam etti ve o noktadan sonra pozisyonların geri kalanı için baskın varyant haline geldi (minör varyantı). Neredeyse tüm pozisyonların unique varyantı (benzersiz varyantı: bir pozisyonda yalnızca bir kez gözlemlenen 9-mer) sahip olmasına rağmen benzersiz varyant motifi, toplam varyantlar arttıkça neredeyse düz bir eğri sergiledi.

Aralık 2022, 76 sayfa.

**Anahtar kelimeler:** İnfluenza A virüsü, çeşitlilik dinamikleri, entropi, H3N2, H7N9, H4N6

# SUMMARY

## M.Sc. THESIS

COMPARATIVE DIVERSITY DYNAMICS OF INFLUENZA A VIRUS SUBTYPES

**Rashid MUKAILA**

**İstanbul University**

**Institute of Graduate Studies in Science and Engineering**

**Department**

**Supervisor : Assoc. Prof. Dr. Çiğdem EROL**

**Co-Supervisor : Assoc. Prof. Asif KHAN**

Influenza A virus subtypes continue to be a pandemic threat to humanity despite having caused multiple pandemics in the past. All influenza A viruses (IAVs) that currently infect humans originated from non-human hosts and achieved the ability to infect humans through diversification factors, such as mutation and reassortment. The World Health Organization redesigns the IAV vaccines twice every year in order to keep up with the high diversity of IAVs that infect humans. The large availability of sequence data of IAVs in public databases enable diversity studies of existing human infecting IAVs, which is necessary to gain better understanding of the diversity, thereby facilitating design of effective interventions, such as vaccines against the IAVs. A quantitative approach to the diversity with all available data, with comparative analyses across subtypes and between host species (avian and human) is much needed.

The thesis herein aimed to conduct a comprehensive inter-host (human and avian) study on the diversity dynamics of three (3) influenza A virus subtypes selected and categorized as: (i) established human (H3N2), (ii) emerging human (H7N9) and (iii) only avian (H4N6). Trends in the variation of protein sequences of these subtypes were comparatively studied by quantifying the diversity through use of Shannon's entropy, as well as calculating the incidences (occurrences) of distinct diversity motifs at each aligned overlapping nonamer (9-mer) position

of the proteins. Nonamers (9-mers) were selected for immunological applications and the 9-mers were generated by use of a sliding-window approach, where each 9-mer overlapped eight amino acids with the next 9-mer, such as 1-9, 2-10, 3-11, *etc*. The diversity motifs comprised of index and its variants (major, minor and unique). The index is the predominant sequence at a nonamer position and all the other nonamers (at that position) are referred to as variants, with one or more amino acid difference from the index. The major variant is the variant nonamer sequence that occurred more than all other variant nonamer sequences, making it the most common variant nonamer, while the unique variants are singletons that occurred only once. The minor variants are those 9-mers that occurred more than once and yet with incidences not up to the incidence of the major variant.

Across all the subtypes, the protein PB1-F2 was observed to be constantly more diverse (unstable) in the avian IAVs compared to the human IAVs. Furthermore, PB1 and PB2 proteins were found to be the most conserved proteins across all the subtypes and hosts. In addition, nearly every 9-mer position had a variant 9-mer, beside the 9-mer with the highest incidence/occurrence (index) at the same position. The entropy values of the 9-mer positions also showed a high positive correlation to total variants, with Pearson correlation coefficient (r) of greater than 0.90 across all the subtypes analysed. Further, each variant motif (major, minor, and unique) had a unique relationship with increasing total variants across the proteome. Major variant exhibited a peak incidence at 50% total variants, beyond which only the minor variants continued to rise and were collectively predominant. The unique variants were observed for nearly all positions, but were of low incidence and showed a uniform trend.

December 2022, 76 pages.

**Keywords:** Influenza A virus, diversity dynamics, entropy, H3N2, H7N9, H4N6

# 1. INTRODUCTION

Each year, the World Health Organization (WHO) redesigns seasonal vaccines for the Influenza virus, and this is done twice in a year (CDC, 2020). This is done to maintain the effectiveness of the influenza vaccines available since the virus changes its "signatures" at a high rate. Despite all these efforts and the relative technological advancements in the health sector, influenza A virus (IAV) remains a huge public health problem and a serious pandemic threat (Eng et al., 2014; Szewczyk et al., 2014). Although the IAV virus is mainly in avian hosts, its greater mutation ability and other variation processes have rapidly created strains that are able to infect humans and have caused both epidemics and serious worldwide pandemics (Tscherne & García-Sastre, 2011). As a result of this threat and the fact that more and new influenza A viral sequences are being deposited in databases, it calls for continuous effort in research to gain more and updated understanding of the diversity dynamics of the Influenza A virus (IAV). The 2019 COVID virus (SARS-CoV-2) has demonstrated the need to not only research on vaccines and viruses that are already causing fatalities, but also viruses that have the potential of at least infecting humans. The SARS-CoV-2 caught the world unprepared against its infection and caused unprecedented casualties, made worse by the long process of manufacturing new vaccines. Also, more IAVs, which were previously known to only infect avian hosts are being reported as gradually adapting and beginning to infect humans too. As such, there is the need to conduct diversity studies comparatively for IAVs that have already adapted and gained stability in infecting humans, versus IAVs that are still regarded as mainly infecting avian hosts. This will help in both better vaccine design and to gain further understanding of this "jumping" mechanism from the diversity point of view, which may pave ways to the possibility of minimizing or controlling IAV infections.

Research on IAVs is made rapidly obsolete by the presence of large amount of new data. Also, there is limited literature on the comparative analysis across subtypes between different hosts. One of the extensive comparative analysis studies on IAVs is the work of (Heiny et al., 2007), which analysed the diversity of major IAVs (human and avian hosts isolates) reported before and during the year 2007.

This thesis aims to conduct a comprehensive inter-host (human and avian) study on the diversity dynamics of three (3) influenza A virus subtypes selected and categorized as: established human (H3N2), emerging human (H7N9) and only avian (H4N6).

## 1.1. INFLUENZA A VIRUS AND DIVERSITY

Viruses are the most dominant biological entities on earth (Moelling & Broecker, 2019). This is exacerbated by the ability of viruses to change (due to mutation and other genetic diversity generating factors), which timelessly can give rise to even more unique viruses on the planet.

It is important to understand that the increasing number of viruses does not necessarily mean an increasing number of human viral infections. This is because not all viruses are harmful to humans. In fact, some viruses are directly or indirectly necessary for the survival of humans. Within the human body system, the food and chemical engineering processes, and the environment as a whole, viruses play some essential roles to complement the survival of humans and animals.

Viruses of RNA genomes, such as the Influenza A virus (IAV), are known to have high rates of mutation due to the lack of proofreading in the enzymes responsible for making copies of the genome (Scroggs et al., 2019), leading to increased diversity in such viruses. This mutation often occurs during replication and recombination or reassortment between genetic materials (Mumford, 2007; Alnaji et al., 2022). Several evidence indicate that mutation through reassortment is common in IAVs (Webster et al., 1992). For example, all past IAV pandemics of humans were caused by subtypes that resulted from reassortment occurring in avian subtypes (Diskin et al., 2020). Although, diversity of species increases the functioning and stability of an ecosystem (Jochum et al., 2020), in the case of the viruses that are deadly to humans, this stability means fatality to humans. Aside from being RNA based, the segmented nature of the IAV genome also contributes to the observed high diversity rates of IAVs (Piret & Boivin, 2021a).

Although diversity occurs in all gene segments of IAVs (Webster et al., 1992), the classification of IAVs is done by various combinations of only the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA), which are among the most variable. In 1976, research led to the

discovery of waterfowl as the main reservoir of IAVs and since then about 96 combinations of HA and NA glycoproteins of IAVs have been recognised (Diskin et al., 2020).

Diversity studies of IAVs is crucial to public health because the diversity complicates the design of vaccines against the virus (Abd Raman et al., 2020). IAV vaccines are prepared to trigger immunity against current circulating and predicted IAV strains (CDC, 2020). However, as those strains evolve over a period of time into new strains, the diversity of which is not captured by the existing vaccines, it rapidly renders such vaccines inefficient against infections caused by these new strains.

### 1.1.1 Antigenic Shift and Antigenic Drift

The surface glycoproteins of the IAVs (HA and NA proteins) are termed as "antigens" because they are detectable by the host immune system and they can cause an immune response to be triggered (CDC, 2021). Due to the absence of proofreading in making copies of the genome during replication, IAVs, just like other RNA virus, have high mutation rates. Sometimes, these mutations may cause changes in the amino acids of the antigens, HA and NA, causing them to have different signatures as opposed to what the current host immunity has of them. This effectively makes them potentially escape recognition by the immune system. Such minor changes occurring in the surface proteins (HA and NA) of IAVs is termed as antigenic drift (Taubenberger & Kash, 2010; CDC, 2021).

The segmented genome nature of IAVs (7-8 segments of genes) makes them prone to reassortments when different IAV strains co-infect the same host cell (Piret & Boivin, 2021a). This results in exchange of gene segments leading to progenies that have gene segments from both parents that took part in the co-infection. When this process leads to exchange or complete change of HA or NA antigens (proteins), it is referred to as antigenic shift (Taubenberger & Kash, 2010). Antigenic shift constitutes a major change resulting in antigens that are not recognized by previous immunity of the host, and this can lead to pandemics; nevertheless, antigenic shift occurs less frequently as compared to antigenic drift (CDC, 2021).

Both **antigenic shift** and **antigenic drift** play a major role in the viral diversity of IAVs. Antigenic shift instantly results in new subtypes when the strain involved gains a new HA or NA protein from another strain of a different subtype. While the diversity created by antigenic drift does not instantly result in a new IAV subtype, it continues to compound over time and

can eventually result in creating distinct strains (CDC, 2021), which can possibly lead to the creation of a new subtype.

## 1.2 PANDEMICS OF IAV

In 2010, during the time in which the H1N1 pandemic was happening, (Hughes *et al*., 2010) in their article made a profound statement that the then pandemic, plus the newly emerging viral strains like the SARS-coronavirus are a reminder to humanity about our vulnerability and inability to predict or prevent such pandemics and similar threats to human health. Since then (2010 to 2019), a lot of advancements and improvements in our health technologies have been achieved, such as improvements in CRISPR and next generation sequencing technologies. However, the current 2019 SARS-CoV-2 (coronavirus) pandemic is yet a reminder that the statement by the authors about our preparedness still holds.

According to WHO, Influenza virus causes 290,000–650,000 deaths every year (Taheri et al., 2021). This is only accounted for by the yearly epidemics of Influenza virus, contrary to over 50 million deaths (Johnson & Mueller, 2002) that a single IAV pandemic, such as the Spanish flu (1918–1919) had caused.

With the majority of IAV pandemics triggered by animal-to-human interhost crossing of viruses (Hughes et al., 2010), it is becoming apparent that this mechanism is the main route of the virus in causing pandemics. However, the journey of a virus from animal host to human host doesn't happen at once, instead, it's a multifactorial series of events (albeit not fully understood yet) that play out resulting in the strain to reach its destination host (Piret & Boivin, 2021a). There are five (5) stages involved in the process of crossing from one specie host to a different species host (Wolfe et al., 2007):

(1) Due to natural factors, the animals get infected by the pathogen.

(2) The pathogen diversifies until it can infect humans directly, but at this stage it cannot cause human-to-human infection.

(3) The pathogen can cause human-to-human infection, but for only a few times.

(4) The pathogen still causes disease in animals and at the same time it's able to sustain several human-to-human infections.

(5) The pathogen causes the disease only in humans, thus, becoming fully human-hosted.

Some of the external factors suggested to be playing a role in modulating cross-species transmission of pathogens include the manner and frequency with which humans come into contact with virus-reservoir animals, land use and climate change (White & Razgour, 2020; El-Sayed & Kamel, 2020; Piret & Boivin, 2021a). Beside the recorded IAV pandemics (1889, 1918, 1957, 1968, 1977 and 2009), IAV epidemics have likely occurred in ancient times (Taubenberger & Morens, 2009; Taubenberger & Kash, 2010). The IAV pandemic of 1889, named: "Russian flu" pandemic, was, till that time, the most well recorded and 'tracked' in real-time (Taubenberger & Morens, 2009). The virus reached the entire globe in about 4 months' time and claimed approximately 1 million lives worldwide (Valleron et al., 2010; Piret & Boivin, 2021b). According to epidemiologic and serologic data, the pandemic was suspected to have been caused by an A/H3N8 virus (Worobey et al., 2014).

The 1918 (1918 - 1919) "Spanish flu" Influenza pandemic, often cited as the worst in recorded history, resulted in over 50 million deaths worldwide (Johnson & Mueller, 2002). The pandemic also uniquely identifies itself by preceding previous pandemics with three waves of appearances and featuring a peak of high mortality in healthy young adults (Morens & Taubenberger, 2018).

In J.K. Taubenberger's laboratory, reconstruction of the viral genome of the pandemic using tissues of several victims demonstrated that the causative agent was a descendant of an avian H1N1 virus, and this has been said to be the "mother" of all the IAV progeny circulating and killing today (Taubenberger et al., 2007; Morens & Taubenberger, 2018). The 1957 – 1958 pandemic, "Asian Flu" was caused by a H2N2 IAV subtype, a descendant of the H1N1 which evolved by acquiring three novel genes (HA, NA and PBI) from an unknown avian virus through reassortment (Taubenberger & Kash, 2010).

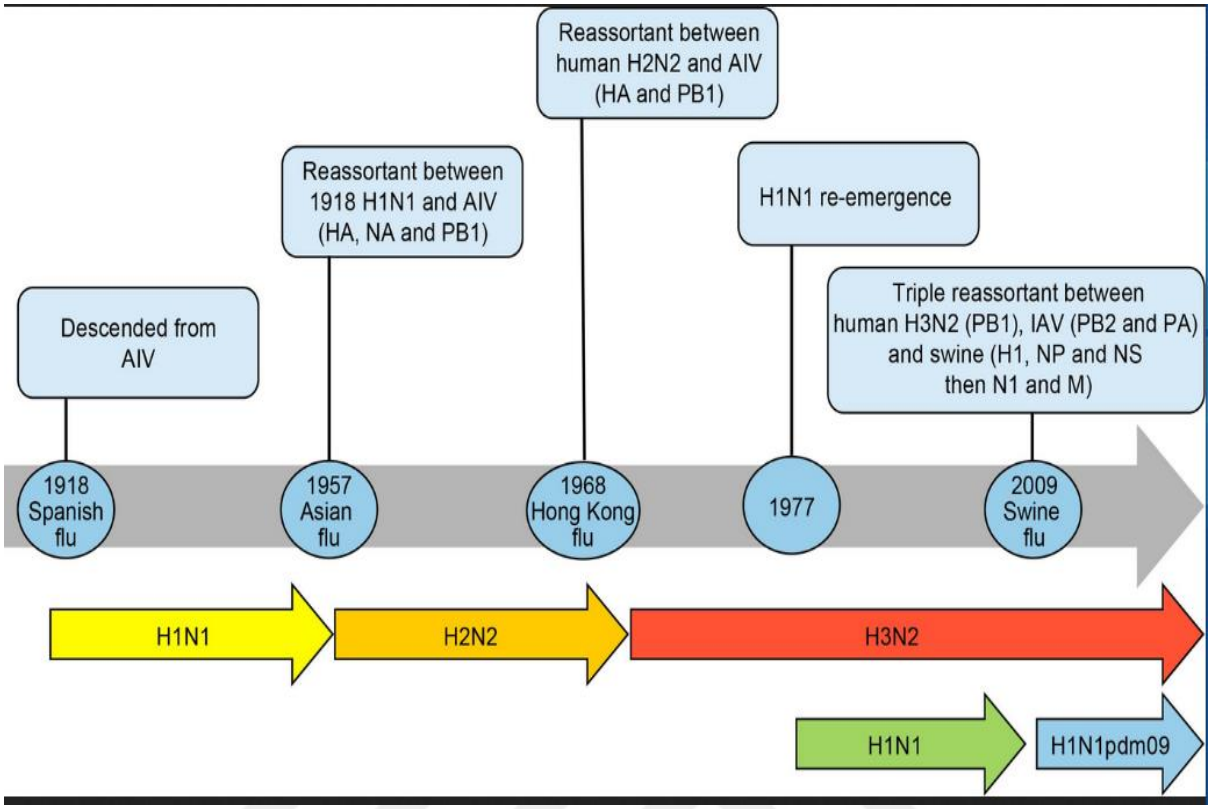


**Figure 1.1:** Timeline of pandemics of influenza A virus (IAV). Taken from (Piret & Boivin, 2021b).

During this reassortment, the HA, NA and PB1 gene segments of the H1N1 subtype were replaced by the HA, NA and PB1 gene segments of an avian-like subtype respectively (Piret & Boivin, 2021b). The 1968 (1968 – 1970) pandemic, named: "Hong Kong flu" killed about 2 million people (Saunders-Hastings & Krewski, 2016). Continuing the lineage of the 1918 virus (H1N1), the Hong Kong flu pandemic was caused by an A/H3N2 virus, a reassortant of A/H2N2 virus, evolved through replacement of the HA and PB1 gene segments (Piret & Boivin, 2021b). As shown in Figure 1.1, the 1977 pandemic (1977 – 1978.) was a "zombie" H1N1 virus, remerging in 1977, 20 years after it had disappeared at the end of the 1918 pandemic. Since it is not very likely that a virus would remain in nature for 20 years without going through mutational changes, it has been suggested that the virus was a frozen strain dating back to the 1950s (Taubenberger & Kash, 2010). As can be seen in Figure 1.1, at the time of this re-emergence, the human H3N2 was still in circulation and thus, both H3N2 and the 1918 origin H1N1 co-circulated until the H1N1 was replaced by the swine-origin H1N1 2009 pandemic virus (Taubenberger & Morens, 2020). In 2009, a swine-origin human H1N1 IAV, sometimes

called H1N1pdm09, caused a pandemic that brought about lots of casualties and economic losses (Dawood et al., 2009). The virus is reported to have first emerged in La Gloria, a village in Mexico, and by June 2009, it was declared by WHO as a pandemic (Moore et al., 2021).

## 1.3 INFLUENZA VIRUS GENOMIC ARCHITECTURE

The influenza virus is an enveloped virus of the *Orthomyxoviridae* family that contains a negative-sense single strand RNA genome, segmented into 7-8 genes (Taubenberger & Kash, 2010; Becker et al., 2021). The genera of the Influenza *Orthomyxoviridae* family include thogotovirus, isavirus, influenza virus A, influenza virus B, influenza virus C (Taubenberger & Kash, 2010) and the most recent influenza virus D (Hause et al., 2014; Liu et al., 2020). Among all the influenza viruses, the most epidemiologically significant, associated with pandemics (Taubenberger & Kash, 2010), and most severe in disease manifestations is IAV (Vasin et al., 2014).

The IAV genome of x segments encodes for at least 11 proteins, which includes the non-structural proteins (NS1 and NS2), the matrix proteins (M1 and M2), the main surface glycoproteins hemagglutinin (HA) and neuraminidase (NA), the polymerase subunits (PA, PB1 and PB2), the nucleocapsid protein (NP), and the quite recently discovered proteins (PB1-F2, N40, PA-X, and M42) (Taubenberger & Kash, 2010; Kumar et al., 2018; Sun et al., 2020). NS2 is also referred to as the nuclear export protein (NEP). The envelop of IAVs, derived from host cell lipid membrane, have three proteins protruding from it (HA, NA, and M2), which can be seen in Fig 3B. The HA functions both as a fusion and receptor binding protein (Taubenberger & Kash, 2010). M1 is the major component of the viral envelope. Found underlining the membrane, the envelope plays several roles in virion assembly and infection (Taubenberger & Kash, 2010; Vasin et al., 2014). It interacts with both the viral ribonucleoprotein (vRNP) complexes and the surface glycoproteins. NA has sialidase enzymatic activity, which is needed for the release of newly created virions and for cleavage off sialic acids (SAs) on viral glycoproteins (to prevent nascent viral particles from aggregating) and SAs of host cell (Taubenberger & Kash, 2010; Khanna et al., 2014). NP forms the main part of the viral RNP complex and has the function of controlling the nuclear cytoplasmic transport of RNA. The RNA segments in virions are wrapped around NP monomers, and, together with the polymerase proteins (PB1, PB2 and PA), they form the RNP complex (Taubenberger & Kash, 2010; Vasin et al., 2014).

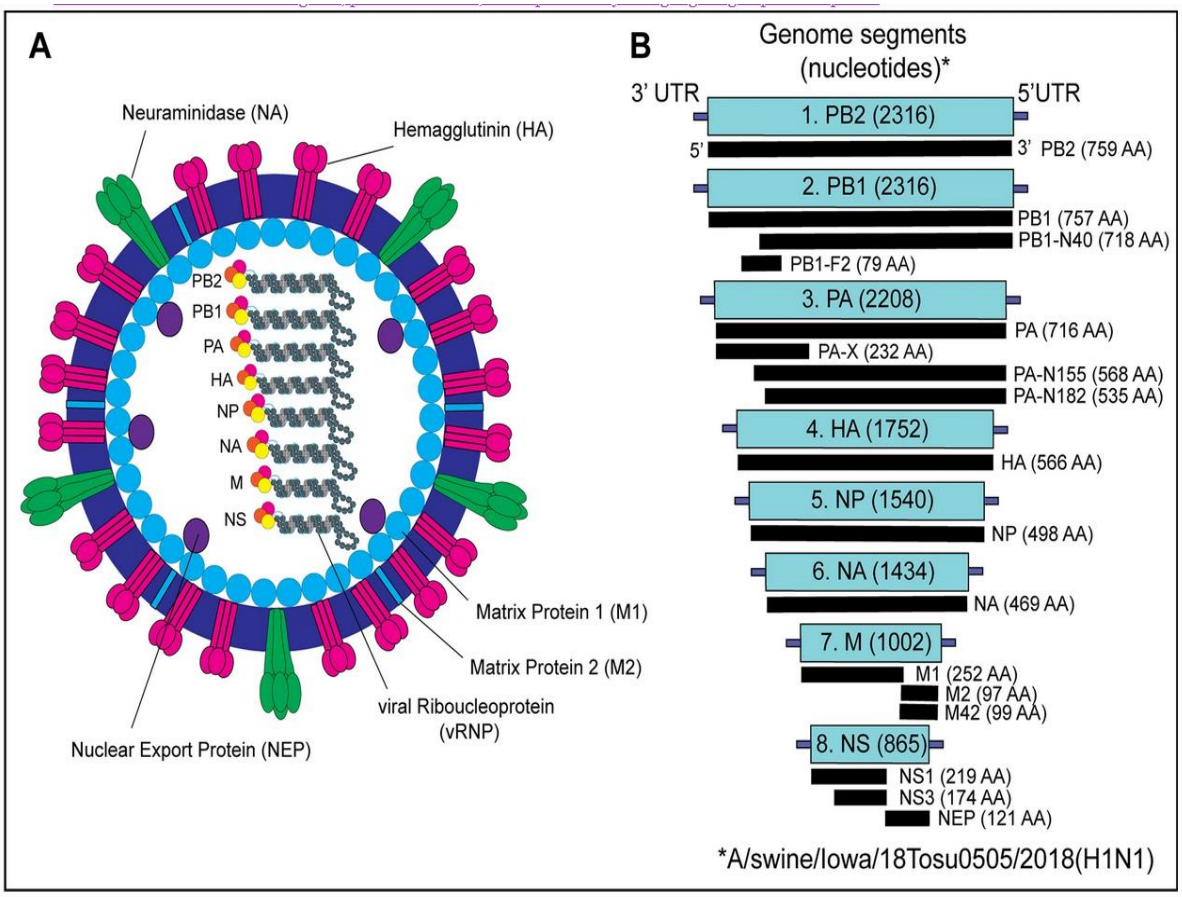


**Figure 1.2**: A: Schematic diagram of IAV envelope and B: IAV gene segments' length breakdown (lengths might vary from that of other strains). Taken from (Chauhan & Gordon, 2022).

During virus entry, M2 facilitates genome unpacking. It forms a proton channel, which gets activated in low pH conditions (Vasin et al., 2014). PB1 interacts with PB2 and PA and acts as a catalyst to the viral RNA-dependent polymerase responsible for RNA chain elongation. It is also necessary for both RNA transcription and replication (Vasin et al., 2014). As depicted in Figure 1.2 A, the PB2, another direct component of the vRNP complex (accompanying PA and PB1), is the first and the longest segment along the IAV genome of 759 amino acids (Eisfeld et al., 2015; Chauhan & Gordon, 2022). More importantly, the PB2 has a major influence in determining viral pathogenicity and virulence of IAVs, although how it achieves this is poorly understood (Long & Fodor, 2016). The PA segment of the IAV genome encodes the PA protein. As can be seen in Fig 3A, the PA protein, which together with PB1 and PB2 forms the vRNP complex, has a mutual function of being a requirement for replication and transcription in the nucleus of host cells (Obayashi et al., 2008; Mehle et al., 2012). PA also functions in activities such as cap-binding, promoter-binding and endonuclease (Hara et al., 2006). NS1 plays

multiple roles in the interaction between the virus and its host cell. These include the regulation of host and viral gene expression as well as enabling the virus to escape antiviral mechanisms of the host (Vasin et al., 2014). Just like NS1, NS2 is also a multifunctional protein but has been identified to play a major role of mediating the binding of the M1 matrix proteins to vRNPs and also facilitates the exportation of vRNPs from the nucleus of host cells (Shimizu et al., 2011; Paterson & Fodor, 2012; Brunotte et al., 2014). Researchers have identified the NS3 protein, however, its function is yet to be well understood (Chauhan & Gordon, 2022).

### 1.4 SIGNIFICANCE OF STUDYING SUBTYPES H7N9, H3N2, AND H4N6

Since the first emergence of the H7N9 IAV subtype human infection in China, which caused human epidemic in the country (To et al., 2014; Rejmanek et al., 2015), there had been increasing human cases of H7N9 spanning five waves of human infections between 2013 and 2017, with 49 deaths out of 159 cases in January 2014 (To et al., 2014), totalling 615 deaths out of 1568 cases from 2013 to 2017 (Yin et al., 2021). By applying an H7N9 vaccine in the poultry in China, further human infections of H7N9 were greatly brought under control since September 2017. However, H7N9 has not been eliminated in poultry, making it one of the lingering potential threats to human health that demands attention (Yin et al., 2021). Furthermore, despite the fact that H7N9 is not popularly known to sustain human-to-human transmission, it has been identified to have alarmingly high mortality rates in infected people (Rejmanek et al., 2015).

The A/H3N2 virus, which killed approximately 2 million in the 1968 Hong Kong flu pandemic, is still around and keeps causing deaths in the form of seasonal cases. Over the years, it has been observed that influenza seasons, where H3N2 cases dominate are more severe, with higher number of deaths and hospitalizations than the seasons dominated by other subtypes (Vemula et al., 2016). For example, the influenza season that took place from the year 2017 to 2018 in the United States had A/H3N2 as the dominant circulating IAV subtype and was highly severe IAV season recording 80,000 deaths with 900,000 hospitalizations, registering it as the IAV season with the third most outpatient visits in the past 20 years for IAV-like illness (Allen & Ross, 2021). Just like other IAV viruses, with different lineages of A/H3N2 still living in avian hosts like the poultry, the world still stands a risk of having any of those lineages crossing over to the human host and causing pandemics. Guan et al. (2019) detected various lineages of H3N2 avian viruses in live poultry markets in China and found that the avian H3N2 viruses have

distinct replication phenotypes in mice, binding to both human-type and avian-type receptors and spreads efficiently to guinea pigs in direct contacts.

Avian H4N6 poses a threat to human health due to its, aside from other factors, resemblance in characteristics with some avian IAV's, which are currently established as causing fatalities in humans like the H3N2. Both avian H3N2 and avian H4N6 have been demonstrated to be able to bind to avian-type and human-type receptors, replicate efficiently in mammals (mice) without adaptation, and through direct contact spreads very easily to guinea pigs (Bui et al., 2012; Guan et al., 2019; Liang et al., 2016; Shi et al., 2016). H4N6 was also found to be easily transmissible through water (VanDalen et al., 2010) and this poses a threat to agriculture as many flocks and other animals in the free-range often share the same source of water in open places.

The above facts about the three IAV subtypes under study and the daily increase in data on the subtypes make it expedient to sustain a continuous study on these subtypes. Above all, while preventing viral infections through vaccines has shown great success in fighting against pandemics, we may achieve better results if we are able to prevent avian viruses from crossing from avian and other hosts to human hosts (zoonosis) through early detection of such a potential. IAVs have been hosted by waterfowls likely for many years before they were discovered (1976) to be mainly originating from waterfowls, and during all those years, they caused less to no casualties to their avian hosts. However, once they jumped and adopted to the human hosts, they became alarmingly fatal to humans. Therefore, preventing avian to human host crossing of IAVs should be considered as a complimentary approach to vaccination. This study herein aims to analyse sequence diversity patterns of these three subtypes, categorises into three host categories: i) established human host (H3N2), ii) emerging human host (H7N9), and iii) only avian/commonly avian host (H4N6). The objective is to gain a better understanding of the transition from avian to human.

# 2. MATERIALS AND METHODS

## 2.1 DATA GATHERING AND PROCESSING

Protein sequences of H7N9 and H3N2 subtypes of human and avian hosts, and protein sequences of H4N6 IAV subtype from only avian host were downloaded from the public repositories GISAID (https://www.gisaid.org; as of March 2022) (Elbe & Buckland-Merrett, 2017; Shu & McCauley, 2017; Khare et al., 2021) and FluDB (https://www.fludb.org; as of March 2022) (Zhang et al., 2017). Both databases allow the download of sequences that are already classified into individual proteins. Based on the number of records (data) available for each protein, only 12 proteins were selected for download and further analyses (NS1, PB2, NA, PB1, M1, PB1-F2, NS2, PA, NP, HA, PA-X, M2). While IAV's have more than 12 proteins, this thesis focused on only the above 12. This was because, the rest of the proteins had significantly less data in the databases and would not guarantee a fair analysis if used together with the 12 proteins selected.

The three (3) subtypes were evaluated for their host specificity based on the available data ratio between human and avian host sequences. As per the notion from the literature, H3N2 has been documented as infecting both humans and avian, and thus was classified as "Established Human". H7N9 was classified as "Emerging human" due to the gradual increase in reports of its presence in both human and avian hosts. H4N6 subtype was classified as "Only avian" because it is majorly reported as infecting mainly avian hosts. The ratio of available data between the hosts supported the classification.

Because the data was downloaded host by host from two databases, for each subtype, human host data from GISAID was added to human host data from FluDB to form a single human host data for that subtype. The same was done with the avian host data. After which, filtering was done to each host data as follows: Sequences shorter than 50 amino acids (aa) and those containing unknown aa character "X" were removed to improve the ease and quality of multiple sequence alignment (MSA). Upon this filtering, each protein's duplicated sequences were removed using the deduplication tool, CD-HIT (Li & Godzik, 2006; Fu et al., 2012) to get the non-redundant (NR) sequences. Default parameters of CD-HIT was used (Table 2.1). Before the resulting NR data was aligned, for the H3N2 and H7N9 subtypes, human host protein sequences of each subtype were appended to their counterpart protein sequences of avian host

in their respective subtypes, eventually resulting in 12 protein dataset per subtype instead of the previously separated datasets of 12 protein human host and 12 protein avian host dataset per each subtype (H3N2 and H7N9). This was done because the thesis is primarily studying the diversity (mutations) which occurred while the virus was transitioning from avian host to human hosts and thus, the multiple sequence alignment (MSA) tool will need data of both hosts sequences at a time to be able to indicate what changes occurred in the sequences of human hosts as compared to avian hosts and vice versa (hosts are separated later). This was not done to the H4N6 subtype since it had only avian host protein sequence data. The resulting datasets of each protein were then aligned using the command line MUSCLE (Edgar, 2004) MSA tool. Default parameters of MUSCLE was used for this thesis (Table 2.1). MUSCLE was chosen for the MSA because it produced better alignment after trying the dataset with MUSCLE and two other MSA tools (MAFFT and Clustal Omega). After the alignment was done the protein dataset of each subtype was separated back into their respective human and avian host NR sequence data. The aligned NR datasets were then submitted to DIMA (Tharanga et al., 2022) to perform entropy calculation and carry out diversity motif analyses as carried out by (Hu et al., 2013) and (Abd Raman et al., 2020). Figure 2.1 provides an overview of the workflow employed.

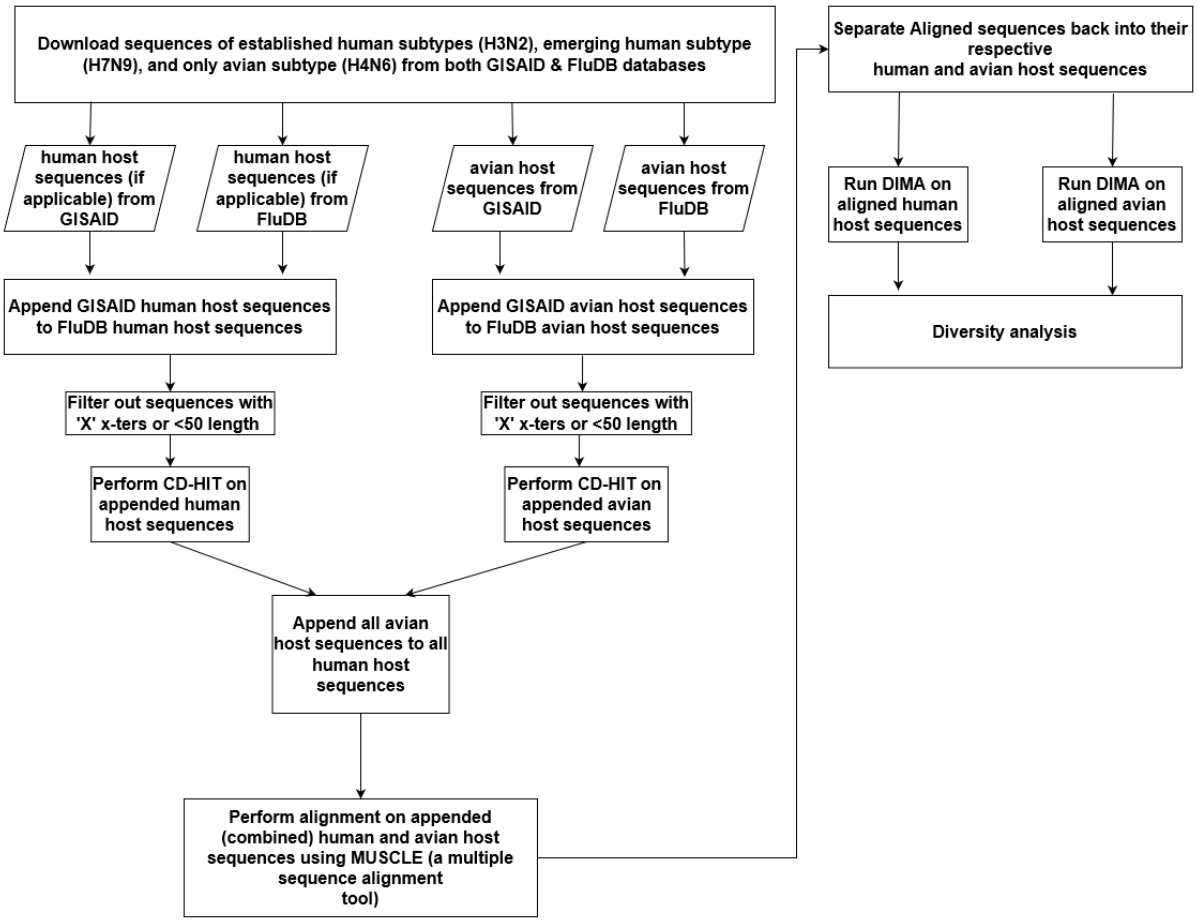


**Figure 2.1:** Thesis workflow diagram.

**Table 2.1:** Information on tools/software used in this thesis.

| Tool | Version | Parameters/Algorithms used | Purpose of Use | Citation |
|---|---|---|---|---|
| Python3 | 3.8.1 | N/A | Main programming language used for all tasks. | (Van Rossum & Drake, 2009) |
| Matplotlib | 3.3.2 | N/A | Plotting of graphs | (Hunter, 2007) |
| seaborn | 0.11.0 | N/A | Plotting of graphs | (Waskom et al., 2017) |

| | | | | |
|---|---|---|---|---|
| MUSCLE | 3.8.1551 | Default parameters were used:<br><br>**Default Parameter: value**<br>-maxiters:16<br>- maxtrees: 1<br><br>-diagonal optimization: disabled<br><br>**Default flags:**<br><br>-le<br><br>(-le stands for the *Log-expectation score* which is the default profile scoring function called) | For performing multiple sequence alignment (MSA) | (Edgar, 2004) |
| CD-HIT | 4.8.1 | Default parameters were used: | For removing duplicate sequences. | (Li & Godzik, 2006; Fu et al., 2012) |
| Biopython | 1.79 | N/A | For programmatically processing fasta files | (Cock et al., 2009) |
| DIMA | 4.1 | Default parameters were used: | For analysing sequence diversity | (Tharanga et al., 2022) |

## 2.2. NONAMER ENTROPY & DIVERSITY MOTIF ANALYSES

Using a sliding k-mer window approach as utilized by (Abd Raman et al., 2020), DIMA (https://dima.bezmialem.edu.tr) chunks the aligned sequences into a user-specified *k*-mer length, in our case, nine (9; nonamer) for immunological applications. Nine (9) is the most common length of peptides or their binding cores that the human leukocyte antigens (HLAs) mostly recognize (Heiny et al., 2007; Rammensee, 1995). The sliding k-mer window approach is such that each 9-mer overlaps the subsequent 9-mer by 8 amino acids, for example: 1-9, 2-10, 3-11, 4-12, 5-13 are five 9-mers chunked from a protein sequence with amino acids that fall within position 1 and 13. For the analysis of the 9-mer incidences, each 9-mer was further grouped into the following categories ( termed as diversity motifs) based on the incidence (in

percentage - %) of that 9-mer: index, major, minor, and unique (Figure 2.2.). DIMA then calculates the entropy of each nonamer using a modified Shannon's entropy calculation formula (Khan et al., 2006). Nonamer positions (1-9, 2-10, 3-11, *etc*) that had less than 30 sequences were termed as positions of low statistical support. Such positions were evaluated for entropy but were not included in the eventual analysis.

The entropy *H(x)* of a nonamer *i* at position *x* is calculated as below:

$$H(x) = -\sum_{i=1}^{n(x)} p(i,x)\log_2 p(i,x) \tag{2.1}$$

***2.2.1. Sequence Diversity Entropy Calculation Formula (Abd Raman et al., 2020).***

Where *H* equals the entropy at a position *x*, *p (i, x)* is the probability of a nonamer *i at x,* and *n(x)* is the number of distinct nonamers at *x*. An increase in the total number of distinct nonamers causes increase in the entropy value (*H*). The relative probability of the nonamers also affects the entropy value such that the nonamer position where one nonamer is largely predominant, the entropy value reduces, which may imply a complete conservation of that position. The lower the value of the entropy at a (nonamer) position, the higher the conservation of the position. Thus, in theory, a position with 0 entropy is regarded to be at the highest level of conservation (complete or 100% conservation), while a nonamer position with an entropy of 39 is regarded to be at the lowest level of conservation. This is only achieved if all possible combinations of the nonamer at that position are observed with equal frequency (Abd Raman et al., 2020). It is a theoretical value given that sequences within a species exhibit a certain degree of conservation. Further, certain combinations of nonamer peptides are sterically not plausible, such a nonamer of only prolines. The highest nonamer entropy value that has been observed thus far is 9.2 for Env protein of clade B human immunodeficiency viruses (HIV-1) (Hu et al., 2013). The total number of sequences in the alignment also influences the entropy calculation, whereby the larger the sample size the lesser the bias. As described by Khan et al. (2008), a size bias resulting from this influence is corrected by using a statistical adjustment.

DIMA was then used to classify the distinct sequences at each nonamer position into diversity motifs based on their frequencies (Figure 2.3): “index” and its total variants (“major”, “minor” and “unique”) (Tharanga *et al*., 2022) (these were defined in the introduction section).

The index is the predominant sequence at a nonamer position and all other nonamers (at that position) with one or more amino acid difference from the index are referred to variants. The major variant is the variant nonamer sequence that occurred more than all other variant sequences, making it the most common variant nonamer., while the unique variants are singletons that occurred only once, and the minor variants are those of incidences in-between the major and the unique (Figure 2.2).

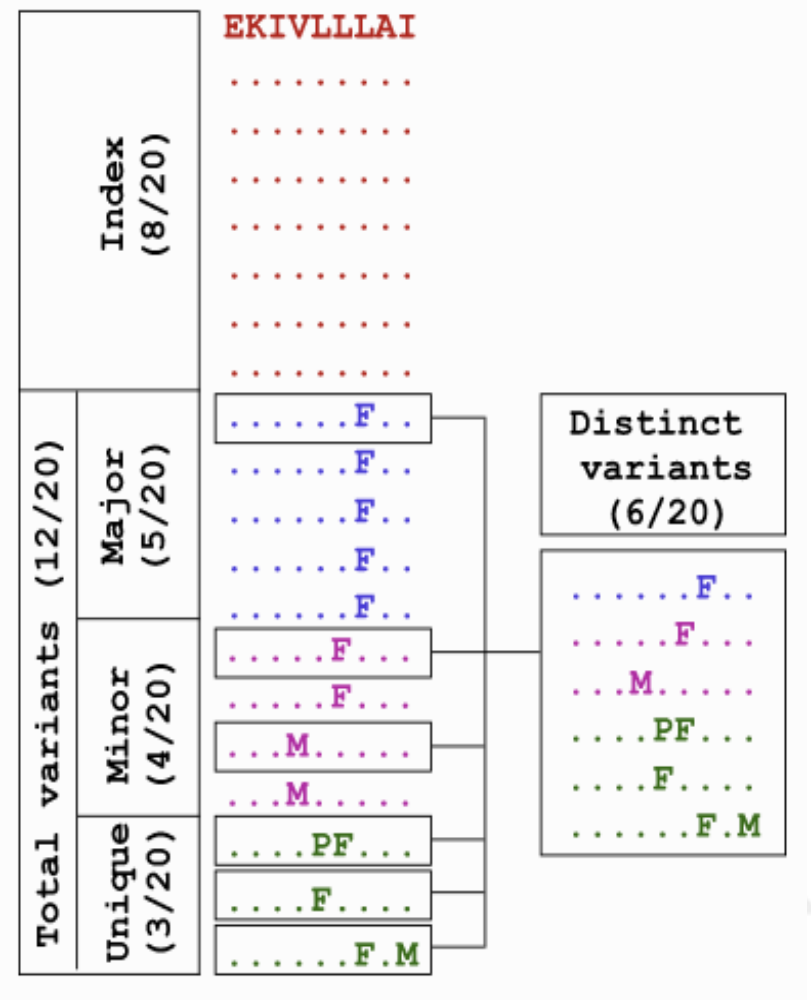


**Figure 2.2:** Diversity motifs[1]. Taken from (Tharanga et al., 2022).

[1] The index is the predominant sequence at a nonamer position and all other nonamers (at that position) with one or more amino acid difference from the index are referred to variants. The major variant is the variant nonamer sequence that occurred more than all other variant sequences, making it the most common variant nonamer, while the unique variants are singletons that occurred only once, and the minor variants are those of incidences in-between the major and the unique, that is, the minor variant is the second most common variant nonamer after the major variant. Distinct variants are the number of all the different variant nonamers at the position (excluding the index nonamer which is not considered a variant). In this figure there are 6 of them. Distinct variants are used to derive "Nonatypes", which is basically the percentage of distinct variants at a position in this case it is 30% (6/20).

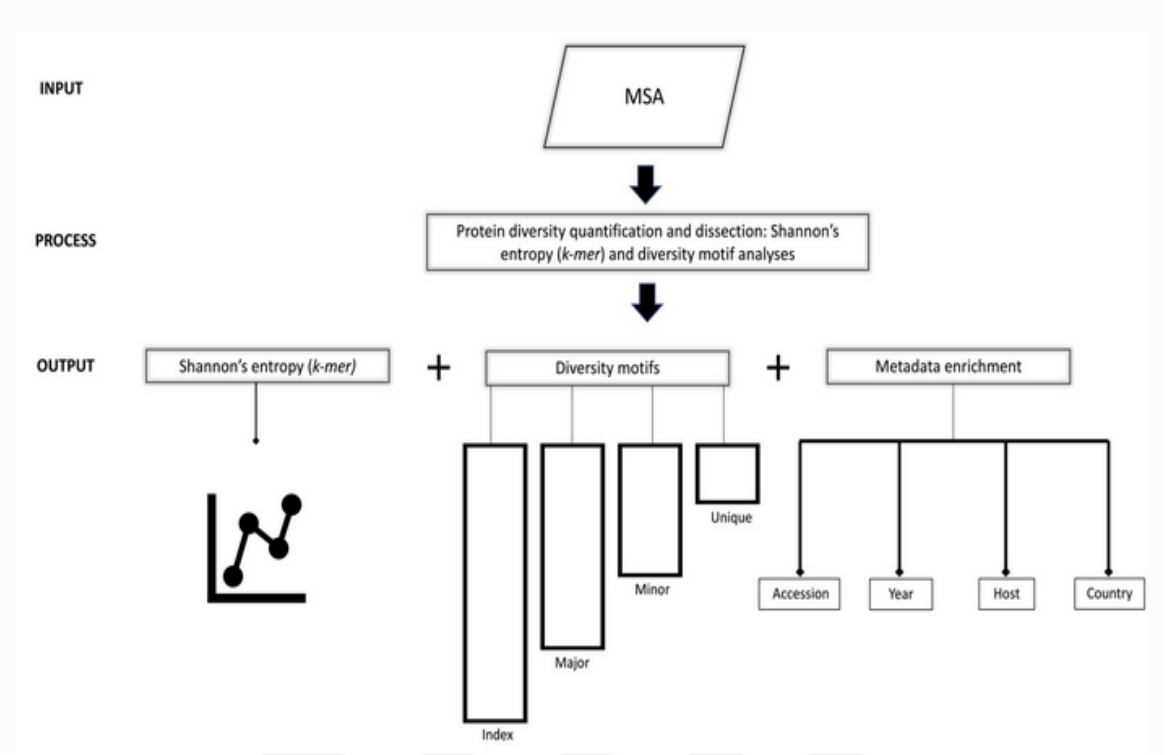


**Figure 2.3:** General workflow of DIMA in calculating entropy and performing the diversity motif analyses. Taken from (Tharanga et al., 2022).

The resulting diversity motif data were further processed and analysed for protein specific and proteome-wide diversity patterns for each subtype and the corresponding host. Visualisations were made by use of matplotlib, seaborn and pandas of python3.

## 2.3. QUANTIFYING SEQUENCE CONSERVATİON ACROSS ALL THREE SUBTYPES

The distribution of the sequence conservation across the proteomes of the avian and human (if applicable) hosts for each subtype was analysed by categorizing the index sequences into the following categories based on its incidence.

i) Extremely diverse: this category contains index sequences with incidence less than 10% (index incidence <10%).

ii) Highly diverse: represents index sequences with incidence greater than or equal to 10% and less than 20% (10% ≤ index incidence < 20%).

iii) Mixed variable: represents index sequence with incidence greater than or equal to 20% and less than 90% (20% ≤ index incidence < 90%).

iv) Highly conserved: represents index sequence with incidence greater than or equal to 90% and less than 100% (90% ≤ index incidence < 100%).

v) Completely Conserved: represents index sequence with incidence exactly equal to 100% (index incidence = 100%).

For each subtype, this distribution was visualized in the form of a bar chart using matplotlib python library.

# 3. RESULTS

## 3.1. DATA

A total of 907086 viral protein sequences (human and avian) were downloaded for the H3N2 subtype, 35183 (human and avian) for H7N9 (human and avian) and 38303 (human and avian) for H4N6. The H3N2 subtype had the highest total number of sequences downloaded, while the H7N9 had the least total number of downloaded sequences. In terms of hosts, the H3N2 subtype again had the highest number of downloaded sequences of the human host, while the H7N9 recorded the least number of downloaded sequences of the human host. As per the methodology, sequences of human hosts were not downloaded for the H4N6 subtype. In the case of avian host however, across all the subtypes, H4N6 had the highest number of downloaded avian host sequences with H3N2 subtype recording the lowest number. The avian-to-human host ratio (A2H) of the number of downloaded sequences for H3N2 subtype was 0.01 and that of the H7N9 subtype was 1.30. (Table 3.1).

After the download, general filtering was done to remove sequences shorter than 50 amino acids (aa) as well as sequences containing unknown aa characters. This filtering was followed by deduplication of sequences eventually resulting in a relatively lower number of final sequences than the originally downloaded number of sequences. After filtering and deduplicating, the originally downloaded data (all subtypes and hosts inclusive) reduced by approximately 98.87%. In the remaining data, H3N2 still had the highest number of sequences (94994) while the H4N6 subtype had the lowest total number of sequences (7436) ready for analysis (Table 3.1).

**Table 3.1:** Number of H3N2, H7N9, and H4N6 sequences analysed.

| Subtype | Downloaded sequences | | | After filtering | | | After deduplication | | | Total analysed |
|---|---|---|---|---|---|---|---|---|---|---|
| | Human | Avian | Total | Human | Avian | Total | Human | Avian | Total | |
| **H3N2** | 896804 | 10282 | 907086 | 895636 | 10262 | 905898 | 91915 | 3079 | 94994 | 94994 |
| **H7N9** | 15284 | 19899 | 35183 | 15261 | 19881 | 35142 | 4133 | 4010 | 8143 | 8143 |
| **H4N6** | N/A | 38303 | 38303 | N/A | 38147 | 38147 | N/A | 7436 | 7436 | 7436 |

Sequences shorter than 50 amino acids (aa) and those containing unknown aa character "X" were removed to improve the ease and quality of multiple sequence alignment (MSA). Shorter sequences could match segments of several full-length(longer) sequences during alignment therefore making it more difficult for the MSA tool to decide the correct position of such short sequences. This increases the chances of the sequences being placed at a wrong position during the alignment. Likewise, Unknown characters ("X") are supposedly amino acids that were not able to be defined during sequencing of the samples in the laboratories. Because they are unknown to the MSA tool, they may end up introducing errors in the alignment process which could give wrong analysis results. Also, deduplication on the other hand removes duplicate sequences (sequences with exact same amino acids arrangements) from the data to ensure that all analysis efforts are not redundantly duplicated.

### 3.2. H3N2

The H3N2 proteome had a total of 1104 and 1120 9-mer positions of low support for viruses of human and avian hosts, respectively. The average (mean) entropy value ranged from ~ 0.28 (PB1) to ~2.51 (PB1-F2), with a proteome-wide mean entropy of ~0.56 for the human host. Separately, the avian host exhibited average (mean) entropy starting from ~0.35 (PB2) to ~3.80 (PB1-F2) and a proteome-wide average of ~0.67, relatively higher than viruses of the human host. The low average proteome entropy (< 1) indicated a high conservation level of H3N2. Correspondingly, the proteome-wide average total variants (summation of all diversity motifs except the index) was low of 9.12% (human host) and 12.05% (avian host) for the respective viruses. However, the incidences of the average individual protein total variants exhibited a broader spectrum, starting from as low as 3.96% (PB1) to 43.26% (PB1-F2) for the human host, while 5.47% (PB2) to 71.28% (PB1-F2) for the avian host.

The highest absolute entropy across the H3N2 proteome of the human host was ~4.6 at position 185 in the HA protein, having a high incidence of total variants as 81.99%. In contrast, the

highest absolute total variants incidence of 82.03% was at position 182, also in the HA protein, with an entropy of 4.53. The H3N2 viruses of the avian host exhibited a higher peak absolute entropy of ~5.41 at position 105 of the NA protein, with a much higher total variants incidence of 91.85%. Comparatively, the maximum absolute total variants incidence was 91.87% at position 106 in the same NA protein, with an entropy of 5.37. Using the Env protein of HIV-1 clade B as a benchmark, the peak absolute entropy of 9.2 and total variants of 98% was observed in the Env protein (Hu et al., 2013).

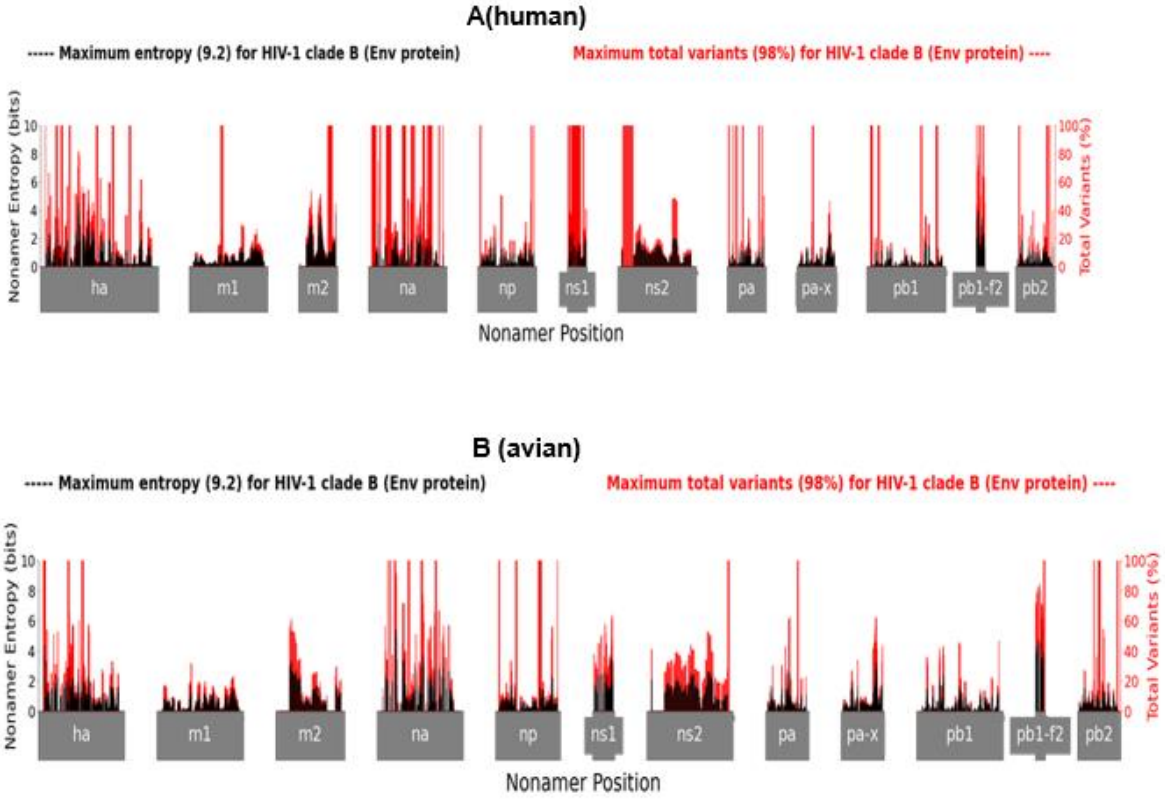


**Figure 3.1:** Shannon's entropy and total variants incidence for the 9-mer positions of each protein of the H3N2 subtype, isolated from human (top) and avian (bottom) host populations.

Proteomes of both hosts of the H3N2 subtype contained numerous positions with a total variants incidence of more than 40%. There were 30 positions with a total variants incidence of more than 60% in the human host (Figure 3.1 A**)**, while 100 positions in viruses of the avian host (Figure 3.1B). As observed, highly diverse 9-mer positions (total variants of ≥ 80%) were almost non-existent in both the avian and human proteomes occurring only in the avian host at 0.08 (human) and 0.38% (avian) of the positions.

The entropy and total variants showed a positive Pearson correlation coefficient (r) of 0.95 for the human host and 0.97 for the avian host (Figure 3.2). For both human and avian hosts of the H3N2 subtype, increase in the total variants causes direct increase in the range of entropies representing the diversity, this continued until a maximum of roughly 50% total variants was reached at which point the range of entropies began to fell although with a rising value. As observed, there were only 3 positions (human host) and 15 (avian host) with total variants higher than ~80%.

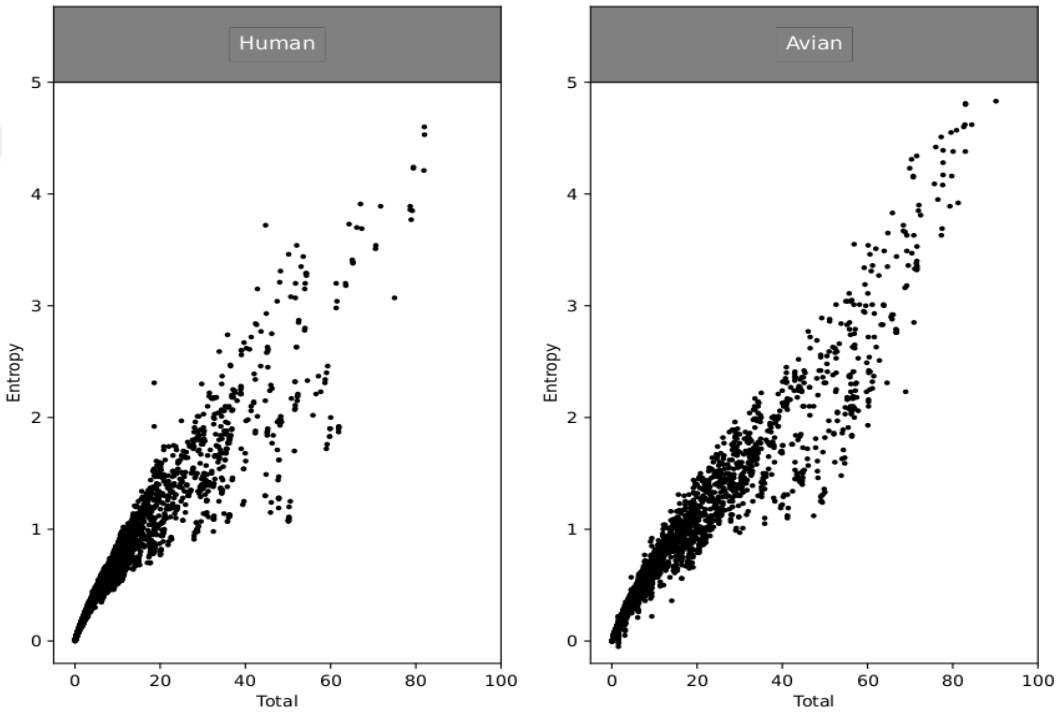


**Figure 3.2:** The relationship between entropy and incidence of total variants for H3N2 viruses of human (left) and avian (right) hosts.

Entropy of 1 was achieved at total variants as low as 11.72% (human host) and 12.04% (avian host). Entropy values of [0, 1, 2] and the highest of 4.6 were achieved for the human host at the total variants shown in Table 3.2, while entropy values of [0, 1, 2, 3] and the highest of 5.41 were achieved for the avian host proteome at the total variants shown in the Table 3.2.

**Table 3.2:** Minimum and maximum total variants at each entropy boundary values observed in both hosts of the H3N2 subtype.

| Entropy (bits) | Minimum total variants | | Maximum total variants | |
|---|---|---|---|---|
| | Human | Avian | Human | Avian |
| 0 | 0.01 | N/A | 0.06 | 0.3 |
| 1 | 11.72 | 12.04 | 28.6 | 23.44 |
| 2 | 32.53 | 35.05 | 60.01 | 56.91 |
| 3 | N/A | 59.6 | N/A | 63.84 |
| 4.6 | 81.99 | N/A | 81.99 | N/A |
| 5.41 | N/A | 91.85 | N/A | 91.85 |

Following the diversity motifs's definitions, in theory, each variant motif was expected to exhibit a characteristic pattern of incidence in relation to the total variants. Figures 3.3 and 3.4 plot the relationship between the incidence of each of the diversity motifs against a backdrop of increasing total variants for the proteome and individual proteins, respectively. Figures 3.5 and 3.6 provide frequency distribution of the diversity motifs incidence by use of violin plots for the proteome of both hosts and individual proteins, respectively.

Throughout the entire proteome, the index nonamer (9-mer) in general was the dominant motif, recording the highest average incidence of approximately 90.88% (human) and 87.95% (avian). The index is the inverse of the total variants, which comprises of the major variant, minor variants and unique variants. The theoretical minimum and maximum index incidence is 0% (inversely total variants of 100%) and 100% (inversely total variants of 0%).

By definition, ranking by the relative percentage incidence, the major variant is the motif that follows immediately after the index motif, and its incidence cannot go beyond the incidence of the index sequence, and thus, by taking the maximum incidence of the major variant into account, it is theoretically expected to exhibit an incidence pattern that follows the shape of a pyramid Due to this limitation created by the incidence of the index motif, the maximum incidence attainable by the major variant occurs at 50% total variants.. Before reaching the peak

incidence, there can be continuous increase in the incidence of both the major and total variants motifs while the index incidence reduces, but, after the peak incidence, the incidence of the major variant begins to fall along with the already decreasing incidence of the index.. The H3N2 major variant pattern for each host was near-pyramidal with a peak incidence of 49.37% (human) and 46.08% (avian) (Figure 3.3 A and B).

Nonamer sequences that occur more than once but are not as prevalent as the major variants are termed as minor variants. That is, the minor variant is the third most common sequence at a nonamer position. The human host proteome's minor variants motif represented a notable motif (Figure 3.3 C**)**, comparable to the major variant for total variants less than 50%. In the avian proteome, the minor variants motif represented the dominant diversity motif, particularly for positions whose total variants is ≥ 50% (Figure 3.3 D).

Unique variants (Figure 3.3 C and D) is the category of sequences that form the remainder of the variant sequence population. They represent nonamers that occurred only once at a position in the entire recorded sequence population. For both hosts, as the total variants increased a progressive increase in the unique variant incidence was observed until it reached a maximum of, ~16.67% (human) and 14.98% (avian) incidence was reached for the subtype.

These variant types (unique, minor, major) were mostly of low (<10%) incidence for both hosts, as illustrated in (Figure 3.3 A, B, C and D). However, the minor variant showed an extent of exception by exhibiting a continuous general increase in incidence and even going beyond the incidence of the index at positions where total variant incidence is more than 50%. Despite this observation the number of positions displaying this pattern of higher incidence beyond the index motif was minimal for the subtype.

Nonatypes (Figure 3.3 E and F), is a percentage measure of how many variant nonamers occurred at a position assuming all duplicated nonamers were removed. Nonatypes were observed for almost all (~99.97%, human; ~91.01%, avian) positions of the H3N2 nonamers. The human host H3N2 viruses exhibited nonatypes (Figure 3.3E) of generally low incidence, with a proteome-wide average of 0.74% and the highest incidence was 25% at position 50, corresponding to a total variants incidence of 75.0% (principally consisted of unique and minor variants). Likewise, for the avian host proteome, nonatypes (Figure 3.3F), generally had low incidences, and the average nonatype incidence across the entire proteome was 2.92% while

the highest individual incidence was 26.72% at position 29, corresponding to a total variants incidence of 83.0% (also majorly consisted of different unique and minor variants).

The diversity motif dynamics pattern at the proteome level (Figures 3.3 and 3.5) was generally consistent with that of the individual protein level (Figures 3.4 and 3.6), with distinctive features of each protein highlighted by the plots. The proteins with the closest proximity to the proteome pattern are the HA and NA in both hosts. These proteins showed the largest range of index incidence (Figures 3.4 and 3.6). On the other hand, a couple of proteins showed a relatively less resemblance to the general diversity dynamics pattern of their respective proteomes. These include M1, PA, PB1, PB2 for the human host and M1 for the avian host. However, the protein with the least resemblance to the proteome pattern is the avian PB1-F2, having data points only at a total variants range of ~55% to 84%.

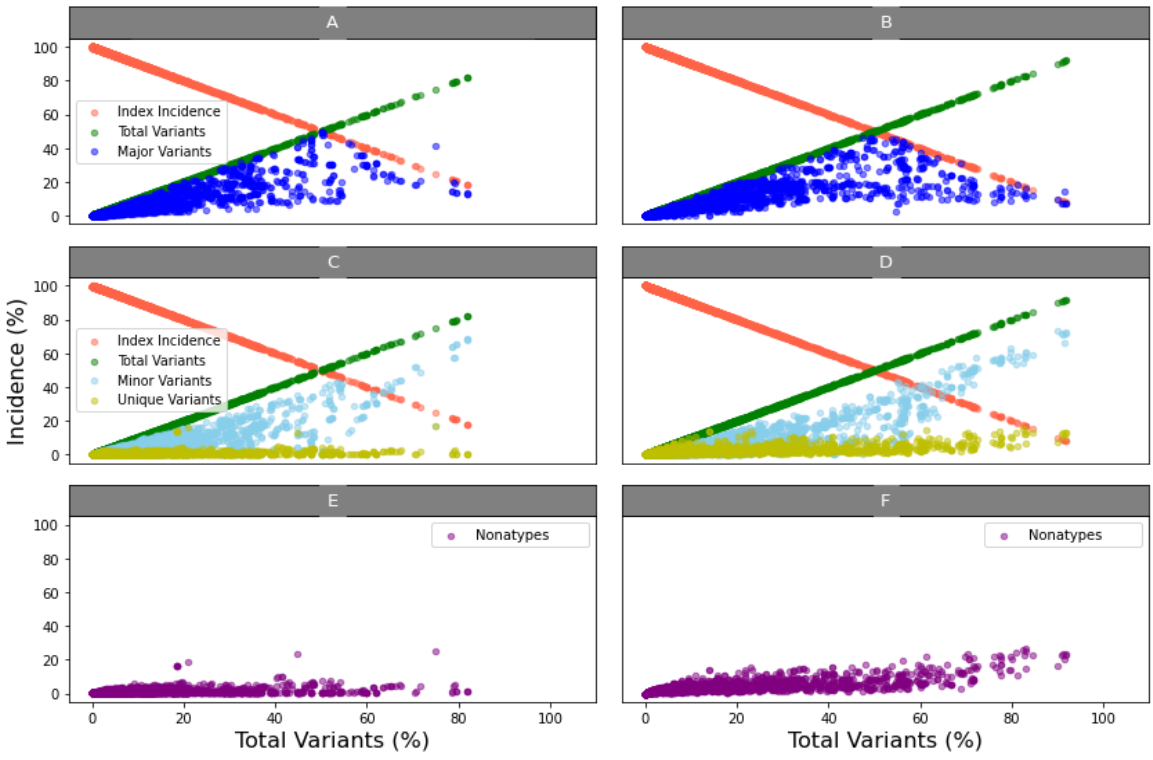


**Figure 3.3:** Proteome-based diversity motif patterns in the H3N2 IAV subtype of human(left) and avian (right) hosts.

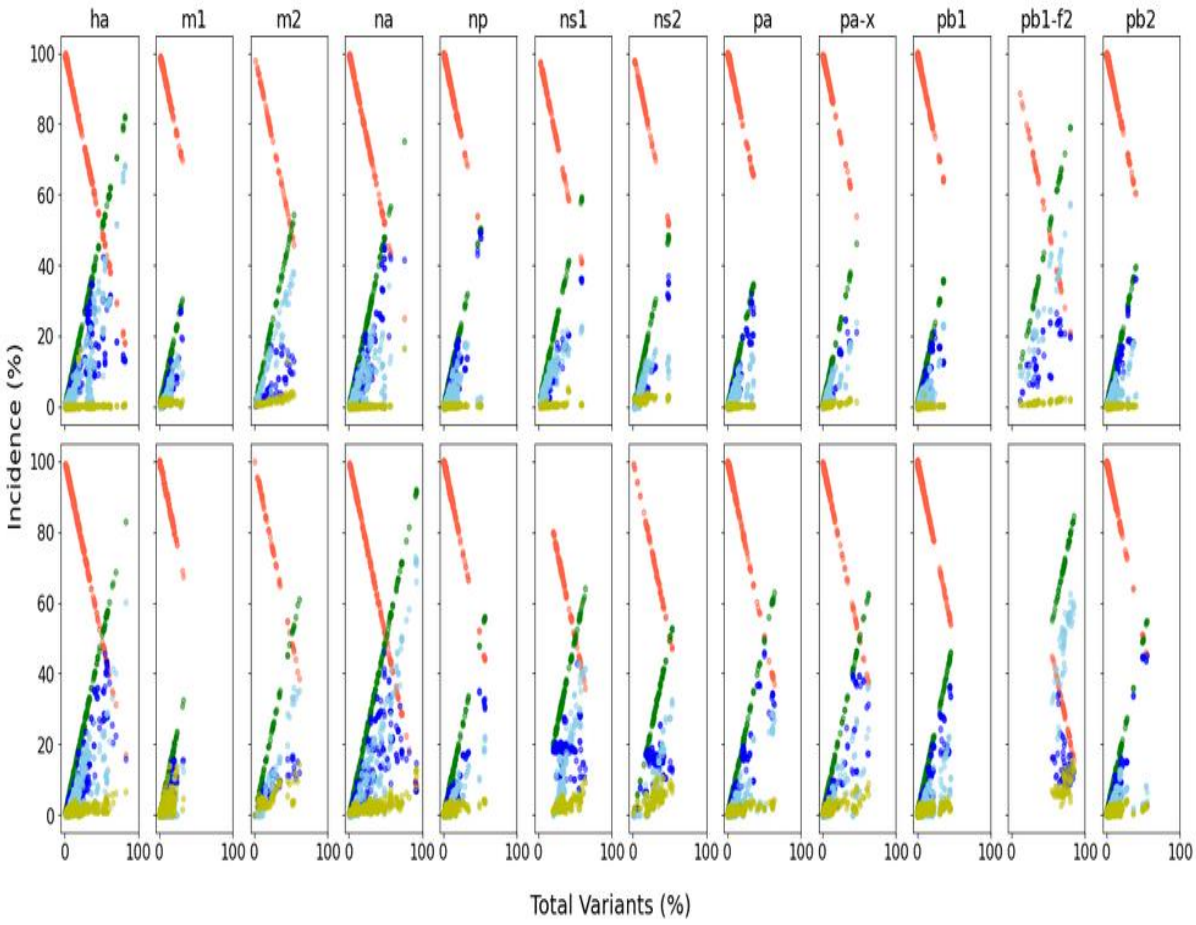


**Figure 3.4:** Protein-based diversity motif patterns in the H3N2 IAV subtype of human (top) and avian (bottom) hosts.

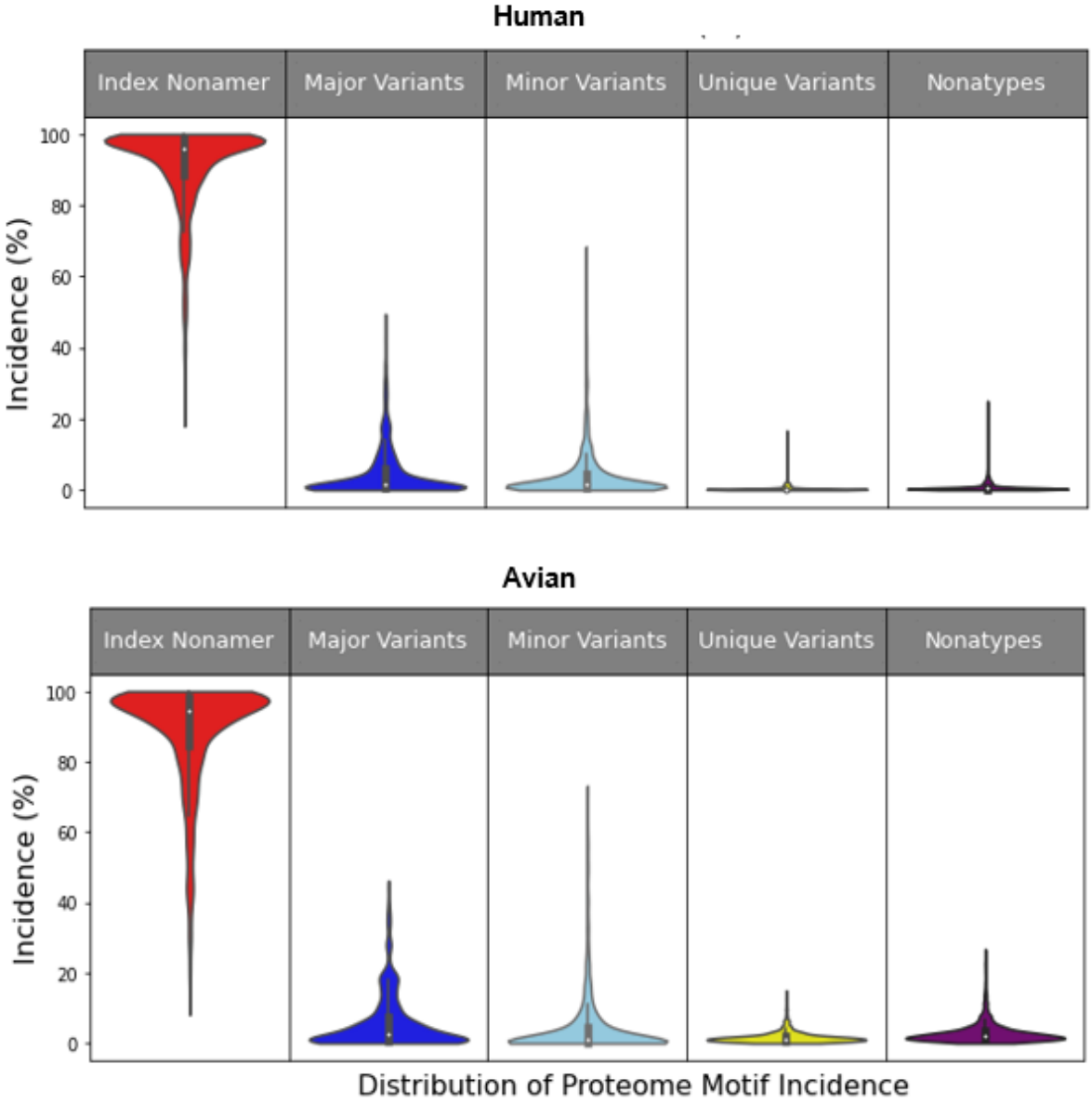


**Figure 3.5:** Diversity motif frequency distribution for the H3N2 IAV proteome of viruses isolated from human (top) and avian (bottom) hosts.

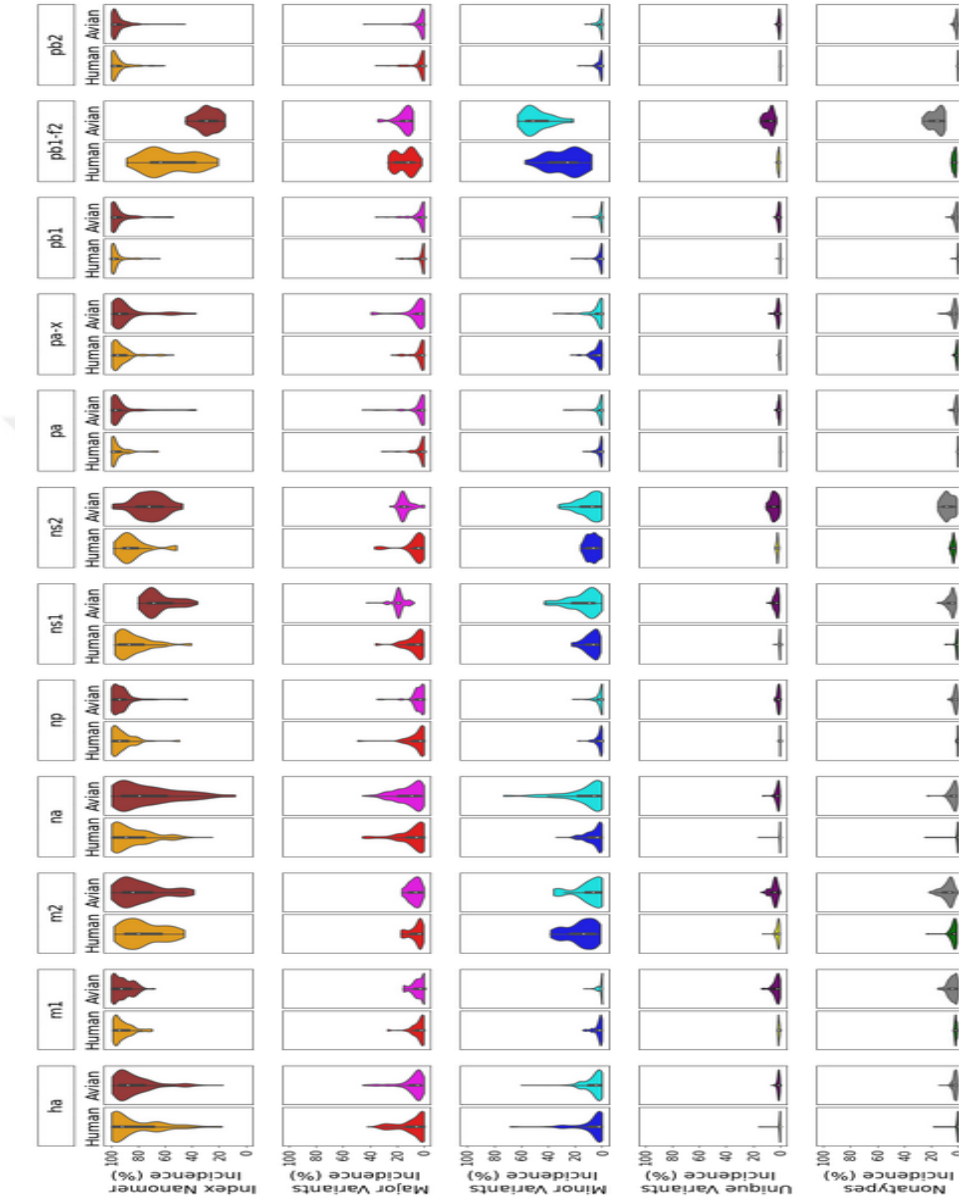


**Figure 3.6:** Diversity motif frequency distribution for the H3N2 IAV individual proteins of viruses isolated from human and avian hosts.

The Figure 3.7 provide a distribution of index sequences across various conservation thresholds for H3N2 viruses of human host. None of the positions were completely conserved, although

majority (71.15%) of the positions were highly conserved (90% ≤ index incidence < 100%). The second largest percentage of the positions (28.78%) constituted the mixed variable category (20% ⩽ index incidence < 90%). Only 0.69% of the positions were highly diversed and this was contributed by just the HA protein. However, none of the positions in the H3N2 of human host were extremely diversed. Likewise, Figure 3.8 provides the conservation distribution for H3N2 viruses of the avian host. Unlike the human host proteome, the avian host proteome had some percentage of positions for all the five levels of conservation. These are: completely conserved (2.82%), highly conserved (63.38%), mixed variable (33.43%), highly diversed (0.23%) and extremely diversed (0.15%). In the case of extremely diversed, it was contributed by only one protein (NA, 2.14%).

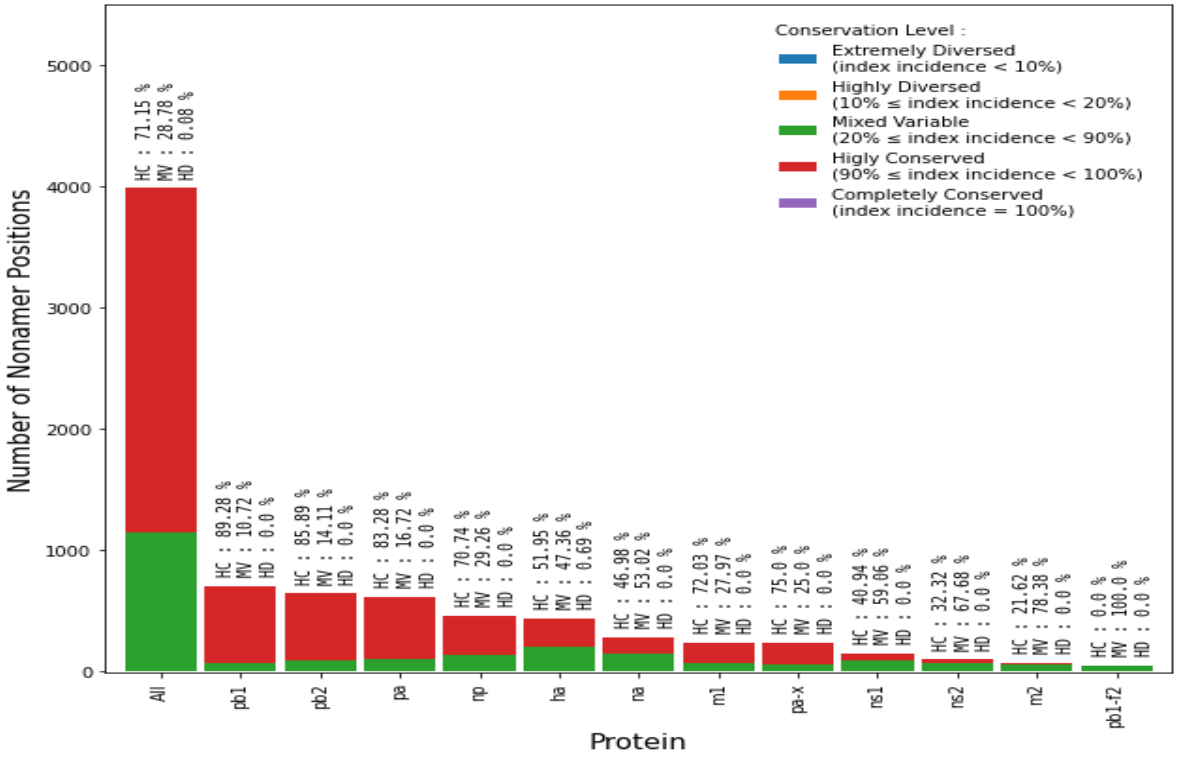


**Figure 3.7:** Distribution of index incidences across various conservation levels for all the nonamer positions of the H3N2 viruses of the human host. All (proteome) refers to the combined distributions of the individual proteins.

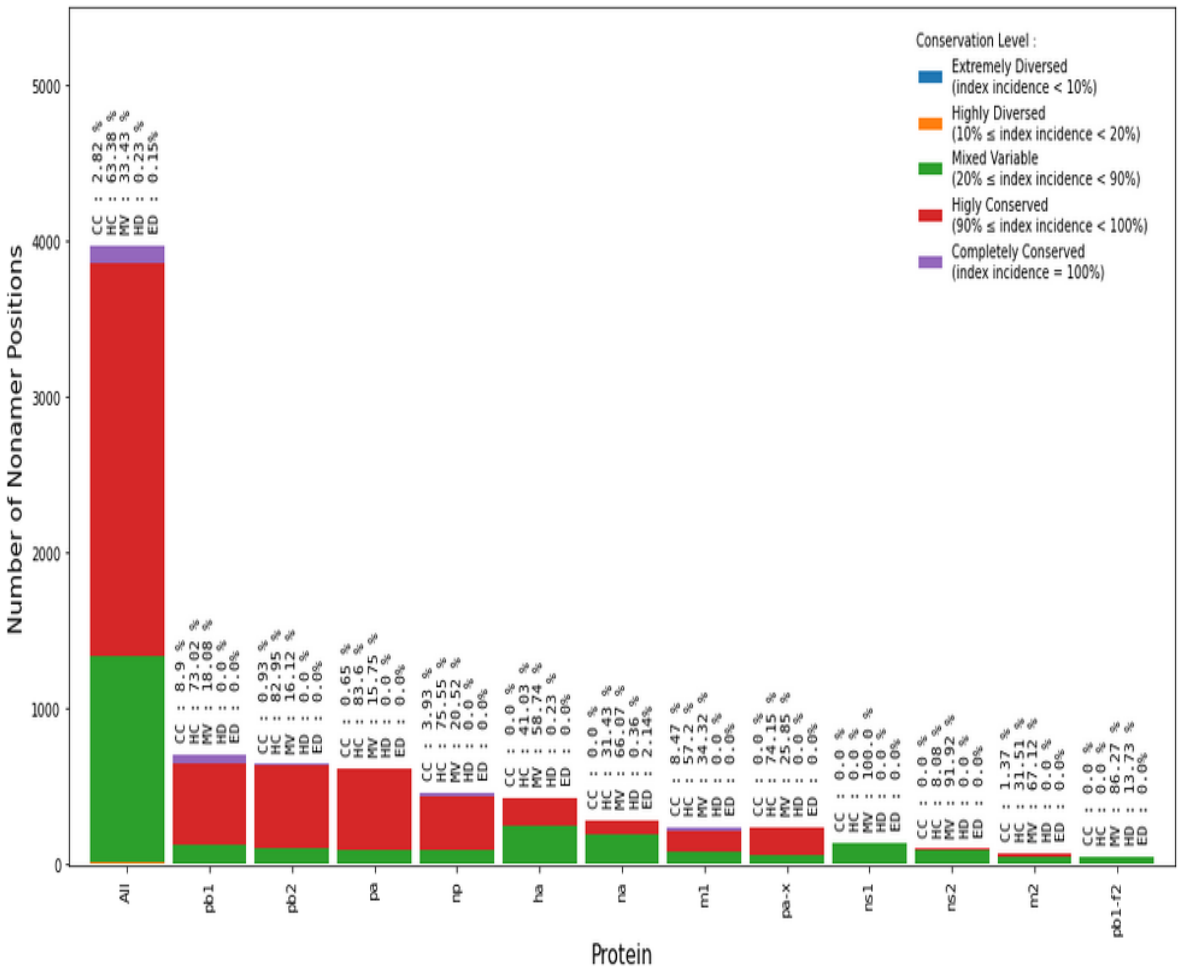


**Figure 3.8:** Distribution of index incidences across various conservation levels for all the nonamer positions of the H3N2 viruses of the avian host. All (proteome) refers to the combined distributions of the individual proteins.

### 3.3. H7N9

The H7N9 proteome had a total of 387 and 252 9-mer positions of low support viruses of human and avian hosts, respectively. The average (mean) entropy value ranged from ~0.29 (PA) to ~1.95 (PB1-F2) having a proteome-wide mean entropy of ~0.43 for the human host. Separately, the avian host exhibited average (mean) entropy starting from ~0.37 (PB1) to ~2.77 (PB1-F2) and a proteome-wide mean of ~0.67, relatively higher than viruses of the human host. Like the H3N2 subtype, the low average proteome entropy (< 1) indicated a high conservation level of H7N9.

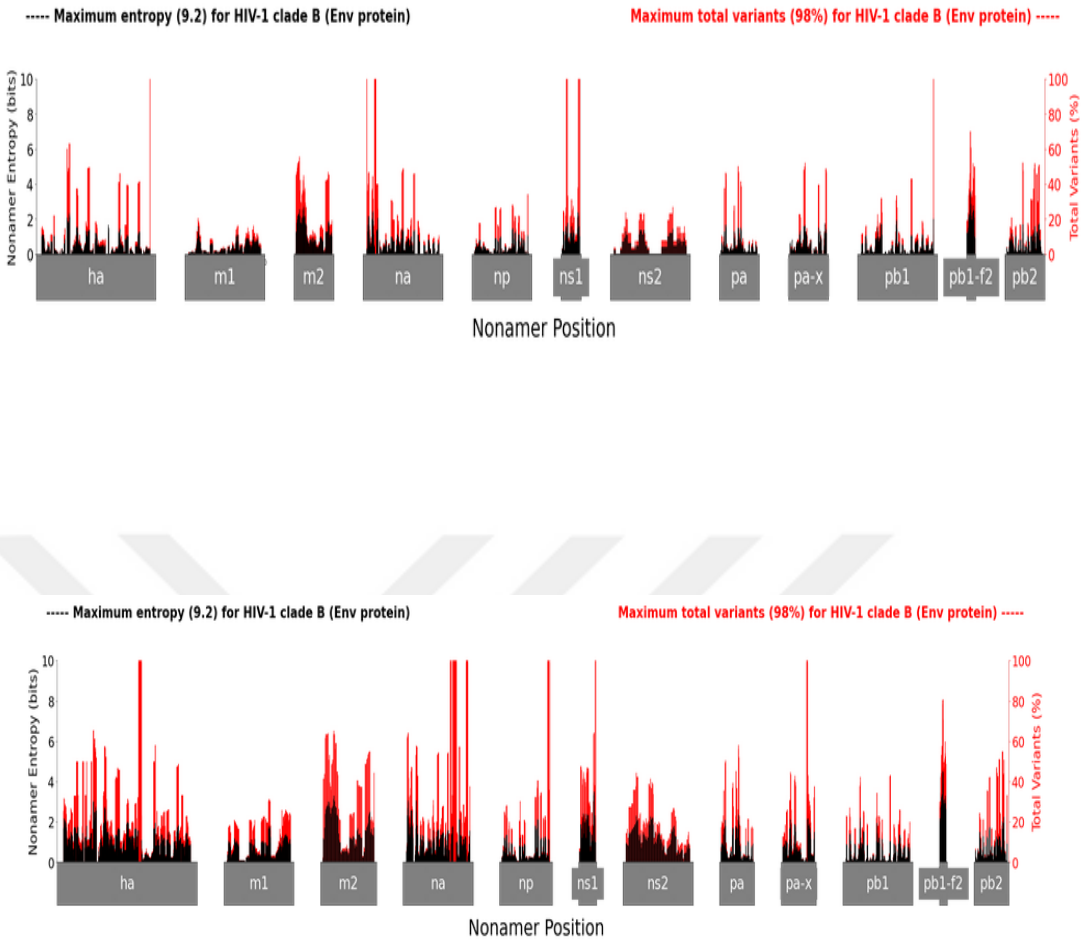


**Figure 3.9:** Shannon's entropy and total variants incidence for the 9-mer positions of each protein of the H7N9 subtype isolated from human (top) and avian (bottom) host populations.

Correspondingly, the proteome-wide average total variants (summation of all diversity motifs except the index) was low of 7.84% (human host) and 12.36% (avian host) for the respective viruses. However, the incidences of the average individual protein total variants exhibited a broader spectrum, starting from as low as 4.67% (PB1) to 34.07% (PB1-F2) for the human host, while 6.61% (PB1) to 46.77% (PB1-F2) for the avian host.

The highest absolute entropy across the H7N9 proteome of the human host was ~3.57 at position 41 in the PB1-F2 protein, with a total variants incidence of 69.96%, and this (69.96%) was, as well, the highest total variant incidence for the H7N9 human host proteome. The H7N9 viruses of the avian host recorded a relatively higher maximum absolute entropy of ~4.54 at position 42 in the PB1-F2 protein, and a total variants incidence of 80.33%. However, the avian

proteome's highest absolute total variants incidence, 80.74% was at position 42 also in the PB1-F2 protein, with an entropy of 4.47. Using the Env protein of HIV-1 clade B as a benchmark, the peak absolute entropy of 9.2 and total variants of 98% was observed in the Env protein (Hu et al., 2013).

Proteomes of both hosts of the H7N9 subtype also had several positions with a total variants incidence of more than 40%, and 11 positions with a total variants incidence of more than 60% (Figure 3.9A) in the human host, while 37 positions in the viruses of the avian host **(**Figure 3.9 B). As observed, highly diverse 9-mer positions (total variants ≥ 80%) were nearly non-existent in both the avian and human proteomes occurring only in the avian host at 0.04% of the positions.

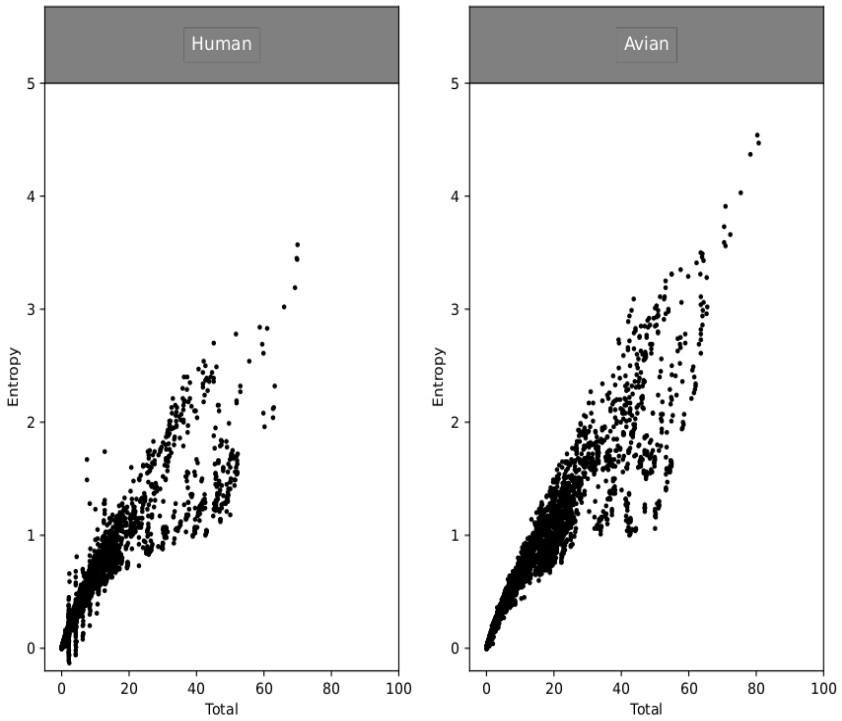


**Figure 3.10:** Human (left) and Avian (right) host visualisations of the relationship between entropy and incidence of total variants for the H7N9 IAV.

The entropy and total variants showed a positive Pearson correlation coefficient (r) of 0.92 for the human host and 0.95 for the avian host (Figure 3.10). For both human and avian hosts of the H7N9 subtype, just like it was in the H3N2, increase in the total variants causes direct increase in the range of entropies representing the diversity, this continued until a maximum of roughly 50% total variants was reached at which point the range of entropies began to fell although with a rising value. As observed, there were no (0) positions (human host) and 2 positions (avian host) with total variants incidence higher than ~80% in respective proteomes of both hosts.

Entropy of 1 was achieved at total variants as low as 14.34 (human host) and 13.89 (avian host) in both hosts. Entropy values of [0, 1, 2] and the highest of 3.57 were achieved for the human host at the total variants as shown in Table 3.3 while entropy values of [0, 1, 2, 3, 4] and the highest of 4.54 were achieved for the avian host proteome at the total variants shown in Table 3.3.

**Table 3.3:** Minimum and maximum total variants at each entropy boundary values observed in both hosts of the H7N9 subtype.

| Entropy (bits) | Minimum total variant incidence | | Maximum total variant incidence | |
|---|---|---|---|---|
| | Human | Avian | Human | Avian |
| 0 | 0.0 | 0.0 | 2.08 | 0.21 |
| 1 | 14.34 | 13.89 | 42.64 | 42.4 |
| 2 | NA | 39.83 | 36.14 | 58.29 |
| 3.57 | 69.96 | 53.95 | 69.96 | 53.95 |
| 4.54 | NA | 80.33 | NA | 80.33 |

Again, Following the diversity motifs's definitions, in theory, each variant motif was expected to exhibit a characteristic pattern of incidence in relation to the total variants. Figures 3.11 and 3.12 plot the relationship between the incidence of each of the diversity motifs against a backdrop of increasing total variants for the proteome and individual proteins, respectively. By

use of violine plots, Figures 3.13 and 3.14 provide frequency distribution of the incidences of each diversity motif for the proteome of both host and individual proteins, respectively.

Throughout the entire proteome, the index nonamer (9-mer) was the dominant motif, recording the highest average incidence of approximately 92.16% (human) and 87.64% (avian). By definition and ranking by the relative percentage incidence, the major variant is the motif that follows immediately after the index motif, and its incidence cannot go beyond the incidence of the index sequence, and thus, by taking the maximum incidence of the major variant into account, it is theoretically expected to exhibit an incidence pattern that follows the shape of a pyramid. Due to this limitation created by the incidence of the index motif, the maximum incidence attainable by the major variant occurs at 50% total variants. Before reaching this peak incidence, there can be continuous increase in the incidence of both the major and total variants motifs while the index incidence reduces, but, after the peak incidence, the incidence of the major variant begins to fall along with the already reducing incidence of the index. The H7N9 major variant pattern for each host was near-pyramidal with a peak incidence of 45.83% (human) and 48.98% (avian) (Figure 3.11 A and B).

Nonamer sequences that occur more than once but are not as prevalent as the major variants are termed as minor variants. That is, the minor variant is the third most common sequence at a nonamer position. For positions with >= 50% total variants, the minor variants motif was relatively less significant in the human host proteome as compared to that of the avian host of the subtype (Figure 3.11 C and D) . Unique variants is the category of sequences that form the remainder of the variant sequence population. It represents nonamers that occurred only once at a position in the entire recorded sequence population. For both hosts, the incidence of unique variants (Figure 3.11 C and D) progressively increased, as total variants increased until it reached a maximum of ~12.78% and 14.75% for the subtype.

These variant types (major, minor and unique) were mostly of low (<10%) incidence for both hosts, as illustrated in (Figure 3.11 A, B, C and D). However, the minor variant showed an extent of exception by exhibiting a continuous general increase in incidence and even going beyond the incidence of the index at positions where total variant incidence is more than 50%. Despite this observation the number of positions displaying this pattern of higher incidence beyond the index motif was minimal for the subtype. Nonatypes (Figure 3.11 E and F), is a

percentage measure of how many variant nonamers occurred at a position assuming all duplicated nonamers were removed.

Nonatypes were observed for most (~88.04%, human; ~94.45%, avian) of the H7N9 nonamer positions. The human host H7N9 viruses exhibited nonatypes (Figure 3.11 E) of generally low incidence with a proteome-wide average of 1.72% with the highest incidence of 18.8% at position 20 corresponding to total variants of 45.11% (principally consisted of unique and minor variants). Likewise, for the avian host proteome, nonatypes (Figure 3.11 F generally had low incidences, and the average nonatype incidence across the entire proteome was 2.45% with the highest incidence of 25.41% at position 42, corresponding to total variants of 80.74% (also majorly comprised of different unique and minor variants).

The diversity motif dynamics pattern at the proteome level (Figures 3.11 and 3.13) was generally consistent at the individual protein level (Figures 3.12 and 3.14), with distinctive features of each protein highlighted by the plots. The proteins in the closest proximity to its proteome patterns are the HA (human and avian hosts), and NA (avian host) of the H7N9 subtype. These proteins showed the largest range of index incidence (Figures 3.12 and 3.14). On the other hand, a couple of proteins showed less proximity to the general diversity dynamics pattern of their respective proteomes. These include M1, NP, NS1, and PB1 for the human host and M1, NP, NS2, PA-X, PB1 for the avian host. However, the protein with the least resemblance to its proteome motif diversity dynamics pattern is the human host NS2 protein, having absolutely no data points at all. This protein had very few sequences for human and avian hosts in the databases used and thus all the positions had less than 30 support (default support threshold in DIMA) so they were automatically filtered out in this plot.

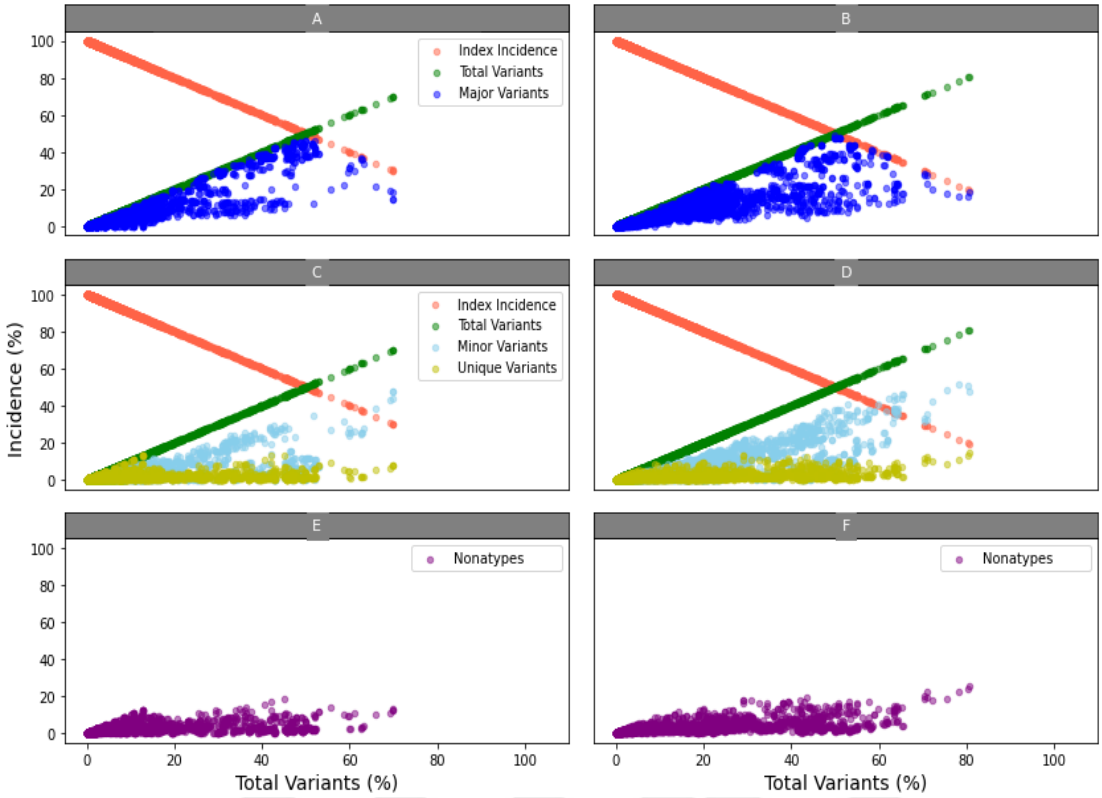


**Figure 3.11:** Proteome-based diversity motif patterns for the H7N9 IAV subtype of human (left) and avian (right) hosts.

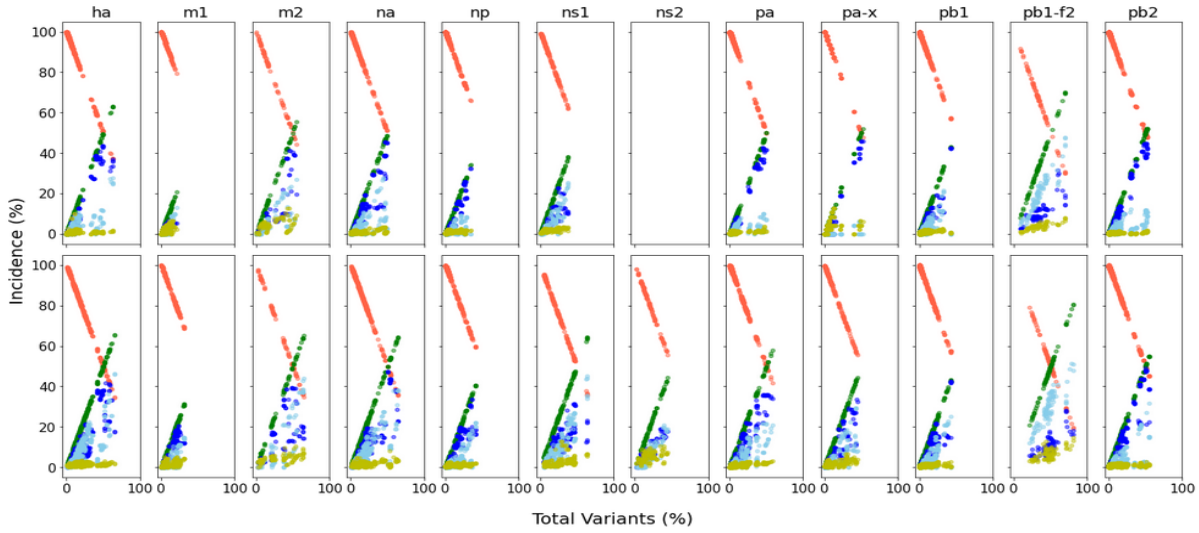


**Figure 3.12:** Protein-based diversity motif patterns in the H7N9 IAV subtype of human host (top) and avian host (bottom).

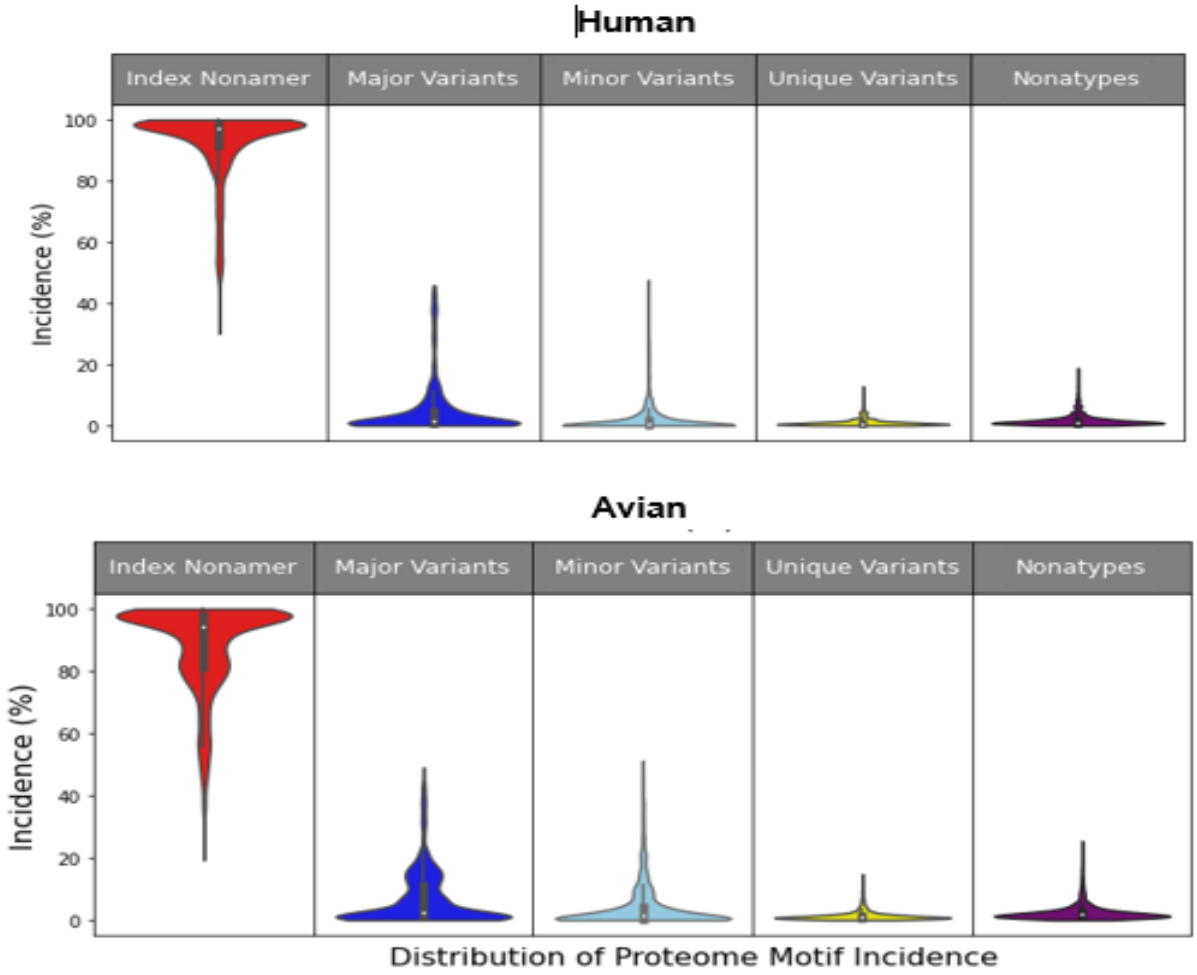


**Figure 3.13:** Diversity motif frequency distribution for the H7N9 IAV proteome of viruses isolated from human (top) and avian (bottom) hosts.

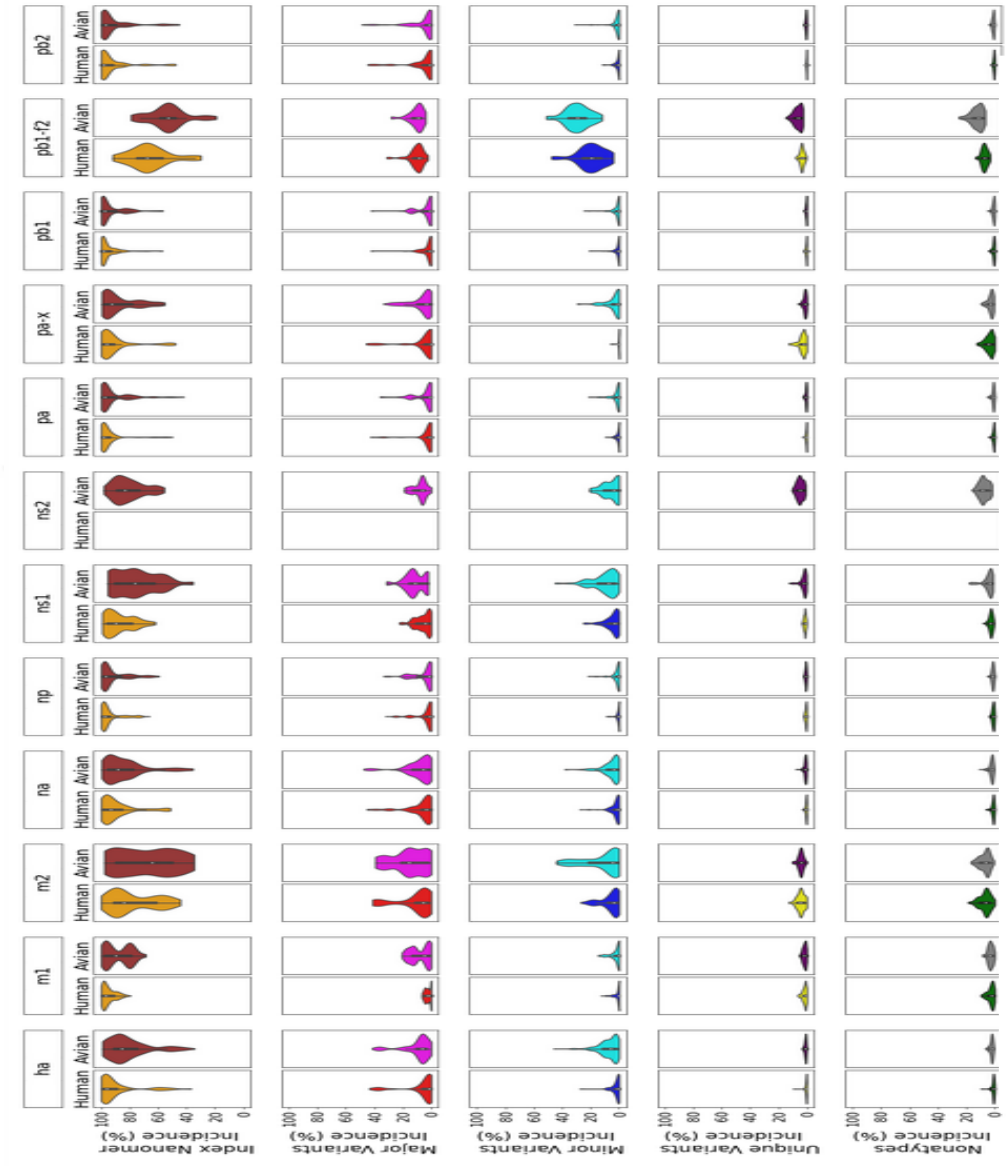


**Figure 3.14:** Diversity motif frequency distribution for the H7N9 IAV individual proteins of viruses isolated from human and avian hosts.

The Figure 3.15 provides a distribution of index sequences across various conservation thresholds for H7N9 viruses of human host. Although a greater percentage (~74.25%) of the positions were Highly conserved (90% ≤ index incidence < 100%) only a few (2.81%) of them were completely conserved (index incidence = 100%). The second largest percentage of the positions (22.94%) fell in the Mixed variable (20% ≤ index incidence < 90%) conservation level. However, none of the positions were highly diversed nor extremely diversed. Likewise, Figure 3.16 provides the conservation distribution for H7N9 viruses of the avian host. the avian host proteome had a slightly different distribution of conservation levels as compared to the human

host. About half of the positions (57.79%) were highly conserved (Figure 3.16), and a bit below half of them (41.08%) fell in the mixed variable conservation level. There were 1.08% completely conserved positions coming from a couple of proteins, however, in the case of highly conserved positions, only one protein (PB1-F2) had positions in the highly conserved conservation level forming about 0.04 % of all the positions of the proteome.

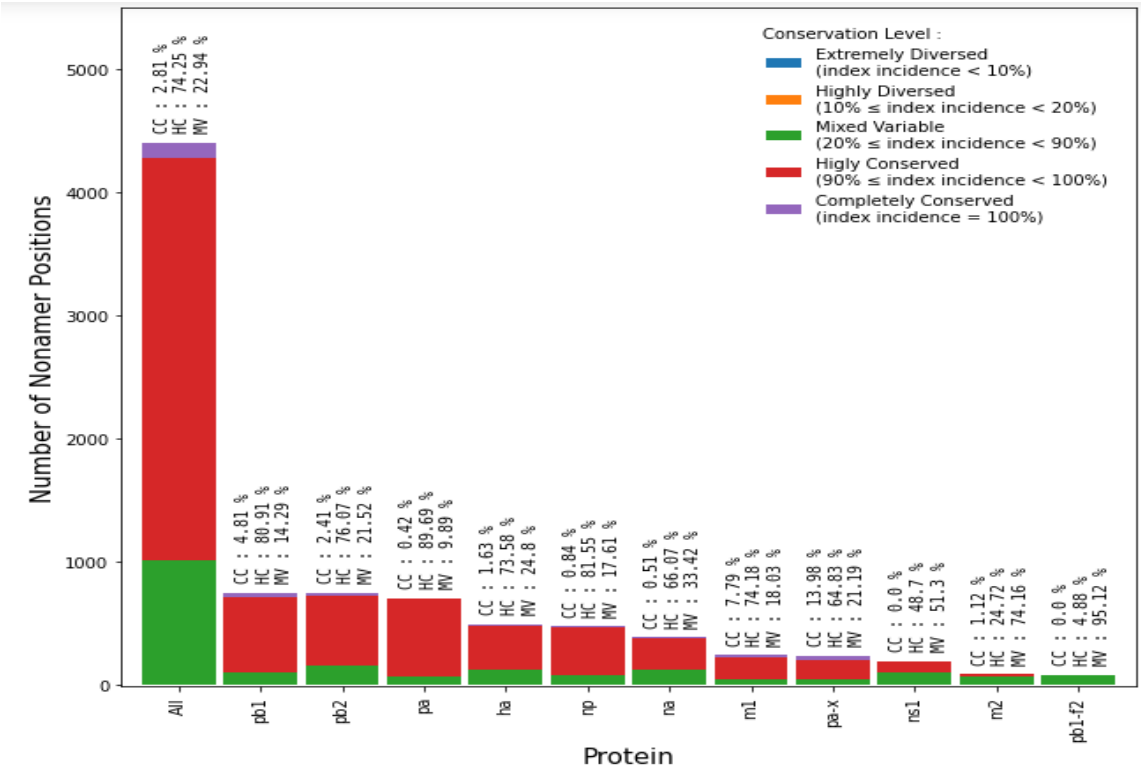


**Figure 3.15:** Distribution of index incidences across various conservation levels for all the nonamer positions of the H7N9 viruses of the human host. All (proteome) refers to the combined distributions of the individual proteins.

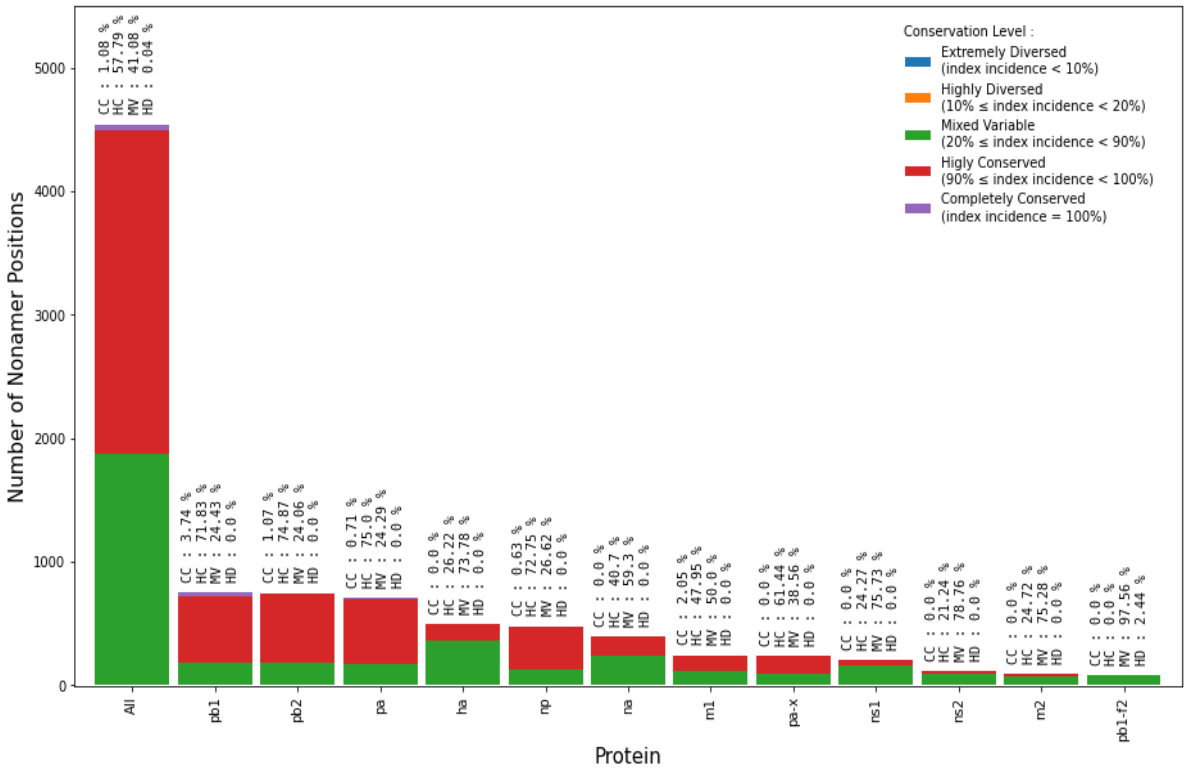


**Figure 3.16:** Distribution of index incidences across various conservation levels for all the nonamer positions of the H7N9 viruses of the avian host. All (proteome) refers to the combined distributions of the individual proteins.

### 3.4. H4N6

As explained earlier, the H4N6 subtype was classified as “Only avian”, and therefore only the H4N6 avian host proteome was analysed. The H4N6 avian host proteome had a total of 189 9-mer positions of low support, with approximately low average (mean) entropy value starting from ~0.26 (PB1) to ~3.54 (PB1-F2), with a proteome-wide mean entropy of ~0.59. Similar to H3N2 and H7N9, the low average proteome entropy (< 1) indicated a high conservation level of H4N6. Correspondingly, the proteome-wide average total variants (summation of all diversity motifs except the index) was low of 11.87%. for viruses of H4N6 avian host. However, the incidences of the average individual protein total variants exhibited a broader spectrum, starting from as low as 4.05% (PB2) to 66.23% (PB1-F2).

The highest absolute entropy across the H4N6 proteome of avian host was ~4.92 at position 36 in the PB1-F2 protein, having a high incidence of total variants as 85.47%. In contrast, the

highest absolute total variants incidence of 86.45%, was at position 35 also in the PB1-F2 protein, with an entropy of 4.9. Using the Env protein of HIV-1 clade B as a benchmark, the peak absolute entropy of 9.2 and total variants of 98% was observed in the protein (Hu et al., 2013).

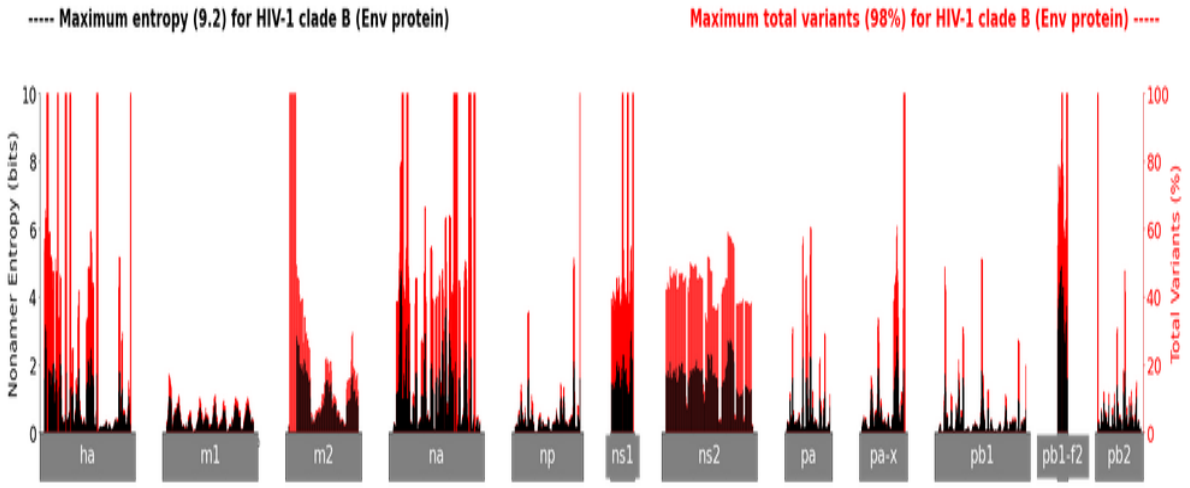


**Figure 3.17:** Shannon's entropy and total variants incidence for the 9-mer positions of each protein of the H4N6 subtype, isolated from human (top) and avian (bottom) host populations.

The avian host H4N6 proteome contained, several positions with a total variants incidence of more than 40%. There were 97 positions with a total variants incidence of more than 60% (Figure 3.17). As observed, highly diverse 9-mer positions (total variants of ≥ 80%) were nearly non-existent in the proteome (0.07% of the positions).

The entropy and total variants showed a positive Pearson correlation coefficient (r) of 0.96 for the H4N6 avian host (Figure 3.18). As total variants increased, the range of possible entropy values describing the diversity of the proteome also increased, with an inflection at approximately 50% total variants, where the entropy range started decreasing, albeit with an increasing value. Notably, there were only 3 positions with total variants higher than ~80%.

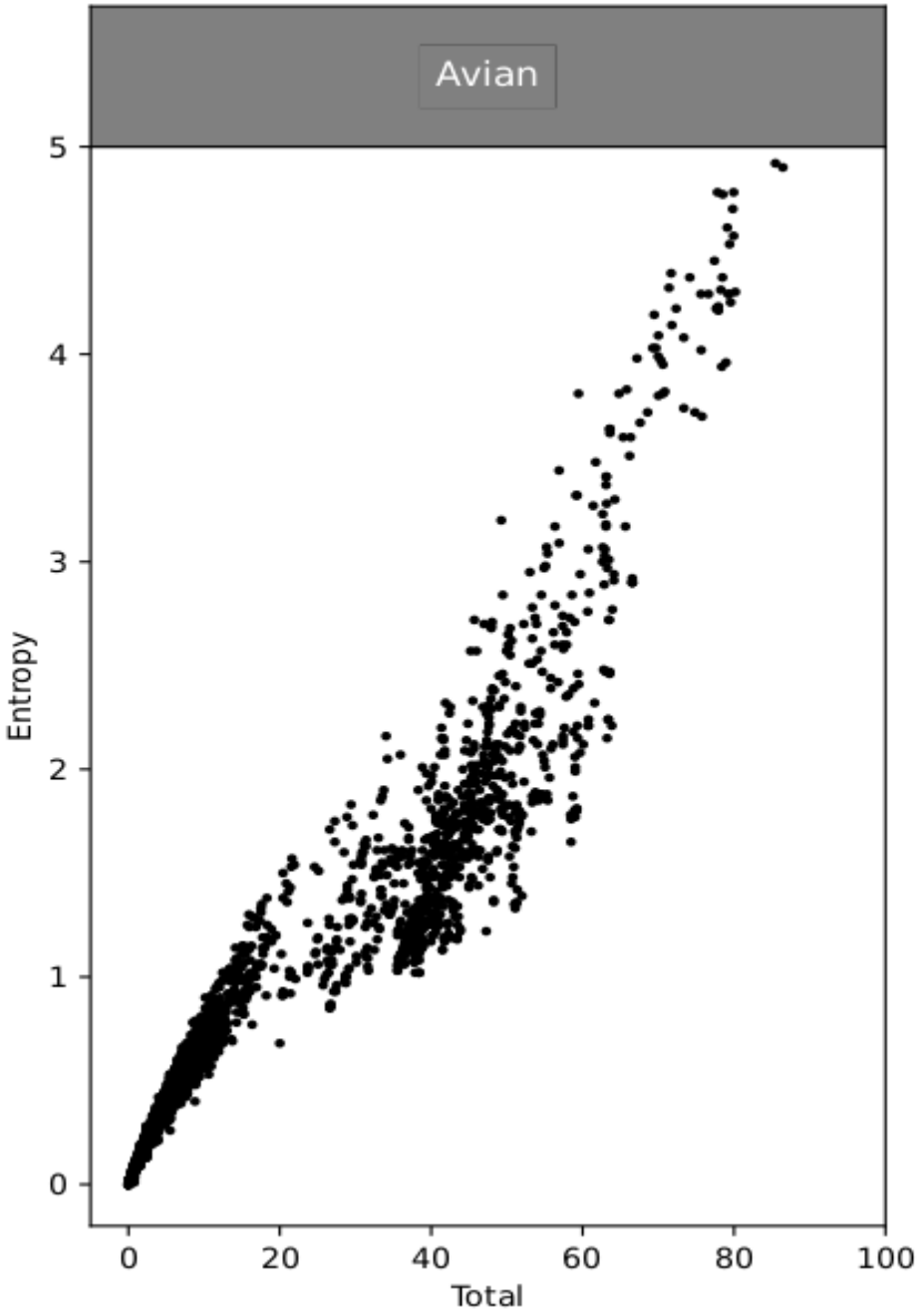


**Figure 3.18:** The relationship between entropy and incidence of avian host total variants for H4N6 subtype.

Entropy of 1 was achieved at total variants as low as 13.73%. Entropy values of [1, 2, 3] and the highest of 4.92 were achieved for the H4N6 avian proteome at the total variants shown in Table 3.4.

**Table 3.4:** Minimum and maximum total variants at each entropy boundary values observed in avian host of H4N6 subtype.

| Entropy (bits) | Minimum Total Variants (%) | Maximum Total Variants (%) |
|---|---|---|
| **1** | **13.73** | **28.7** |
| **2** | **44.2** | **45.63** |
| **3** | **62.63** | **62.63** |
| **4.92** | **85.47** | **85.47** |

Following the diversity motifs's definitions, in theory, each variant motif was expected to exhibit a characteristic pattern of incidence in relation to the total variants. Figures 3.19 and 3.20 plot the relationship between the incidence of each of the diversity motifs against a backdrop of increasing total variants for the proteome and individual proteins, respectively. By use of violine plots, Figures 3.21 and 3.22 provide frequency distribution of the incidences of each diversity motif for the proteome and individual proteins, respectively

Just like the H3n2 and H7N9, throughout the entire proteome of the H4N6, the index nonamer (9-mer) was the dominant motif, recording the highest average incidence of approximately 81.13% (avian). By definition and ranking by the relative percentage incidence, the major variant is the motif that follows immediately after the index motif, and its incidence cannot go beyond the incidence of the index sequence, and thus, by taking the maximum incidence of the major variant into account, it is theoretically expected to exhibit an incidence pattern that follows the shape of a pyramid. The H4N6 major variant pattern for the avian host near-pyramidal with a peak incidence of 47.31% (Figure 3.19 A)

Nonamer sequences that occur more than once but are not as prevalent as the major variants are termed as minor variants. That is, the minor variant is the third most common sequence at a nonamer position. The minor variants motif of the avian host proteome represented the

dominant diversity motif, particularly for positions whose total variants was ≥ 50% (Figure 3.19 B).

Unique variants (Figure 3.19 B) is the nonamer category of sequences that form the remainder of the variant sequence population. They represent nonamers that occurred only once at a position in the entire recorded sequence population. For avian hosts, as the total variants increased a progressive increase in the unique variant incidence was observed until it reached a maximum of ~10.16% across the avian host proteome.

These variant types (unique, minor, major), were mostly of low (<10%) incidence, as illustrated in (Figure 3.19 –A, B and C). However, the minor variant showed an extent of exception by exhibiting a continuous general increase in incidence and even going beyond the incidence of the index at positions where total variant incidence is more than 50%. Despite this observation the number of positions displaying this pattern of higher incidence beyond the index motif was minimal for the proteome.

Nonatypes (Figure 3.19 C), is a percentage measure of how many variant nonamers occurred at a position assuming all duplicated nonamers were removed. Nonatypes were observed for almost all (~97.03%) of the H4N6 avian host nonamer positions. The proteome exhibited nonatypes (Figure 3.19 C) of generally low incidence, with a proteome-wide average of ~1.81% and the highest incidence of 15.71% at position 44 corresponding to total variants of 79.83% (principally consisted of different unique and minor variants).

The diversity motif dynamics pattern at the proteome level (Figures 3.19 and 3.21) was generally consistent at the individual protein level (Figures 3.20 and 3.22), with distinctive features of each protein highlighted by the plots. The proteins with the closest proximity to the proteome pattern are the HA and NA proteins. These proteins showed the largest range of index incidence (Figures 3.21 and Figure 3.22). On the other hand, a couple of proteins showed a relatively less resemblance to the general diversity dynamics pattern of their respective proteomes. These include M1, M2, NP, PB1 and PB2. However, the protein with the least resemblance to and relatively significant divergence from its proteome motif diversity dynamics pattern is the avian PB1-F2 protein, having data points only at a total variants range of ~48% to ~86% (Figure 3.20).

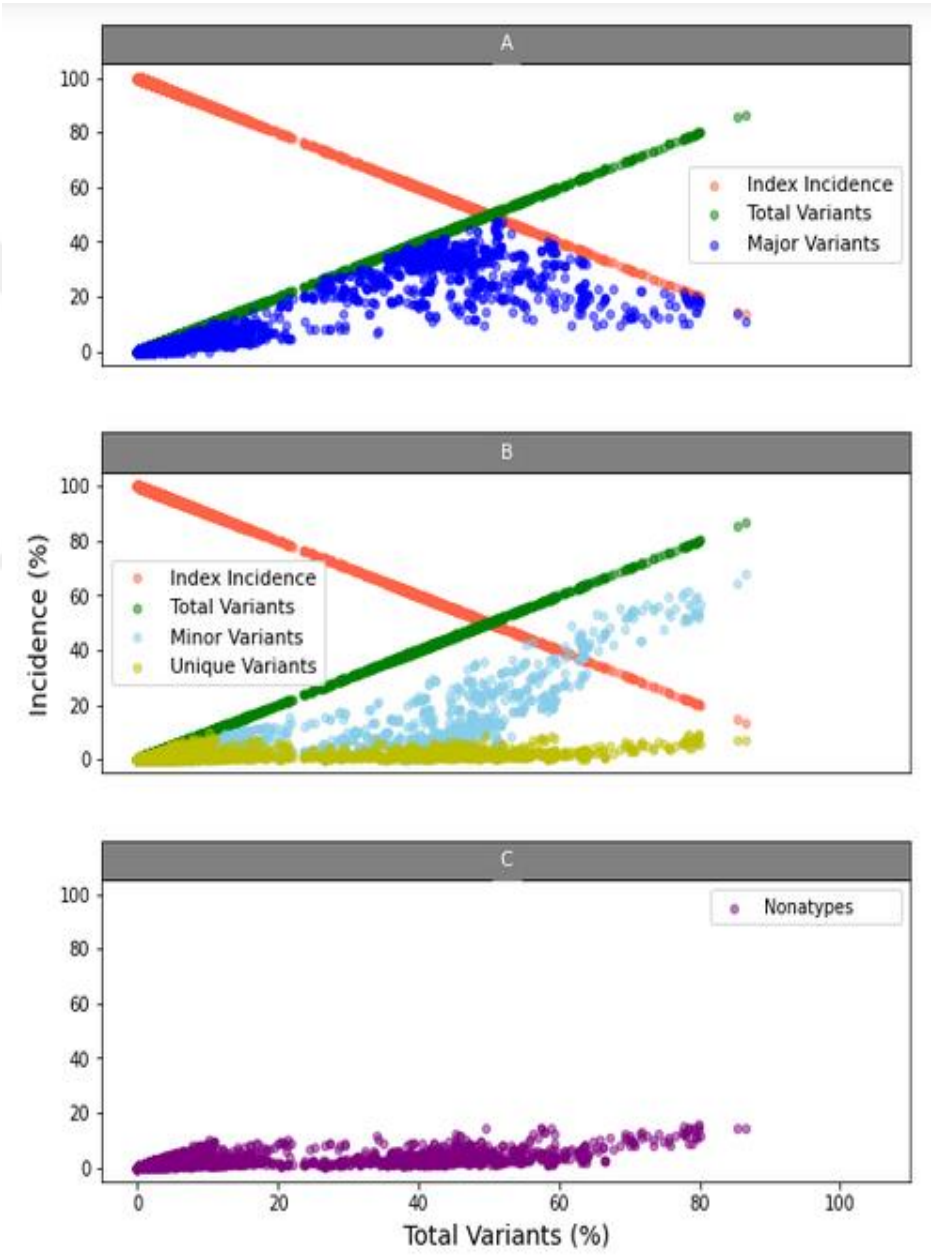


**Figure 3.19:** Proteome-based diversity motif patterns in the H4N6 IAV subtype of avian host.

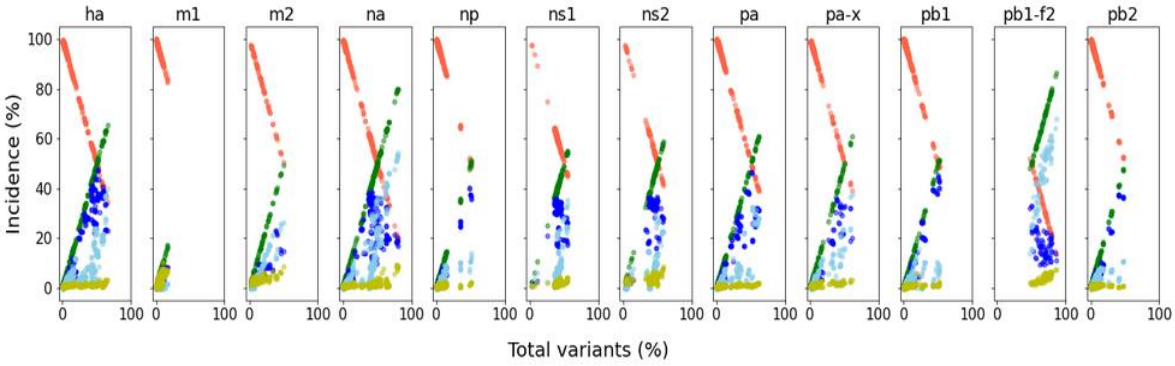


**Figure 3.20:** Protein-based diversity motif patterns in the H4N6 IAV subtype of avian host.

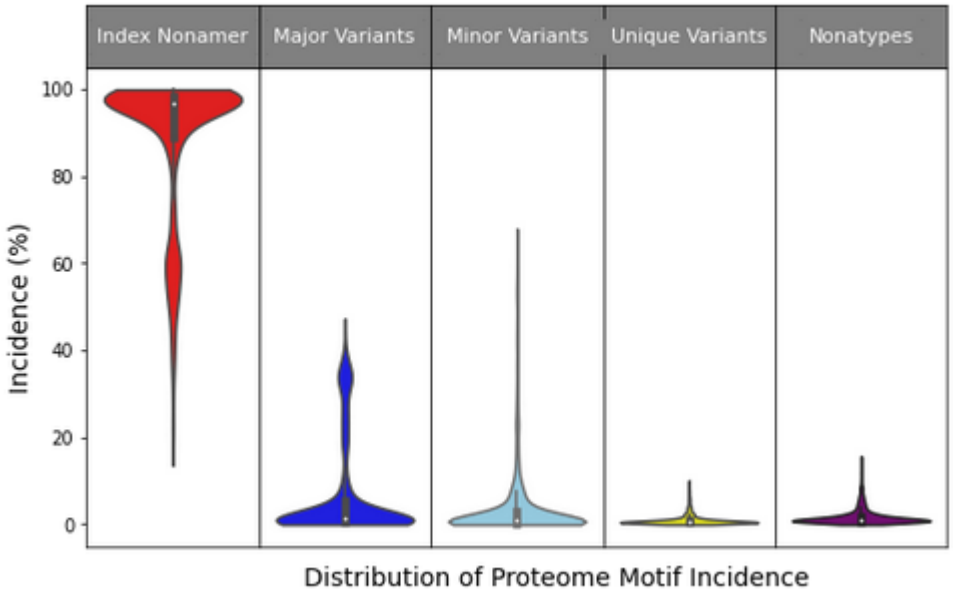


**Figure 3.21:** Diversity motif frequency distribution for the H4N6 IAV proteome of viruses isolated from avian hosts.

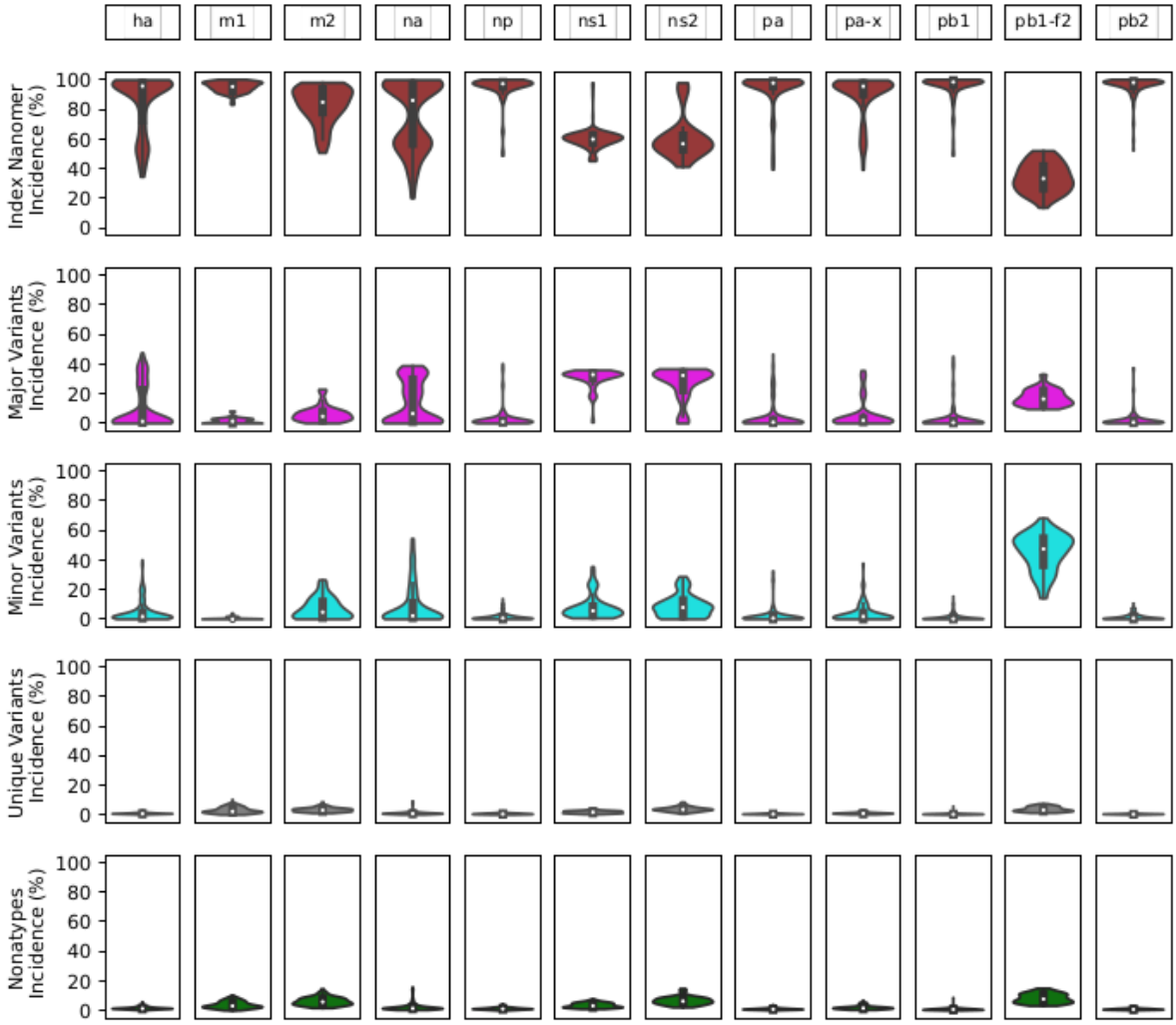


**Figure 3.22:** Diversity motif frequency distribution for the H4N6 IAV individual proteins of viruses isolated from avian hosts.

The Figure 3.23 provides a distribution of index sequences across various conservation thresholds for the H4N6 viruses of the avian host. Although a greater percentage (~73.06%) of the positions of the proteome were highly conserved (90% ≤ index incidence < 100%) only a few (0.48%) of them were completely conserved (index incidence = 100%). The second largest percentage of the positions (26.39%) constituted the mixed variable (20% ≤ index incidence < 90%) category. The least percentage of the positions (0.07%) were highly diversed occurring only in the PB1-F2 protein.

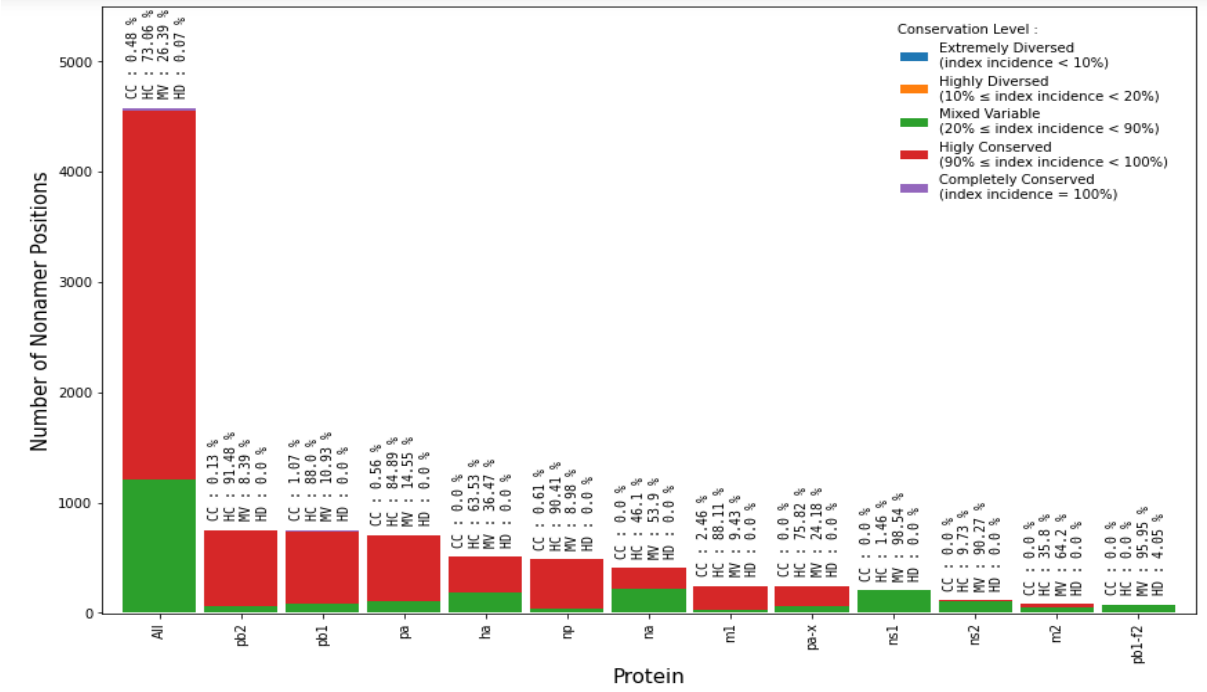


**Figure 3.23:** Distribution of index incidences across various conservation levels for all the nonamer positions of the H4N6 viruses of the avian host. All (proteome) refers to the combined distributions of the individual proteins.

**Table 3.5:** Summary statistics of entropy and diversity motif distributions.

| IAV Subtype | Proteome (Human host) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Absolute entropy range (Bits)** | **Average (mean) entropy (Bits)** | **Average (mean) index incidence (%)** | **Average (mean) total variants incidence (%)** | **Average (mean) major variant incidence (%)** | **Average (mean) minor variants incidence (%)** | **Average (mean) unique variants incidence (%)** | **Average (mean) nonatype incidence (%)** |
| H3N2 | 4.6 | 0.56 | 90.87 | 9.12 | 4.75 | 3.99 | 0.38 | 0.74 |
| H7N9 | 3.32 | 0.43 | 92.16 | 7.84 | 4.86 | 1.95 | 1.03 | 1.72 |
| **IAV Subtype** | **Proteome (Avian host)** | | | | | | | |
| | **Absolute entropy range (Bits)** | **Average (mean) entropy (Bits)** | **Average (mean) index incidence (%)** | **Average (mean) total variants incidence (%)** | **Average (mean) major variant incidence (%)** | **Average (mean) minor variants incidence (%)** | **Average (mean) unique variants incidence (%)** | **Average (mean) nonatype incidence (%)** |
| H3N2 | 5.36 | 0.66 | 87.95 | 12.05 | 5.93 | 4.39 | 1.73 | 2.91 |
| H7N9 | 4.52 | 0.67 | 87.64 | 12.36 | 6.87 | 4.12 | 1.38 | 2.45 |
| H4N6 | 4.92 | 0.60 | 88.13 | 11.87 | 6.98 | 3.86 | 1.03 | 1.81 |

# 4. DISCUSSION

Using in-silico methods, this thesis work evaluated the diversity of IAV protein sequences across three (3) IAV subtypes. The study worked on human and avian protein sequences of H3N2, H7N9 and the avian protein sequences of H4N6, resulting in evaluation of five (5) proteomes (Table 3.5)**,** the data of which were retrieved from public databases, GISAID EpiFlu and FluDB.

Across all the subtypes, the total concentration of mutations (total variants) was inversely proportional to the index sequence (which represents sequence conservation). Based on all the applied diversity metrics (entropy, total variants, nonatypes and conservation levels), the human H7N9 proteome proved to be the most conserved proteome across the 3 subtypes (5 proteomes). However, on the spectrum of diversity, the avian H3N2 proved to be the most diverse subtype. Specifically, the HA protein of both the human and avian H3N2 exhibited three (3) nonamer positions (human: 184 - 192, 185 - 193, and 186 - 194) and one (1) nonamer position (avian: 109 - 117) in the highly diversed region, and more importantly the NA protein of the avian H3N2 had six (6) nonamer positions (103 - 111, 104 - 112, 105 - 113, 106 - 114, 107 - 115, and 108 - 116) in the extremely diversed region. These proteins (HA and NA) define the subtype of an IAV (Diskin et al., 2020), and thus the more diverse they are, the more diverse the entire subtype. (Hu et al., 2013) pointed out that positions in the highly diversed/extremely diversed regions are likely to impact immune recognition by the host and may result in immune escape. Amidst other factors, this may partly explain why (Vemula et al. 2016) observed that IAV seasons that include H3N2 as one of the circulating variants are particularly more fatal than IAV seasons without H3N2.

The proteins PB1 and PB2 were found to be the most conserved proteins across the five (5) proteomes (human and avian host proteomes of H3N2, H4N6 and avian proteome of H4N6). Amongst the five conservation levels (CC, HC, MV, HD, and ED), PB1 and PB2 maintained a relatively higher fraction of the combined upper conservation levels (CC + HC) as compared to the rest of the proteins; although PA, M1 and NP proteins had a good percentage of conserved sequences. This is also in line with the works of others (Heiny et al., 2007; P. T. Tan et al., 2011) who also reported that along with other proteins, PB1 and PB2 to be highly conserved proteins. These proteins (PB1 and PB2) are non-structural in nature and although the immune

system recognizes structural proteins (Rawal et al., 2021), it has been shown that non-structural proteins can also trigger immune response and thus these proteins (PB1 and PB2) are potential candidates as vaccine targets.

The least conserved (or the most diversed) protein was PB1-F2, with the highest fraction of the combined lower conservation levels (MV and HD) across all the five (5) proteomes. Therefore, PB1-F2 is not suitable as vaccine target. However, convention vaccines that use the whole virus would include this highly diverse protein and may negatively impact the duration of the immunity.

While all other proteins nearly had similar diversity dynamics between the hosts, PB1-F2 was the only protein that showed significant change in diversity dynamics (difference) between avian and human subtypes. The PB1-F2 in the avian host had a higher diversity level (HD+ED) than those of the human host. This is in congruence with the observation by (Abd Raman et al., 2020) that the avian PB1-F2 of H5N1 subtype was the only protein that had all its nonamers in the mixed variable (MV) conservation level, while other proteins, including the human PB1-F2 had some of the nonamer positions in the highly conserved (HC) conservation level. It would be important to investigate what may have accounted for this significant difference between the PB1-F2 protein of the human and avian hosts. This may further shed light on the role of PB1-F2 in viruses jumping from the avian host to the human host.

As more avian subtypes are making their way into the human host and continuing to diversify within the human host, existing IAV vaccines against human infection become obsolete at high rate and in order to cope with this situation, IAV vaccines have to be continuously redesigned at the same rate at which the old vaccines become obsolete (Salzberg, 2008; Blanton et al., 2015; CDC, 2020). This is obviously tedious, expensive and given the long process of vaccine design with our current technologies, it is possible that some vaccines could get obsolete before reaching the market based on how rapid IAV mutate. This has called for several efforts to come up with a universal IAV vaccine (Bangaru et al., 2018; J. Tan et al., 2018; Krammer & Palese, 2019). The mitigation of the avian to human viral jump will require comprehensive understanding of the diversity dynamics of the virus as it journeys as a purely avian subtype, such as H4N6, to as a well-established human subtype, such as H3N2, and others that are in-between these two extremes, such as H7N9. One of the key components of a successful universal vaccine is the identification of target proteins (sequences) that do not change (are

completely conserved) across all current and future IAV subtypes (Nachbagauer & Krammer, 2017). Vaccines designed upon such proteins and their regions within are expected to be effective against a large number of IAV variants, if not all of them. Based on the results of this thesis and similar results of others, PB1 and PB2 are attractive as targets for a universal vaccine that aims to target the non-structural proteins (Thura et al., 2021) compared to existing vaccines that target the structural proteins, such as the HA and NA proteins, which are less conserved.

This thesis has simultaneously studied and elucidated the diversity dynamics of a purely avian subtype (H4N6), a purely human subtype (H3N2), and that of a subtype that is in-between the two extremes (H7N9). Among the five (5) proteomes studied from these three subtypes, the most diverse was H3N2 subtype of avian host and the most conserved was H7N9 subtype of human host. The fact that nonatypes were observed for almost all the positions analysed shows that sequence change is spread out across the length of the proteome of each of the subtypes. This supports the notion of the general highly diversed nature of IAVs and for that matter requires continuous, real-time diversity analysis of IAV data deposited in public repositories.

Notably, with about more than 96 different subtypes of IAV in existence (Diskin et al., 2020), this study on just three (3) IAV subtypes might not be representative enough for the entire population of the existing IAV subtypes. Thus, it will be beneficial for future work to include as many subtypes as possible in such a study. Also, due to time and resource limitations, the MSAs (Multiple Sequence Alignments) of the protein sequences were not manually checked, thoroughly for potential sites of misalignments. However, the MSA tools used are generally reliable and widely used, plus general checks indicated reliable alignment of the IAVs.

Data redundance is noted to be a concern for public repositories. For example, there were nearly more than 60% duplicates in the downloaded sequences of all the three (3) subtype data. For example, the original number of sequences for the human host H3N2 subtype downloaded dropped from around 800,000 to just 91,000 sequences after the duplicates were removed. This duplicates unnecessarily overburden the databases and compute power needed to process the data. This could be a major bottleneck for real time sequence analysis. Thus, it may be useful if these online sequence repository platforms do a check on data that is uploaded and if they are duplicates, then only the metadata (country, ID, laboratory, etc) is saved in the database against the already existing reference sequence. The same can be done with the pre-existing data with

the metadata annotations made on a selected reference sequence. This will help optimize storage space, compute power, and eventually reduce data pre-processing time during research.

# 5. CONCLUSION AND RECOMMENDATIONS

This study provided an overview of the diversity dynamics of three IAV subtypes across two different hosts, avian and human. The subtypes were selected to evaluate the diversity difference in terms of adaptability to the host, with H3N2 being a well-established human virus and H4N6 mainly an avian virus, while H7N9 somewhat in between. The results provided insight into the difference in adaptation dynamics across the subtypes and the hosts. This study also classified the index sequences of each proteome according to five conservation levels. PB1 and PB2 were the most conserved and are attractive as candidates for vaccine design, in particular as universal vaccine target. PB1-F2 was the most diverse and thus, should be excluded for vaccine design strategies. The approach reported herein is generic and can be utilised for the study of diversity dynamics of many other IAV subtypes. Future work may include characterisation of the highly conserved sequences identified as immune targets.

## REFERENCES

Abd Raman, H., Tan, S., August, J., & Khan, M. A. (2020). Dynamics of Influenza A (H5N1) virus protein sequence diversity. *PeerJ*, 7, e7954. https://doi.org/10.7717/peerj.7954

Allen, J. D., & Ross, T. M. (2021). Next generation methodology for updating HA vaccines against emerging human seasonal influenza A(H3N2) viruses. *Scientific Reports*, 11, 4554. https://doi.org/10.1038/s41598-020-79590-7

Alnaji, F. G., Bentley, K., Pearson, A., Woodman, A., Moore, J., Fox, H., Macadam, A. J., & Evans, D. J. (2022). Generated Randomly and Selected Functionally? The Nature of Enterovirus Recombination. *Viruses*, 14(5), 916. https://doi.org/10.3390/v14050916

Bangaru, S., Zhang, H., Gilchuk, I. M., Voss, T. G., Irving, R. P., Gilchuk, P., Matta, P., Zhu, X., Lang, S., Nieusma, T., Richt, J. A., Albrecht, R. A., Vanderven, H. A., Bombardi, R., Kent, S. J., Ward, A. B., Wilson, I. A., & Crowe, J. E. (2018). A multifunctional human monoclonal neutralizing antibody that targets a unique conserved epitope on influenza HA. *Nature Communications*, *9*(1), 2669. https://doi.org/10.1038/s41467-018-04704-9

Becker, T., Elbahesh, H., Reperant, L. A., Rimmelzwaan, G. F., & Osterhaus, A. D. M. E. (2021). Influenza Vaccines: Successes and Continuing Challenges. *The Journal of Infectious Diseases*, 224(Suppl 4), S405–S419. https://doi.org/10.1093/infdis/jiab269

Blanton, L., Kniss, K., Smith, S., Mustaquim, D., Steffens, C., Flannery, B., Fry, A. M., Bresee, J., Wallis, T., Garten, R., Xu, X., Elal, A. I. A., Gubareva, L., Wentworth, D. E., Burns, E., Katz, J., Jernigan, D., & Brammer, L. (2015). Update: Influenza Activity--United States and Worldwide, May 24-September 5, 2015. MMWR. *Morbidity and Mortality Weekly Report*, 64(36), 1011–1016. https://doi.org/10.15585/mmwr.mm6436a4

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

Brunotte, L., Flies, J., Bolte, H., Reuther, P., Vreede, F., & Schwemmle, M. (2014). The Nuclear Export Protein of H5N1 Influenza A Viruses Recruits Matrix 1 (M1) Protein to the Viral Ribonucleoprotein to Mediate Nuclear Export *. *Journal of Biological Chemistry*, 289(29), 20067–20077. https://doi.org/10.1074/jbc.M114.569178

Bui, V. N., Ogawa, H., Xininigen, Karibe, K., Matsuo, K., Awad, S. S. A., Minoungou, G. L., Yoden, S., Haneda, H., Ngo, L. H., Tamaki, S., Yamamoto, Y., Nakamura, K., Saito, K., Watanabe, Y., Runstadler, J., Huettman, F., Happ, G. M., & Imai, K. (2012). H4N8 subtype avian influenza virus isolated from shorebirds contains a unique PB1 gene and causes severe respiratory disease in mice. *Virology*, 423(1), 77–88. https://doi.org/10.1016/j.virol.2011.11.019

CDC. (2020, October 26). Selecting Viruses for the Seasonal Flu Vaccine. *Centers for Disease Control and Prevention*. https://www.cdc.gov/flu/prevent/vaccine-selection.htm [Last visit: 24 April, 2021]

CDC. (2021, September 21). How Flu Viruses Can Change. *Centers for Disease Control and Prevention*. https://www.cdc.gov/flu/about/viruses/change.htm [Last visit: 31 May, 2022]

Celik, M. (2021). *Motif SWitch Analyser (MoSwA) (1.0.0) [Computer software]*. https://moswa.bioinfo.perdanauniversity.edu.my/ [Last visit: 25 April, 2022]

Chauhan, R. P., & Gordon, M. L. (2022). An overview of influenza A virus genes, protein functions, and replication cycle highlighting important updates. *Virus Genes*. https://doi.org/10.1007/s11262-022-01904-w

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Dawood, F., Jain, S., Finelli, L., Shaw, M., Lindstrom, S., Garten, R., Gubareva, L., Xu, X., Bridges, C., & Uyeki, T. (2009). Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in HumansNovel Swine-Origin Influenza A (H1N1) Investigation Team. *New England Journal of Medicine*, 360, 2605–2615.

Diskin, E. R., Friedman, K., Krauss, S., Nolting, J. M., Poulson, R. L., Slemons, R. D., Stallknecht, D. E., Webster, R. G., & Bowman, A. S. (2020). Subtype Diversity of Influenza A Virus in North American Waterfowl: A Multidecade Study. *Journal of Virology*, 94(11), e02022-19. https://doi.org/10.1128/JVI.02022-19

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Eisfeld, A. J., Neumann, G., & Kawaoka, Y. (2015). At the centre: Influenza A virus ribonucleoproteins. Nature Reviews. *Microbiology*, 13(1), 28–41. https://doi.org/10.1038/nrmicro3367

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1), 33–46. https://doi.org/10.1002/gch2.1018

El-Sayed, A., & Kamel, M. (2020). Climatic changes and their role in emergence and re-emergence of diseases. *Environmental Science and Pollution Research*, 27(18), 22336–22352. https://doi.org/10.1007/s11356-020-08896-w

Eng, C. L. P., Tong, J. C., & Tan, T. W. (2014). Predicting host tropism of influenza A virus proteins using random forest. *BMC Medical Genomics*, 7 Suppl 3, S1. https://doi.org/10.1186/1755-8794-7-S3-S1

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Guan, L., Shi, J., Kong, X., Ma, S., Zhang, Y., Yin, X., He, X., Liu, L., Suzuki, Y., Li, C., Deng, G., & Chen, H. (2019). H3N2 avian influenza viruses detected in live poultry markets in China bind to human-type receptors and transmit in guinea pigs and ferrets. *Emerging Microbes & Infections*, 8(1), 1280–1290. https://doi.org/10.1080/22221751.2019.1660590

Hara, K., Schmidt, F. I., Crow, M., & Brownlee, G. G. (2006). Amino acid residues in the N-terminal region of the PA subunit of influenza A virus RNA polymerase play a critical role in protein stability, endonuclease activity, cap binding, and virion RNA promoter binding. *Journal of Virology*, 80(16), 7789–7798. https://doi.org/10.1128/JVI.00600-06

Hause, B. M., Collin, E. A., Liu, R., Huang, B., Sheng, Z., Lu, W., Wang, D., Nelson, E. A., & Li, F. (2014). Characterization of a Novel Influenza Virus in Cattle and Swine: Proposal for a New Genus in the Orthomyxoviridae Family. *MBio*, 5(2), e00031-14. https://doi.org/10.1128/mBio.00031-14

Heiny, A. T., Miotto, O., Srinivasan, K. N., Khan, A. M., Zhang, G. L., Brusic, V., Tan, T. W., & August, J. T. (2007). Evolutionarily Conserved Protein Sequences of Influenza A Viruses, Avian and Human, as Vaccine Targets. *PLOS ONE*, 2(11), e1190. https://doi.org/10.1371/journal.pone.0001190

Hu, Y., Tan, P. T., Tan, T. W., August, J. T., & Khan, A. M. (2013). Dissecting the Dynamics of HIV-1 Protein Sequence Diversity. *PLOS ONE*, 8(4), e59994. https://doi.org/10.1371/journal.pone.0059994

Hughes, J. M., Wilson, M. E., Pike, B. L., Saylors, K. E., Fair, J. N., LeBreton, M., Tamoufe, U., Djoko, C. F., Rimoin, A. W., & Wolfe, N. D. (2010). The Origin and Prevention of Pandemics. *Clinical Infectious Diseases*, *50*(12), 1636–1640. https://doi.org/10.1086/652860

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jochum, M., Fischer, M., Isbell, F., Roscher, C., van der Plas, F., Boch, S., Boenisch, G., Buchmann, N., Catford, J. A., Cavender-Bares, J., Ebeling, A., Eisenhauer, N., Gleixner, G., Hölzel, N., Kattge, J., Klaus, V. H., Kleinebecker, T., Lange, M., Le Provost, G., … Manning, P. (2020). The results of biodiversity–ecosystem functioning experiments are realistic. *Nature Ecology & Evolution*, 4(11), 1485–1494. https://doi.org/10.1038/s41559-020-1280-9

Johnson, N. P. A. S., & Mueller, J. (2002). Updating the Accounts: Global Mortality of the 1918-1920 “Spanish” Influenza Pandemic. *Bulletin of the History of Medicine*, 76(1), 105–115. https://doi.org/10.1353/bhm.2002.0022

Khan, A. M., Miotto, O., Heiny, A. T., Salmon, J., Srinivasan, K. N., Nascimento, E. J. M., Marques, E. T. A., Brusic, V., Tan, T. W., & August, J. T. (2006). A systematic bioinformatics approach for selection of epitope-based vaccine targets. *Cellular Immunology*, 244(2), 141–147. https://doi.org/10.1016/j.cellimm.2007.02.005

Khan, A. M., Miotto, O., Nascimento, E. J. M., Srinivasan, K. N., Heiny, A. T., Zhang, G. L., Marques, E. T., Tan, T. W., Brusic, V., Salmon, J., & August, J. T. (2008). Conservation and Variability of Dengue Virus Proteins: Implications for Vaccine Design. *PLOS Neglected Tropical Diseases*, 2(8), e272. https://doi.org/10.1371/journal.pntd.0000272

Khanna, M., Sharma, S., Kumar, B., & Rajput, R. (2014). Protective Immunity Based on the Conserved Hemagglutinin Stalk Domain and Its Prospects for Universal Influenza Vaccine Development. *BioMed Research International*, 2014, e546274. https://doi.org/10.1155/2014/546274

Khare, S., Gurry, C., Freitas, L., Schultz, M. B., Bach, G., Diallo, A., Akite, N., Ho, J., Lee, R. T., Yeo, W., Curation Team, G. C., & Maurer-Stroh, S. (2021). GISAID's Role in Pandemic Response. *China CDC Weekly*, 3(49), 1049–1051. https://doi.org/10.46234/ccdcw2021.255

Koo, Q. Y., Khan, A. M., Jung, K.-O., Ramdas, S., Miotto, O., Tan, T. W., Brusic, V., Salmon, J., & August, J. T. (2009). Conservation and Variability of West Nile Virus Proteins. *PLOS ONE*, 4(4), e5352. https://doi.org/10.1371/journal.pone.0005352

Kumar, B., Asha, K., Khanna, M., Ronsard, L., Meseko, C. A., & Sanicas, M. (2018). The emerging influenza virus threat: Status and new prospects for its therapy and control. *Archives of Virology*, 163(4), 831–844. https://doi.org/10.1007/s00705-018-3708-y

Krammer, F., & Palese, P. (2019). Universal Influenza Virus Vaccines That Target the Conserved Hemagglutinin Stalk and Conserved Sites in the Head Domain. *The Journal of Infectious Diseases*, 219(Suppl 1), S62–S67. https://doi.org/10.1093/infdis/jiy711

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liang, L., Deng, G., Shi, J., Wang, S., Zhang, Q., Kong, H., Gu, C., Guan, Y., Suzuki, Y., Li, Y., Jiang, Y., Tian, G., Liu, L., Li, C., & Chen, H. (2016). Genetics, Receptor Binding, Replication, and Mammalian Transmission of H4 Avian Influenza Viruses Isolated from Live Poultry Markets in China. *Journal of Virology*, 90(3), 1455–1469. https://doi.org/10.1128/JVI.02692-15

Liu, R., Sheng, Z., Huang, C., Wang, D., & Li, F. (2020). Influenza D virus. *Current Opinion in Virology*, 44, 154–161. https://doi.org/10.1016/j.coviro.2020.08.004

Long, J. C. D., & Fodor, E. (2016). The PB2 Subunit of the Influenza A Virus RNA Polymerase Is Imported into the Mitochondrial Matrix. *Journal of Virology*, 90(19), 8729–8738. https://doi.org/10.1128/JVI.01384-16

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), 1750–1757. https://doi.org/10.1128/JVI.06203-11

Miotto, O., Heiny, A. T., Albrecht, R., García-Sastre, A., Tan, T. W., August, J. T., & Brusic, V. (2010). Complete-Proteome Mapping of Human Influenza A Adaptive Mutations:

Implications for Human Transmissibility of Zoonotic Strains. *PLOS ONE*, 5(2), e9025. https://doi.org/10.1371/journal.pone.0009025

Moelling, K., & Broecker, F. (2019). Viruses and Evolution – Viruses First? A Personal Perspective. *Frontiers in Microbiology*, 10. https://doi.org/10.3389/fmicb.2019.00523

Moore, T. C., Fong, J., Rosa Hernández, A. M., & Pogreba-Brown, K. (2021). CAFOs, novel influenza, and the need for One Health approaches. *One Health*, 13, 100246. https://doi.org/10.1016/j.onehlt.2021.100246

Morens, D. M., & Taubenberger, J. K. (2018). The Mother of All Pandemics Is 100 Years Old (and Going Strong)! *American Journal of Public Health*, 108(11), 1449–1454. https://doi.org/10.2105/AJPH.2018.304631

Mumford, J. A. (2007). Vaccines and viral antigenic diversity. *Revue Scientifique Et Technique (International Office of Epizootics)*, 26(1), 69–90.

Nachbagauer, R., & Krammer, F. (2017). Universal influenza virus vaccines and therapeutic antibodies. *Clinical Microbiology and Infection*, 23(4), 222–228. https://doi.org/10.1016/j.cmi.2017.02.009

Obayashi, E., Yoshida, H., Kawai, F., Shibayama, N., Kawaguchi, A., Nagata, K., Tame, J. R. H., & Park, S.-Y. (2008). The structural basis for an essential subunit interaction in influenza virus RNA polymerase. *Nature*, 454(7208), 1127–1131. https://doi.org/10.1038/nature07225

Paterson, D., & Fodor, E. (2012). Emerging Roles for the Influenza A Virus Nuclear Export Protein (NEP). *PLOS Pathogens*, 8(12), e1003019. https://doi.org/10.1371/journal.ppat.1003019

Piret, J., & Boivin, G. (2021a). Pandemics Throughout History. *Frontiers in Microbiology*, 11. https://doi.org/10.3389/fmicb.2020.631736

Piret, J., & Boivin, G. (2021b). Pandemics Throughout History. *Frontiers in Microbiology*, 11. https://doi.org/10.1371/journal.ppat.1003019

Rawal, K., Sinha, R., Abbasi, B. A., Chaudhary, A., Nath, S. K., Kumari, P., Preeti, P., Saraf, D., Singh, S., Mishra, K., Gupta, P., Mishra, A., Sharma, T., Gupta, S., Singh, P., Sood, S., Subramani, P., Dubey, A. K., Strych, U., … Bottazzi, M. E. (2021). Identification of vaccine targets in pathogens and design of a vaccine using computational approaches. *Scientific Reports*, 11(1), 17626. https://doi.org/10.1038/s41598-021-96863-x

Rejmanek, D., Hosseini, P. R., Mazet, J. A. K., Daszak, P., & Goldstein, T. (2015). Evolutionary Dynamics and Global Diversity of Influenza A Virus. *Journal of Virology*, 89(21), 10993–11001. https://doi.org/10.1128/JVI.01573-15

Sankaran, N., & Weiss, R. A. (2021). Viruses: Impact on Science and Society. *Encyclopedia of Virology*, 671–680. https://doi.org/10.1016/B978-0-12-814515-9.00075-8

Saunders-Hastings, P. R., & Krewski, D. (2016). Reviewing the History of Pandemic Influenza: Understanding Patterns of Emergence and Transmission. *Pathogens*, 5(4), 66. https://doi.org/10.3390/pathogens5040066

Scroggs, S. L. P., Grubaugh, N. D., Sena, J. A., Sundararajan, A., Schilkey, F. D., Smith, D. R., Ebel, G. D., & Hanley, K. A. (2019). Endless Forms: Within-Host Variation in the Structure of the West Nile Virus RNA Genome during Serial Passage in Bird Hosts. *MSphere*, 4(3), e00291-19. https://doi.org/10.1128/mSphere.00291-19

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shi, Y., Cui, H., Wang, J., Chi, Q., Li, X., Teng, Q., Chen, H., Yang, J., Liu, Q., & Li, Z. (2016). Characterizations of H4 avian influenza viruses isolated from ducks in live poultry markets and farm in Shanghai. *Scientific Reports*, 6, 37843. https://doi.org/10.1038/srep37843

Shimizu, T., Takizawa, N., Watanabe, K., Nagata, K., & Kobayashi, N. (2011). Crucial role of the influenza virus NS2 (NEP) C-terminal domain in M1 binding and nuclear export of vRNP. *FEBS Letters*, 585(1), 41–46. https://doi.org/10.1016/j.febslet.2010.11.017

Shu, Y., & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Salzberg, S. (2008). The contents of the syringe. *Nature*, 454(7201), 160–161. https://doi.org/10.1038/454160a

Sun, W., Luo, T., Liu, W., & Li, J. (2020). Progress in the Development of Universal Influenza Vaccines. *Viruses*, 12(9), 1033. https://doi.org/10.3390/v12091033

Szewczyk, B., Bienkowska-Szewczyk, K., & Krol, E. (2014). Introduction to molecular biology of influenza A viruses. *Acta Biochimica Polonica*, 61. https://doi.org/10.18388/abp.2014_1857

Taheri, M., Nemattalab, M., Mahjoob, M., Hasan-alizadeh, E., Zamani, N., Nikokar, I., Evazalipour, M., Soltani Tehrani, B., & Shenagari, M. (2021). Toward a universal influenza virus vaccine: Some cytokines may fulfill the request. *Cytokine*, 148, 155703. https://doi.org/10.1016/j.cyto.2021.155703

Tan, J., Asthagiri-Arunkumar, G., & Krammer, F. (2018). Universal influenza virus vaccines and therapeutics: Where do we stand with influenza B virus? *Current Opinion in Immunology*, 53, 45–50. https://doi.org/10.1016/j.coi.2018.04.002

Tan, P. T., Khan, A. M., & August, J. T. (2011). Highly conserved influenza A sequences as T cell epitopes-based vaccine targets to address the viral variability. *Human Vaccines*, 7(4), 402–409. https://doi.org/10.4161/hv.7.4.13845

Taubenberger, J. K., Hultin, J. V., & Morens, D. M. (2007). Discovery and characterization of the 1918 pandemic influenza virus in historical context. *Antiviral Therapy*, 12(4 Pt B), 581–591.

Taubenberger, J. K., & Kash, J. C. (2010). Influenza Virus Evolution, Host Adaptation, and Pandemic Formation. *Cell Host & Microbe*, 7(6), 440–451. https://doi.org/10.1016/j.chom.2010.05.009

Taubenberger, J. K., & Morens, D. M. (2009). Pandemic influenza – including a risk assessment of H5N1. *Revue Scientifique et Technique (International Office of Epizootics)*, 28(1), 187–202.

Taubenberger, J. K., & Morens, D. M. (2020). The 1918 Influenza Pandemic and Its Legacy. *Cold Spring Harbor Perspectives in Medicine*, 10(10), a038695. https://doi.org/10.1101/cshperspect.a038695

Tharanga, S., Hu, Y., Unlu, E. S., Sjaugi, M. F., Celik, M. A., Hekimoglu, H., Miotto, O., Oncel, M. M., & Khan, A. M. (2022). DiMA: Sequence Diversity Dynamics Analyser for Viruses *(arXiv:2205.13915). arXiv*. https://doi.org/10.48550/arXiv.2205.13915

Thura, M., Sng, J. X. E., Ang, K. H., Li, J., Gupta, A., Hong, J. M., Hong, C. W., & Zeng, Q. (2021). Targeting intra-viral conserved nucleocapsid (N) proteins as novel vaccines against SARS-CoVs. *Bioscience Reports*, 41(9), BSR20211491. https://doi.org/10.1042/BSR20211491

To, K. K. W., Tsang, A. K. L., Chan, J. F. W., Cheng, V. C. C., Chen, H., & Yuen, K.-Y. (2014). Emergence in China of human disease due to avian influenza A(H10N8) – Cause for concern? *Journal of Infection*, 68(3), 205–215. https://doi.org/10.1016/j.jinf.2013.12.014

Tscherne, D. M., & García-Sastre, A. (2011). Virulence determinants of pandemic influenza viruses. *The Journal of Clinical Investigation*, 121(1), 6–13. https://doi.org/10.1172/JCI44947

Valleron, A.-J., Cori, A., Valtat, S., Meurisse, S., Carrat, F., & Boëlle, P.-Y. (2010). Transmissibility and geographic spread of the 1889 influenza pandemic. Proceedings of the *National Academy of Sciences*, 107(19), 8778–8781. https://doi.org/10.1073/pnas.1000886107

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

VanDalen, K. K., Franklin, A. B., Mooers, N. L., Sullivan, H. J., & Shriner, S. A. (2010). Shedding Light on Avian Influenza H4N6 Infection in Mallards: Modes of Transmission and Implications for Surveillance. *PLoS ONE*, 5(9), e12851. https://doi.org/10.1371/journal.pone.0012851

Vasin, A. V., Temkina, O. A., Egorov, V. V., Klotchenko, S. A., Plotnikova, M. A., & Kiselev, O. I. (2014). Molecular mechanisms enhancing the proteome of influenza A viruses: An overview of recently discovered proteins. *Virus Research*, 185, 53–63. https://doi.org/10.1016/j.virusres.2014.03.015

Vemula, S. V., Zhao, J., Liu, J., Wang, X., Biswas, S., & Hewlett, I. (2016). Current Approaches for Diagnosis of Influenza Virus Infections in Humans. *Viruses*, 8(4), 96. https://doi.org/10.3390/v8040096

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., Ruiter, J. de, Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., … Qalieh, A. (2017). Mwaskom/seaborn: V0.8.1 (September 2017). *Zenodo*. https://doi.org/10.5281/zenodo.883859

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., & Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiological Reviews*, 56(1), 152–179. https://doi.org/10.1128/mr.56.1.152-179.1992

White, R. J., & Razgour, O. (2020). Emerging zoonotic diseases originating in mammals: A systematic review of effects of anthropogenic land-use change. *Mammal Review*, 50(4), 336–352. https://doi.org/10.1111/mam.12201

Wolfe, N. D., Dunavan, C. P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, 447(7142), 279–283. https://doi.org/10.1038/nature05775

Worobey, M., Han, G.-Z., & Rambaut, A. (2014). Genesis and pathogenesis of the 1918 pandemic H1N1 influenza A virus. *Proceedings of the National Academy of Sciences*, 111(22), 8107–8112. https://doi.org/10.1073/pnas.1324197111

Yin, X., Deng, G., Zeng, X., Cui, P., Hou, Y., Liu, Y., Fang, J., Pan, S., Wang, D., Chen, X., Zhang, Y., Wang, X., Tian, G., Li, Y., Chen, Y., Liu, L., Suzuki, Y., Guan, Y., Li, C., … Chen, H. (2021). Genetic and biological properties of H7N9 avian influenza viruses detected after application of the H7N9 poultry vaccine in China. *PLoS Pathogens*, 17(4), e1009561. https://doi.org/10.1371/journal.ppat.1009561

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C. N., Lee, A. J., Li, X., Macken, C., Mahaffey, C., Pickett, B. E., Reardon, B., Smith, T., Stewart, L., Suloway, C., Sun, G., … Scheuermann, R. H. (2017). Influenza Research Database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1), D466–D474. https://doi.org/10.1093/nar/gkw857

# CURRICULUM VITAE

| Personal Information | |
|---|---|
| Name Surname | Rashid Mukaila |
| Nationality | ❒ T.C. ☑ Other: |

| Educational Information | |
|---|---|
| **B. Sc.** | |
| University | University of Cape Coast |
| Faculty | Faculty of Physical Science |
| Department | Department of Computer science and Information Technology |
| Graduation Year | 26.05.2017 |

| M. Sc. | |
|---|---|
| University | Istanbul University |
| Institute | Institute of Graduate Studies in Science and Engineering |
| Department | Informatics |
| Programme | Informatics |

| Publications |
|---|
| Proceeding<br><br>Mukaila, R., Öncel, M. M., Karimi, H., Engin, E. Ö., Karakuş, D. H., Üstünel, F., Işık, B. E., Gülen, B., Gültepe, B., Karaaslan, K., Koç, M. M., Küçük, S. Ö., Okyaltırık, N. F., Özdemir, R., Aydın, T., Tuncay, I., Khan, A. M., Kazancıoğlu, R. (2021). Integrating Biomedical Data For Health Analytics, A Case Study On COVID-19 Patients. *29th annual conference of the Intelligent Systems for Moleculer Biology/European Conference on Computational Biology (ISMB/ECCB)*. |