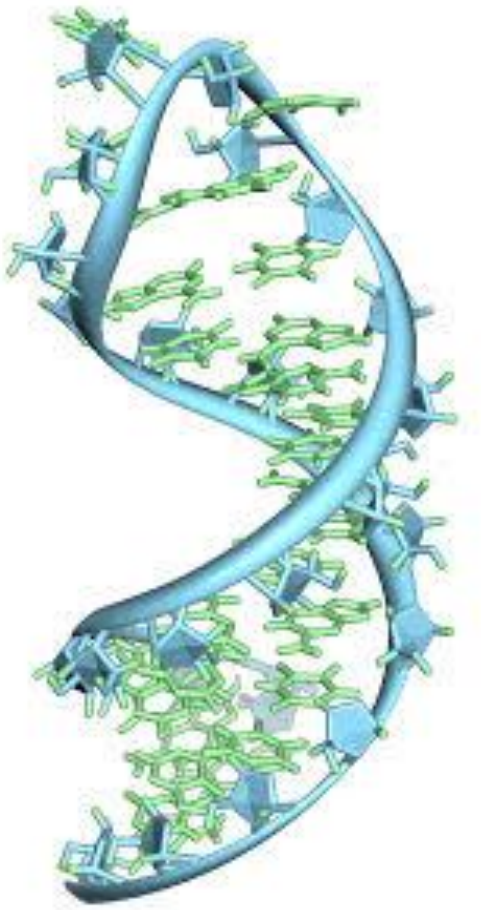
Chapter

1



INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 The genetic code and codon usage bias

Translational mapping from nucleobase triplets to amino acids is dictated by the genetic code. The process of translation of gene to protein involves decoding the information contained in the form of nucleotide triplets (codons) into amino acid sequences of the protein. There are 20 standard amino acids which are encoded by 61 sense codons. The codons can be grouped into 20 disjoint families, one for each amino acid. Three codons namely TAA, TAG and TGA (Wisconsin type) act as termination signal of polypeptide chain (Behura and Severson, 2013).

The presence of more codons compared to the number of amino acids has made the genetic code degenerate in nature. Within the standard genetic code, all amino acids except Met and Trp are coded by more than one codon. The codons encoding the same amino acid are called 'synonymous codons'. Studies have shown that the usage of synonymous codons is a non-random process and they are not used with equal frequency. These biases arising from unequal usage of synonymous codons are the consequence of natural selection during evolution. Extensive studies have shown that synonymous codon usage biases are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions (D'Onofrio *et al*., 2002; Gu *et al*., 2004b; Wan *et al*., 2004).

The preferential use of codons in the translational process is known as ‘codon usage bias’ or simply ‘codon bias’. Molecular evolutionary investigations suggest that codon usage bias is widespread across genomes and may contribute to genome evolution in a profound manner (Sharp and Matassi, 1994). With the rapid availability of vast number of sequences in the post whole genome sequencing era of a large number of species, scientists are now trying to look at the codon bias phenomenon in a holistic manner. Different authors have studied specific genes as well as whole genomes in context of the codon bias (Plotkin and Kudla, 2011).

The codon usage is attributable to the equilibrium between natural selection and mutation pressure (Sharp *et al*., 1993; Shackelton *et al*., 2006). Mutation, selection, and random drift represent the three major evolutionary forces that shape codon usage bias among species (Sharp and Matassi, 1994; Akashi *et al*., 1998; Rocha, 2004; Vicario *et al*., 2007). Studies have revealed that mutation bias may be a more important factor than natural selection in determining codon usage bias of some viruses and vertebrate DNA viruses (Zhong *et al*., 2007; Tao *et al*., 2009; Fu, 2010). It was also suggested that codon usage is related to gene function (Chiapello *et al*., 1998; Epstein *et al*., 2000; Ma *et al*., 2002) and protein secondary structure (Chiusano *et al*., 1999; Gupta *et al*., 2000; Gu *et al*., 2004b)

Moreover, optimal codon pairs can be replaced with synonymous codons in viral genomes in the development of attenuated viral vaccines (McArthur and Fong, 2010). These and many other areas of research and applications in modern biology signify the importance of accurate and meaningful analysis of codon bias in the organisms of interest (Mueller *et al*., 2006; Fletcher *et al*., 2007).

The exponential increase in the volume of sequence information during the early 90s facilitated for the first time the detailed statistical analyses of codon usage. Multivariate analysis techniques were applied to the analysis of the codon usage in mammalian, viral, bacterial, mitochondrial and lower eukaryotic genes (Grantham *et al*., 1980a; Grantham *et al*., 1980b).

1.2 Factors affecting codon usage bias

Apparently silent, codon usage bias at synonymous positions may have important implications on the protein product and other crucial cellular processes. Prominent investigators like Clarke (1970), Ikemura (1981a) were of the opinion that the codon usage setup of an organism is linked to its tRNA pool. Nevertheless, there is no clear-cut theory is available till date to explain the mechanism which steers synonymous variations in an organism's genome. A handful of hypotheses tried to explain the possible factors driving synonymous codon usage; however none has succeeded to solve the riddle that continues puzzling the stalwarts in molecular and evolutionary biology.

Simply put, codon usage is shaped by the equilibrium between two key forces: mutational bias and natural selection.

1.2.1 Mutational bias and codon usage

The mutational bias theory posits that the non-arbitrariness in the underlying mutational patterns is responsible for the existence of codon bias. According to this school of thought, certain codons mutate more and consequently possess minor equilibrium frequencies. Mutational biases vary amid various organisms, thereby bringing about disparity in the codon bias signatures across organisms (Plotkin and Kudla, 2011). That is why mutational bias models are often cited to explain

interspecific codon usage variation. Mutational bias appears to be neutral *i.e.* it does not affect the fitness of an organism and act globally on the whole nucleotide sequences of that organism (Knight *et al*., 2001).

1.2.2 Selection and codon usage

The followers of selection theory conceive that the fitness of an organism is swayed by the synonymous mutations in its genome which thereby get upheld or suppressed in the course of evolution. Selective theories are often taken up to explain the variations across a genome/gene. This, however, does not rule out any possibility of its attributability in explaining interspecific variations (Plotkin and Kudla, 2011). Mutation and selection are not reciprocally exclusive forces *i.e.* both can operate at the same time in a genome. For example, mutation might be operative in certain genes in an organism at a time where some other genes of the same organism might be under selection pressure. Xia (1996) opined that in an mRNA where selection operates for optimized transcription the most abundant nucleobases are transcribed at a rapid pace than the rest (Xia, 1996).

It has been hypothesized that in lower organisms such as bacteria with large population size the codon usage is mostly influenced by selection pressure rather than mutation (Ikemura, 1981; Sharp *et al*., 1993). On the contrary, in organisms with lower population size, such as mammals, the effect of selective forces is too miniscule to make a mark (Sharp *et al*., 1993; Duret, 2002). In viruses, especially ssRNA viruses like influenza A virus, mutational pressure has been shown to be the prime force driving the codon usage in these organisms (Greenbaum *et al*., 2008; Wong *et al*., 2010; Goni *et al*., 2012).

1.2.3 Gene expression

It has been well established that the level of gene expression is related to the background codon usage. It was experimentally validated using *E. coli* as a model organism that genes tend to optimize its codon usage towards higher expression. However, it was found out that stronger codon usage bias is related to both highly expressed and lowly expressed genes (Grantham *et al*., 1980a; Gouy and Gautier, 1982; Sharp and Li, 1986). Selection for fidelity and accuracy is another factor that is associated with gene expressivity. In *E. coli* the *Asn* codon AAT is shown to be misread 8-10 times more than its optimal counterpart AAC (Parker *et al*., 1983). Likewise, TTT codon for *Phe* is often mistranslated as *Leu* codon (Parker *et al*., 1992). It was observed by some investigators that conserved amino acids in genes which show meek codon bias, had elevated codon bias than those of non-conserved counterparts (Akashi, 1994). An exploration of homologous genes in *E. coli* and *S. typhimurium* however, led to the counter argument that there was no significant distinction in codon choice in these organisms for conserved and non-conserved amino acids (Hartl *et al*., 1994).

Apart from these main factors there are numerous other factors which play role in dictating an organism's codon choice. These include nucleotide composition (Osawa *et al*., 1988), gene length (Moriyama and Powell, 1998), hydrophobicity (Romero *et al*., 2000), environment effect (Xiang *et al*., 2015) etc. Location in the genome has also some relation with codon usage (Sueoka, 1988; Sharp *et al*., 1993).

1.2.4 Mutation-selection-drift theory in codon usage bias

Two major paradigms- natural selection and mutational bias have been invoked to explain the codon usage strategies in an organism. Although natural selection plays

crucial role in many organisms, it is always difficult to fish out in what form exactly the selective forces are operating, thus making it ineffective to explain the codon choice alone. There are organisms, on the other hand, where mutation alone dictates bulk of the codon usage discrepancy with drift playing a minor part. Bulmer (1987) and Akashi (1994) were of the opinion that observed codon discrepancy is a result of the balance between the selective forces favouring the fixation of beneficial codons and genetic drift causing the fixation of detrimental codons (Bulmer, 1987; Akashi, 1994). The constant urge for developing a combined theory to explain codon usage forced scientists to arrive at a rationally working hypothesis which has become popular in the name of mutation-selection-drift theory (Bulmer, 1987; Akashi, 1994; Hartl *et al*., 1994; Sharp *et al*., 2010).

1.3 Applications of codon usage study

Codon usage studies are useful in many areas of modern biology. From being used for detection of open reading frames (ORFs) and finding protein coding genes in genomes, codon usage bias studies have come a long way forward. It is codon usage analysis that enabled the molecular biologists to go deeper into the use of rare codons in the context of identification of pseudo-genes and DNA sequencing errors within coding sequences (Gribskov *et al*., 1984). Codon usage bias, being different across organisms, is particularly useful in studying horizontal gene transfer event (Carbone *et al*., 2003; Cortez *et al*., 2005; Bodilis and Barray, 2006). Functional conservation of genes and their expression among various organisms can be studied in the light of codon usage bias (CUB) studies.

Perhaps the most important utility of codon usage studies lies in predicting the gene expression level and optimization of the same for better protein productivity.

Heterologous protein production using codon optimized genes are promising avenues for food and biopharmaceutical industries, especially in the field of DNA vaccine, biosimilar production etc (Lithwick and Margalit, 2005; Ruiz *et al*., 2006; Roth *et al*., 2012).

Codon usage bias is also important in the context of protein folding. Synonymous codon substitution can modulate the rate of translation and grant more time to the N-terminal end to fold correctly. Synonymous substitutions involving the rare codons play a critical part in protein folding. The presence of rare codons halts the translation rate and thereby reduces the possibility of protein misfolding. Replacing the rare codons of mRNA with more common synonymous codons implicates faster translation process (Sun *et al*., 2001).

Codon usage study is useful in medical genetics as well. Earlier it was believed that only non-synonymous mutations are responsible for causing diseases. Although most of the disease related SNPs are non-synonymous in nature, a good number of synonymous SNPs are also shown to be involved in triggering various diseases (Sauna and Kimchi-Sarfaty, 2011). In addition, the pathogenic synonymous mutations are able to change certain motifs important for alternative splicing (Faa *et al*., 2010; Daidone *et al*., 2011).

1.4 Influenza A virus

The influenza viruses are a class of single stranded RNA virus with a genome of negative polarity and belong to the family *Orthomyxoviridae*. There are three major types of influenza viruses named as type A, type B and type C. Influenza virus types B and C infects only humans while influenza A viruses can infect humans, birds and some other mammals like pigs, dogs etc. Influenza A virus (IAV) alone infects

roughly 1.5 million people annually across the globe, causing significant mortality and morbidity worldwide. It poses a threat to health and causes significant negative economic impacts on society every year. Based on the variants of two main surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), the IAVs are classified into different subtypes, e.g. H1N1, H1N2, etc. The last century witnessed 3 Influenza A pandemics: H1N1 in 1918, H2N2 in 1957 and H3N2 in 1968 (Webster *et al*., 1992; Cox and Subbarao, 2000). The recent influenza pandemic of 2009 was caused by an H1N1 type (Webby and Webster, 2003; Dawood *et al*., 2009).

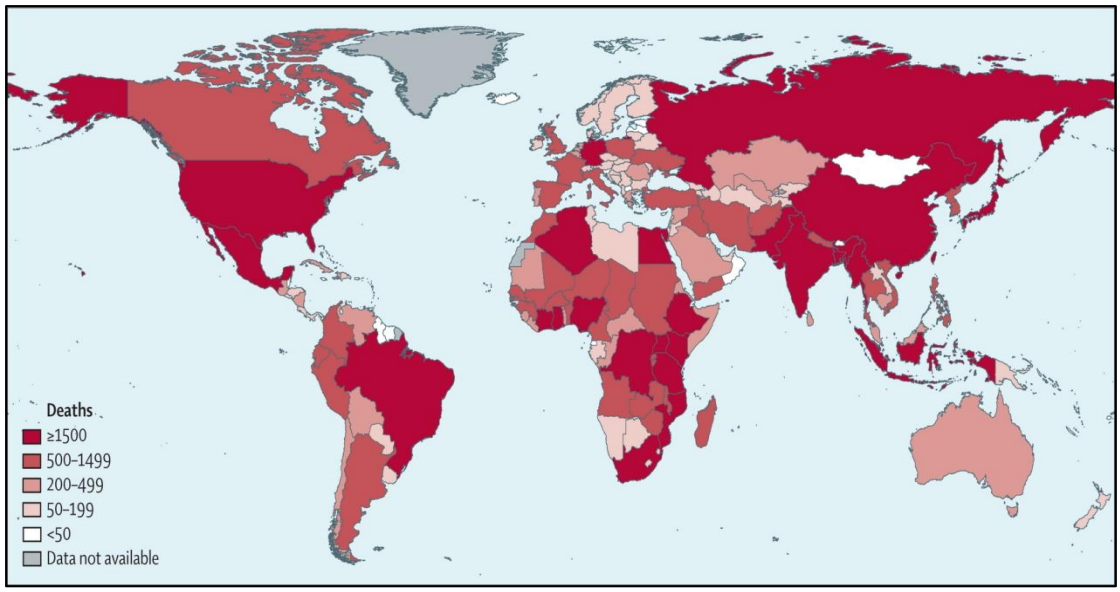


Fig. 1.4.1 Global burden of the 2009 H1N1 subtype of influenza A virus in the first year of infection [Image source: (Dawood *et al*., 2012)]

The replication cycle of the influenza virus depends on host machinery and the virus utilizes host cellular components for its protein synthesis. The genome of the virus is divided into eight segments of negative-sense RNA, which are required to be converted into positive strand in order to replicate inside the host cells upon infection. In order to evade the host immune response, human seasonal influenza virus utilizes a unique phenomenon called antigenic drift by which it changes its antigenicity by introducing novel mutations in its surface proteins (Webster *et al*., 1982). The

proteins HA and NA are the components of the antigenic determinants of the virus that are recognized as pathogen signature by the host's immune system, and thereby elicit an immune response. Owing to the diverse subtypes of these two IAV surface glycoproteins, the human immune system is recurrently confronted with new antigens. Mutations in the HA and NA genes may result in changes in antigenicity that permits the virus to invade people who were previously exposed to the virus. This phenomenon of moving antigenicity is termed as antigenic drift. Antigenic drift occurs when there is a re-assortment of the surface protein segments between viruses, resulting in a virus that is immunologically novel to humans (Gething *et al.*, 1980).

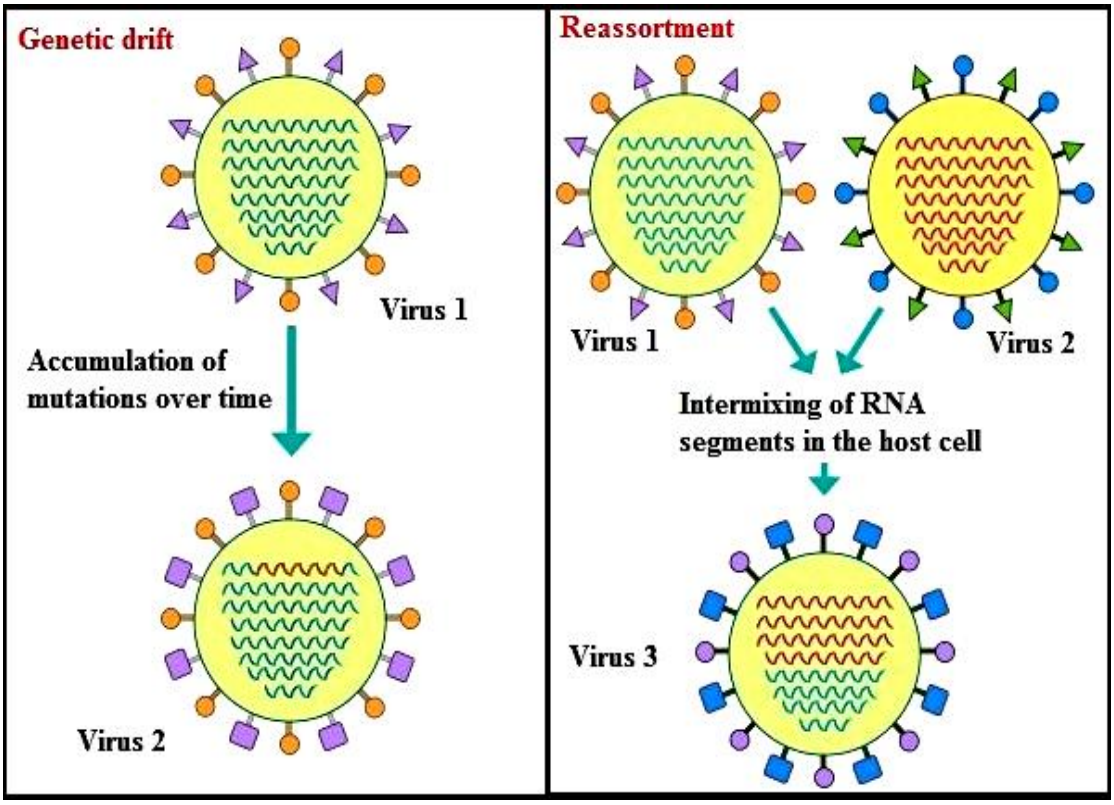


Fig. 1.4.2 Schematic representation of genetic drift and genetic shift giving rise to novel IAV strains

Another process, called re-assortment, is also considered a major force in the evolution of IAV (Hilleman, 2002). It occurs when the virus acquires an HA and/or NA of a different IAV subtypes (via re-assortation) of one or more gene segments.

This process has been the basis of the devastating influenza pandemics that occurred several times in the last century (Ferguson *et al*., 2003). As the IAV genome is segmented into eight parts, the coding sequences are harbored on individual RNA strands. As a result, genome segments get readily shuffled between different strains of the virus that invade the same host cells. For example, let us consider a cell that is infected with IAV from different species. In the cell, as a consequence of reassortment, progeny viruses may arise that contain some genes from strains that are usually known to infect birds and some genes from strains generally infecting the humans. This may lead to the establishment of novel strains that have not been encountered previously. Furthermore, as there are at least 16 different HA subtypes and 9 different NA subtypes characterized till date, numerous combinations of these capsid proteins are possible. Among these, 3 subtypes of hemagglutinin (H1, H2, and H3) and two subtypes of neuraminidase (N1 and N2) have been reported to cause a number of epidemics in the human population globally (Palese, 2004).

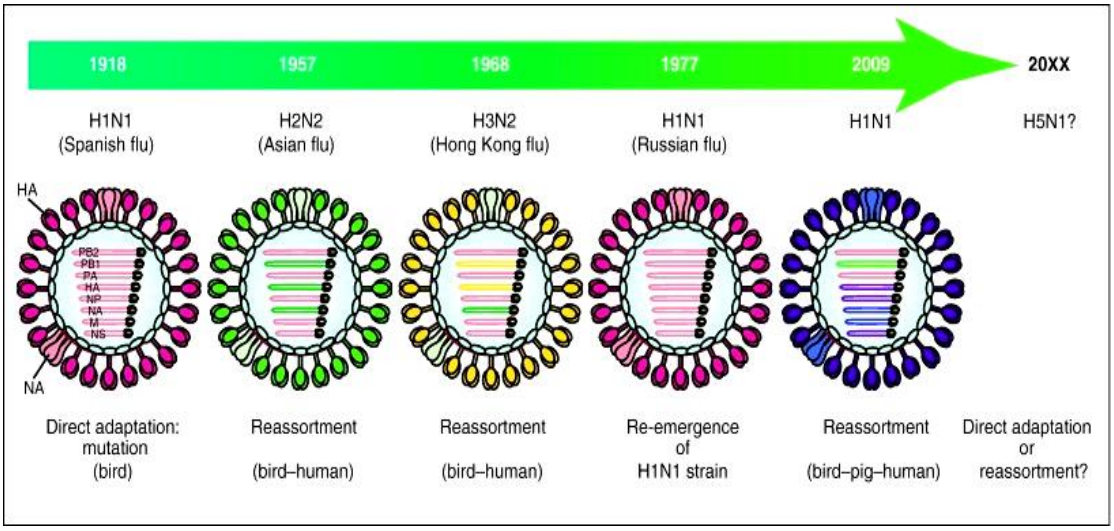


Fig. 1.4.3 Evolutionary timeline of the pandemic influenza A viruses [Image source: (Watanabe *et al*., 2012)]

1.4.1 Influenza A virus subtypes infecting the humans

As mentioned earlier, there are many different subtypes which have already been reported to infect different range of hosts including the humans but not all are known to invade the latter. Generally hemagglutinin subtypes H1, H2 and H3 along with the N1 and N2 subtypes of neuraminidase have recorded human invasion history. Apart from these, H5N1 subtype which is predominantly an avian IAV strain has recently been reported to break the species barrier to infect human populations (Ungchusak *et al*., 2005; Korteweg and Gu, 2008).

1.4.1.1 H1N1

Known as the "Spanish flu", this particular subtype wreaked havoc in 1918 worldwide, killing over 50-100 million people. This is by far the worst pandemic of any influenza A virus known till date. This H1N1 subtype had its origin in a purely avian reservoir (Kobasa *et al*., 2004; Tumpey *et al*., 2005). In 1976, a novel H1N1 strain of swine origin was reported to infect soldiers at Fort Dix, New Jersey (Gaydos *et al*., 2006). In 1977–1978, another strain of H1N1 caused the Russian flu epidemic (Finkelman *et al*., 2007).

In 2009, a novel strain of swine origin H1N1 made its appearance in Mexico which gradually spread to all parts of the world and eventually resulted into a global pandemic claiming around 3.8 lakh lives. This particular strain was later confirmed to be a triple reassortant swine influenza virus (Dawood *et al*., 2009; Garten *et al*., 2009). The WHO declared it to be the first global pandemic since the 1968 Hong Kong flu.

1.4.1.2 H1N2

This subtype is believed to be the result of reassortment of the gene segments from H1N1 and H3N2 subtypes due to similarity in its hemagglutinin protein with that of H1N1; while its neuraminidase protein resembles that of H3N2 subtype. H1N2 is presently endemic in swine and human populations. Although there have not been any recorded pandemics of H1N2, it is considered as a potential threat owing to its close similarity with the H1N1. There are a few cases of H1N2 infecting humans, the first being in 1988-89 from China (Guo *et al*., 1992). The first H1N2 reports from India came in 2001 during the 2001-02 flu season of the northern hemisphere. In 2002, there were a few reported cases of H1N2 from United Kingdom, Israel and Egypt (Xu *et al*., 2002).

1.4.1.3 H2N2

Some researchers affirmed that the 1889-90 “Russian flu” was the result of an H2N2 subtype of IAV, thus making it to be the first IAV outbreak. According to the reports available, about 1 million people died in that pandemic that originated in Russia and later spread to entire Europe, North America, Latin America and Asia (Makarova *et al*., 1999; Tsuchiya *et al*., 2001). An avian origin variant of H2N2 was responsible for the pandemic that originated in China in 1957 and later spread out to other parts of the world. During this pandemic known as "Asian flu" the death toll rose to 1-1.5 million people (Viboud *et al*., 2016).

1.4.1.4 H3N2

H3N2 is a descendent of H2N2 which is considered to be the result of antigenic shift and is currently endemic in the swine as well as the humans. It was responsible for pandemic known as the “Hong Kong flu” that occurred during 1968-1969 and claimed 7.5 lakh lives (Lindstrom *et al*., 2004). These H3N2 viruses had avian origin

hemagglutinin and polymerase 1 (PB1) gene segments whereas rest of the genes originated from human H2N2 viruses (Kawaoka *et al*., 1989). However, these strains of H3N2 are thought to be extinct in the wild and in the later phases predominant reassortment paved the way for adaptation and successful establishment of novel H3N2 lineages in the humans (Lindstrom *et al*., 2004).

1.4.1.5 H5N1

Since 1996, several instances suggest that this avian H5N1 subtype caused outbreaks in different parts of the world, specifically in the Asian countries. In 1997, highly pathogenic avian influenza A (H5N1) virus infected both poultry and humans simultaneously. Although the death toll was not large enough but this was the first instance of an avian virus getting direct entry to humans. This subtype, in fact, crossed species barrier to infect humans, cats, tigers as well as lions which makes H5N1 more deadly than any of the other subtypes of IAV (Skeik and Jabr, 2008; Chen *et al*., 2016). According to the WHO reports the death rate of 1997 outbreak was more than 50% (191 off 317 cases) (Skeik and Jabr, 2008). H5N1 infections of humans and poultry were again reported in China and Hong Kong, Thailand and Vietnam in 2003. In 2005, from five countries *viz.* Cambodia, Indonesia, Thailand, China and Vietnam there were reports of 98 human H5N1 cases with 43 deaths. This was followed by reports of H5N1 outbreaks from other countries like Egypt, Azerbaijan, Iraq, Turkey (2006), and Laos, Nigeria, Pakistan, Myanmar (2007) apart from China, Cambodia, Thailand, Vietnam, Indonesia where majority of the H5N1 cases were reported (Korteweg and Gu, 2008).

1.4.2 Genetic structure of influenza A virus

Influenza A virus genome is a 8-segmented negative stranded RNA which encodes eleven known proteins. Of these eleven proteins, nine are located in the virion while

the remaining two *i.e.* non-structural protein (NS1) and PB1-F2 are encoded in the host cells upon infection by the virion (Schulze, 1970; Palese and Shaw, 2007). Hemagglutinin (HA) and neuraminidase (NA) are the two glycoproteins that are found embedded in the lipid envelope of the virion. These two proteins are responsible for most of the variations observed in different influenza A virus strains (Palese and Shaw, 2007). With a molecular weight of 76,000 approximately, HA is embedded in the lipid membrane in such a way that the major part of it, containing five antigenic domains, is exposed at the outer surface. HA acts as a receptor by attaching to sialic acid (N-acetylneuraminic acid) and facilitates the penetration of the core of the viral particle by the process of membrane fusion (Kamps *et al*., 2006).

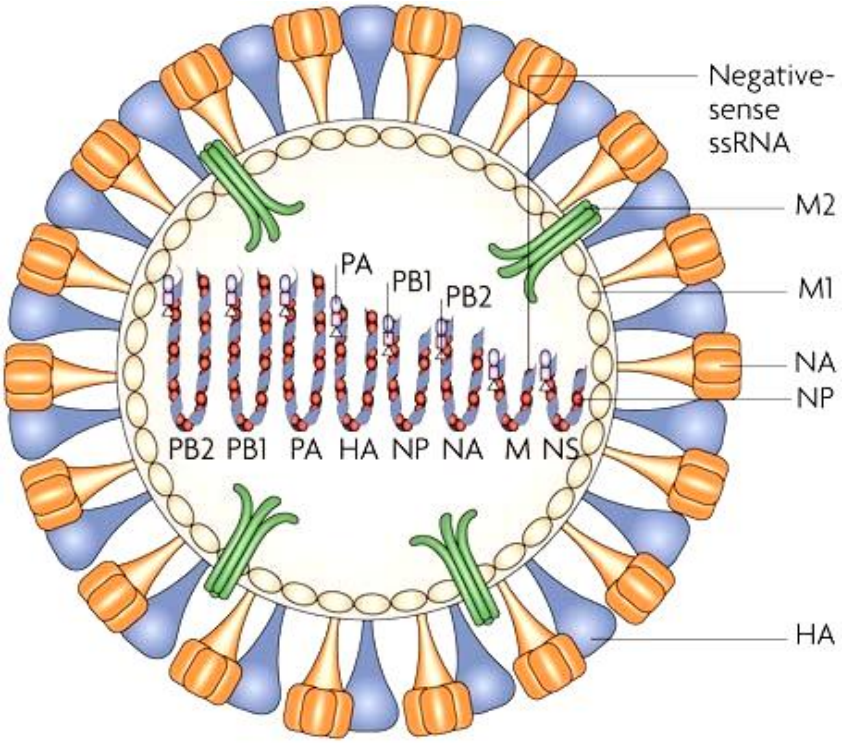


Fig. 1.4.4 Schematic representation of the structure of influenza A viruses [Image source: (Nelson and Holmes, 2007)]

The NA protein is found as protrusions on the viral envelope and forms a tetramer having an average molecular weight of 220,000 approximately. NA spans the lipid layer with its core part projecting outwards from the envelope, with a small cytoplasmic tail towards the interior. NA serves as an enzyme, cleaving the sialic acid residue from HA, other NA molecules and the surface glycoproteins and glycolipids. It also acts as a key antigenic site, and seems to be crucial for the virus entry through the respiratory epithelium in the host (Kamps *et al*., 2006).

Beneath the viral envelope lies the layer of matrix protein 1 (M1) which encircles the 8 segmented ribonucleoproteins composed of the RNA encrusted with the nucleoprotein (NP). The matrix protein 2 (M2) acts as the ion channel by inducing a low pH during viral entry to the host. It disrupts the HA-M1 binding opening up the viral particle as a result. This is followed by the release of viral ribonucleoproteins into the cytoplasm of the host cells which kicks off the viral RNA synthesis. The trimers of polymerase proteins (PA, PB1 and PB2) are found attached to these ribonucleoproteins. Nuclear export protein (NEP) is another protein found in the virus particles (Compans *et al*., 1970; Skehel and Schild, 1971).

1.5 Statement of the problem

Codon usage is a significant feature of almost all the organisms, which influences their genetic make-up. Thus codon usage analysis is immensely helpful in understanding the genetic and evolutionary characteristics of an organism. Influenza A virus is entirely dependent on its host cellular machinery for replication of its genome and consequent establishment of the virus inside the latter for successful invasion. Thus the viral genome has to adapt to the host conditions and find its way to escape the host immunity to survive. The host in turn tries to prevent such adaptation

and this phenomenon leads to the host selection pressure which is crucial in the context of virus genome evolution. It is thus extremely important to get insights into the viral genome signatures which are reflected in its codon usage. Influenza A virus has high mutation rates which enables it to escape host immune system. Previously many studies have been carried out involving different strains of IAV in different hosts. But there is still scope of detailed studies regarding this ever-changing viral pathogen which is constantly increasing its range of hosts coupled with generation of novel strains that can cross species barrier. Recently an avian strain of H5N1 broke the species barrier to infect humans and recorded more than 50% mortality in humans.

This study seeks to understand the variation in codon usage of five IAV subtypes that have been found circulating among the humans in varying degree. Sporadic studies have been carried out to study IAV codon usage in human hosts; however, no attempt has been made to perform a comparative study involving these five subtypes together. We believe such studies would help us gain insights into the evolutionary aspects of this immensely important viral entity that has been a serious global threat both in terms of mortality and morbidity.

1.6 Rationale of the study

- ✓ The replication cycle of the influenza virus depends on host machinery and the virus utilizes host cellular components for its protein synthesis. Therefore codon usage in this virus and its hosts could be expected to affect viral replication.
- ✓ The effect of selection pressure imposed by the human host on the codon usage of human influenza viruses and trends in viral codon usage over time needs to be investigated.

- ✓ A detailed understanding of the basic biology of this virus, especially its evolution and methods for host adaptation, is needed to prevent future pandemics.

1.7 Objectives

Keeping the above points under consideration, the present study was formulated with the following objectives:

1. To analyze the codon usage pattern of the genomes of different influenza *A* virus subtypes

2. To investigate the overall nucleotide and codon position specific nucleotide composition in the coding sequences of influenza *A* virus genes

3. To compare the codon usage patterns across different genes of influenza *A* virus

4. To predict the expression of important genes using nucleotide determinants in influenza *A* virus

5. To characterize the basic biochemical properties of proteins (GRAVY score, hydrophilicity, pI) encoded by the influenza *A* virus genes