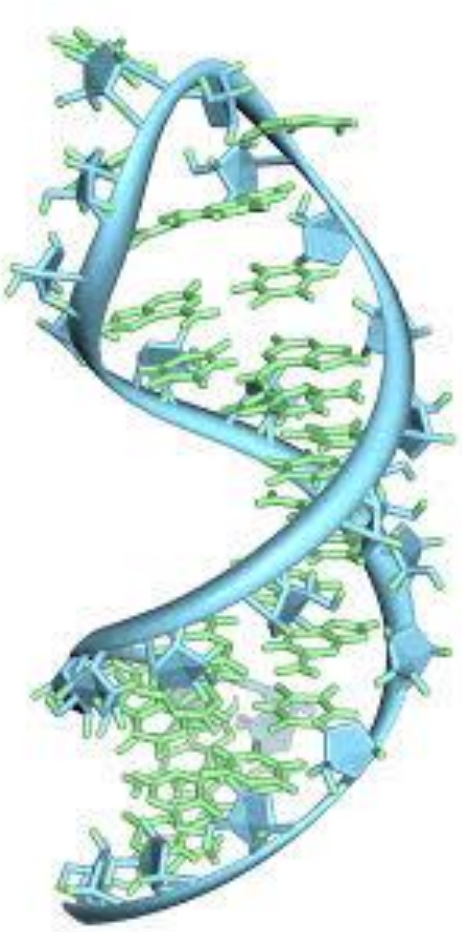
Chapter

2



REVIEW OF LITERATURE

CHAPTER 2

REVIEW OF LITERATURE

Analyses of synonymous codon usage in organisms are vital to unravel fundamental characteristics of evolution at fine scale and also provide clues to the practical molecular biological applications. Until 1970 it was believed that variations at molecular level do not affect fitness of an organism, following the controversial 'neutral evolution' hypothesis (Kimura, 1968). In the year 1970, Clarke reported unequal usage of codon usage (Clarke, 1970). In the year 1980, Richard Grantham came with the 'genome hypothesis' which states that different genes have their own codon usage strategies (Grantham *et al.*, 1980b).

In the 1980s the codon usage model was largely limited to the model organism *E. coli* because of the availability of its complete genome. Back then it was believed that codon usage bias was more pronounced in the highly expressed genes and it was also experimentally proved in *E. coli.* First it was the works of Ikemura (1981) and later some more investigators consolidated the observation (Ikemura, 1981; Gouy and Gautier, 1982). Ikemura (1982) opined that in yeast, the codon usage bias is mainly anchored by the abundance of isoaccepting tRNA pool (Ikemura, 1982). This has lead to the concept that in most in organisms the genomes tend to select translationally optimal codons.

In the late 1980s, the reason for selection of optimal codons was seriously debated by many investigators. The traditional view was that use of optimal codons was directly linked to the efficient translation process (Kurland and Ehrenberg, 1984; Andersson and Kurland, 1990). It was believed that the optimal codons are translated quicker than the non-optimised codons and hence the ribosomes would move faster along the

mRNA where optimal codons are more abundant. As a result those ribosomes would be released quicker and available for the next mRNA for translation (Sørensen and Pedersen, 1991; Kudla *et al*., 2009).

The second view states that use of optimal codons facilitates translation process to be accurate. In *Drosophila melanogaster* it was observed that the highly expressed genes displayed higher codon usage bias (Shields *et al*., 1988; Duret and Mouchiroud, 1999). Akashi (1994) showed that codon usage bias was stronger in the conserved amino acids in the fruit fly (Akashi, 1994). Later a variety of investigations consolidated the view of translational accuracy to be the primary evolutionary force exerting on the codon usage, especially in bacteria and eukaryotes (Stoletzki and Eyre-Walker, 2007; Drummond and Wilke, 2008).

Most of the reports in the 1980s advocated for translationally optimal codons and level of gene expression to be the driving force behind codon usage bias. However, most of these investigations were covering the bacteria and yeast (Grosjean and Fiers, 1982; Ikemura, 1985; Bulmer, 1987; Sharp and Li, 1987). According these hypotheses, the highly expressed genes tend to use the so-called 'preferred codons' or 'major codons' *i.e.* the synonymous codons with high availability of the isoaccepting tRNA molecules. On the contrary, the lowly expressed genes seemed lack these codons and instead use the 'minor' codons. Thus, the nucleobase composition in the highly expressed genes deviates from the expected composition and the mutational equilibrium.

Although similar observation was reported in some multicellular organisms like *Drosophila* (Shields *et al*., 1988) and *Bombyx mori* (Chevallier and Garel, 1979; Garel, 1982), the same hypothesis, however, did not seem to be complete enough to

explain the codon usage variations shown by different tissues and developmental stages in the higher organisms (Comeron and Aguade, 1998). The warm blooded vertebrates having genomic isochores showed different nucleobase composition which was believed to be the result of mutational biases (Aota and Ikemura, 1986; Bernardi and Bernardi, 1986; Filipski, 1987). This resulted in difference in codon usage among the different levels of organization in the same organism.

A number of investigators tried to quantify the degree of codon usage bias among the genes of a species as well as across different species using various measures available at that time. These different methods of codon usage study fell broadly into two main categories. The first category followed the null hypothesis that use of synonymous codons was a random process. Scaled chi-square (Shields *et al*., 1988), the effective number of codons or ENc (Wright, 1990), codon bias index (Morton, 1993) and intrinsic codon bias index or ICDI in short (Freire-Picos *et al*., 1994) come under this category. The second category of codon bias indices includes the frequency of optimal codons (Fop) (Ikemura, 1981) and the codon adaptation index (CAI) (Sharp and Li, 1987). These indices estimate codon usage bias by a comparison of the observed frequency of the various synonymous codons and the frequency of the preferred codons obtained from a reference gene set.

Xia (1996) proposed transcriptional selection to be the prime evolutionary force shaping the codon usage at the mRNA level. He was of the opinion that the genes with more abundant codons are transcribed faster. The presence or absence of optimal codons can affect the mRNA folding and decay (Xia, 1996).

Codon usage studies have been carried out in almost all kinds of living beings starting from bacteria to higher organisms including humans. While in bacteria, the highly expressed genes strongly favours the use of optimal codons and consequently bringing selection into play (Gouy and Gautier, 1982; Sharp and Li, 1986; Akashi and Eyre-Walker, 1998; Moriyama and Powell, 1998), the mammalian genomes tend to be under mutational pressure (Wolfe *et al*., 1989; Sharp *et al*., 1993; Francino and Ochman, 1999). The reason for mutational bias in mammals could be due to dependencies of the genes in various chromosomal locations. As the population size of the mammals is much less in comparison to the likes of bacteria, selection pressure becomes too minuscule to exert a profound effect in mammalian genome. The comparatively undersized population means that genetic drift will lead the mutational evolutionary dynamics that vary only marginally in fitness (Jenkins and Holmes, 2003).

According to Drake and Holland (1999), the rate of spontaneous mutations is a crucial factor in determining the genetic framework of a population which reveal much of their evolutionary development (Drake and Holland, 1999). Mutation imparts the platform for the necessary variations for evolution which is manoeuvred by selection, recombination and random genetic drift. The RNA viruses have much higher rates of mutation rates as compared to the DNA viruses. Thus, mutation could be the prime force of evolution in the RNA viruses rather than selection for optimal translation.

Hembert and Berkhout (1995) reported inclination of HIV-1 genome towards over-use of adenine (A). They demonstrated an inverse correlation between the codon bias and translation rate in these viruses. It was observed that the inclination toward A-rich codons was more prominent in *pol* genes; while less adenine abundance was observed

in *tat-rev-env* reading frames as well as in the overlapping genes of the regulatory sequences in *nef*-LTR region. They proposed that the availability of aminoacyl-tRNA in the host cell confines the lentiviral partiality for codons rich in A-content.

Karlin *et al.* (1994) opined that virtually all small eukaryotic viruses tend to suppress their CpG content in the genome (Karlin *et al*., 1994). They attributed the lower usage of CpG to the lack of proofreading as well as mismatch repair system in their genome. This phenomenon, according to them, helps the viral genomes to replicate rapidly and efficiently and consequently facilitating their rapid evolution.

In papillomavirus, codon usage bias has been linked to the availability of tRNA (Zhou *et al*., 1999). Papillomavirus is a DNA virus which shows much less rate of mutation as opposed to the RNA counterparts. They took help of a codon replacement experiment using green fluorescent protein (*gfp*) to support the hypothesis that differences in translational efficiency was associated with codon usage rather than mRNA structure. They speculated that availability of matching tRNA to codon usage might be a factor in restricting the expression of papillomavirus genes.

Haas *et al.* (1996) suggested codon re-engineering as a potential method of enhancing protein production of certain beneficial poorly expressed mammalian genes. While working on the expression of HIV-1 surface glycoprotein they observed that selection in background codon usage was responsible for inefficient protein production in the virus. A model protein Thy-1 wa re-engineered with the most preferred HIV-1 codons which resulted in reduced Thy-1 expressivity. Changing the coding region of the *gfp* protein of jellyfish into preferred codons of highly expressive proteins of the humans however substantially increased the efficiency of expression (Haas *et al*., 1996).

In RNA viruses, mutation pressure gains the upper hand over selection to shape up the codon usage of these viruses. Given the high population sizes of the RNA viruses, it looks absurd. Brown and Leigh (1997) provided an explanation of this observation. According to them, mutation rates in RNA viruses are too high owing to the lack of proofreading activity and mismatch repair mechanisms. Also, fitness effect of the synonymous codon preferences is too weak for translational selection to act efficiently. Thus, tremendous mutation rate overpowers the influence of selection in these viruses. Further, in some cases, the effective population size *i.e.* the number of contributing progeny viruses to the next generation, might be inadequate for genetic drift to take control of molecular evolution.

Rima and McFerran (1997) did a comprehensive study on the use of dinucleotides in 64 single stranded RNA viruses (Rima and McFerran, 1997). They found a strong inverse correlation between the dinucleotide frequency and the GC-content in these viral genomes contradicting the findings of Karlin (Karlin *et al.*, 1994). According to them, the lack of CpG was coupled with a corresponding elevation of the CpA and UpG frequencies. Additionally, they opined that the odds ratios calculated for dinucleotides are in fact not independent variables and the over-representation of CpA and UpG in their findings were as a consequence of UpA and CpG containment in the viral genomes.

Mutational pressure against selection was also projected as the major determinant of codon usage in the nucleopolyhedroviruses (Levin and Whittome, 2000). According to their findings GC-mutational bias plays a significant role in the codon usage profile in these viruses. They did not find any striking difference in the codon usage fingerprints of the homologous genes encoded by different nucleopolyhedroviruses.

They also ruled out any significant correlation between the length of genes and codon bias.

Guan *et al.* (2000) reported some 'H5N1-like' genome complex in a H9N2 subtype that was found infecting humans in Hong Kong (Guan *et al*., 2000). They compared these strains with the H9N2 strains circulating in poultry in the region in 1999. The study established that the two lineages of viruses (Qa/HK/G1/97 and Dk/HK/Y280/97) were prevalent in the poultry in Hong Kong and the Qa/HK/G1/97-like viruses were prevailing in the poultry markets. The study provided initial characterization of these viral strains in their respective hosts and highlighted the need for sustained scrutiny of the viruses in poultry and mammals having the H5N1-like replicative complex.

Jenkins and his co-workers (2001) observed difference in base composition and the codon usage profiles of the flaviviruses opting for different vectors (Jenkins *et al*., 2001). Flaviviruses can alternate between host with the aid of vectors like ticks, mosquitoes or by direct transfers between hosts. They had found that the tick-associated flaviviruses were low on GC-content as compared to those transferring directly between hosts. The difference was thoroughly distributed across the genome at all codon positions. The mosquito-borne flaviviruses showed intermediate GC-content between the two aforesaid counterparts. The variations were, however, as a result of weak selection operating at the silent sites. The authors concluded that there could be an association of the codon usage with the vector specificity albeit different from associations previously reported in other virus families.

A comprehensive analysis of the codon usage profiles of 50 RNA viruses including the influenza viruses were compiled by Jenkins *et al.* (2003) which advocated that the overall degree of codon discrepancy in RNA viruses is low and these viruses do not display much variation in between genes (Jenkins and Holmes, 2003). They also opined that there exists strong correlation between the base composition and codon usage bias as well as dinucleotide usage frequency which means that mutational pressure, rather than natural/translational selection, plays crucial role in the codon usage of these viruses. However, they also found some relationships between structural and environmental factors with codon usage in certain viruses.

In a research work involving severe acute respiratory syndrome *Coronavirus* (SARSCoV) virus, Wanjun Gu *et al.* compared the codon usage of SARSCoV to a 10 evolutionarily related viruses (Gu *et al.*, 2004a). Although they found differences in codon usage among the viral groups, the variations were too slight. They concluded that mutational constraint was the determining force working in the genomes of these viruses. Further, they mentioned that the bias was phylogenetically conserved but not host-specific. Gene function was also found to have alliances with the codon usage bias in SARACoV.

Tong Zhou and his fellow co-workers of the Southeast University, Nanjing, China, documented the codon usage profiles of the H5N1 subtype of influenza A virus (Zhou *et al.*, 2005). They compared this subtype with related influenza A viruses and influenza B virus and found that codon usage bias in H5N1 is slight and mainly occurred due to mutational pressure and compositional bias at synonymous position. Further, the synonymous codon usage in H5N1 and other influenza A virus genes were found to be phylogenetically conservative, however it was not strain-specific.

Synonymous codon usage in the genes from different influenza A viruses were genus conservative.

Goni *et al.* (2012) investigated the codon usage profiles of the 2009 H1N1 pandemic strains of influenza A virus (Goni *et al.*, 2012). They had reported high codon usage bias in the amino acids alanine, arginine, proline, threonine and serine. The general association of codon usage and nucleobase composition along with lack of optimization to host tRNA pool suggested that mutational pressure could be the dominant evolutionary constraint exerting on these strains.

An important model for the studies seeking insights into the viral persistence and the response imposed by the host immune system during lentiviral infection is the equine infectious anemia virus (EIAV). Yin *et al.* (2013) studied the codon usage in EIAV and its hosts in view of the viral evolution. They took 29 whole genomes of EIAV for the study and found slight bias in them. Mutation pressure was found to be dominant among these viruses, while other environmental factors were also hinted to be present in shaping up the viral evolution. However, they also mentioned that the data sample size was too small to make any conclusive statements about EIAV evolution.

Youhua Chen (2013) made a comparative investigation on some RNA and DNA viruses in context of their differences in codon choices (Chen, 2013). The results of the study comprising about 38000 ORFs stated that approximately 27% and 21% of the total variation in codon usage could be attributable to mutational pressure whereas about 5% and 6% were due to natural selection in the DNA and RNA viruses respectively.

Cheng *et al.* (2013) studied CpG usage in the RNA viruses and compared the odds ratio of CpG usage with genome polarity, nucleobase composition and various other factors along with the host conditions (Cheng *et al*., 2013). They found host selection pressure, underlying codon usage bias and the nucleobase compositional pressure to be the main driving force behind the usage of CpG in the RNA viruses. They attributed the under-presentation of CpG in negative ssRNA viruses to the base compositional constraint imposed due to AU-rich genome. On the other hand +ssRNA viruses were opined to be under host selection pressure. The +ssRNA viruses tend to mimic their host codon usage profiles according to them.

The Russian-German duo of Ilya S. Belalov and Alexander N. Lukashev (2013) investigated the causes and implications of discrepancy in codon usage in the RNA viral genomes (Belalov and Lukashev, 2013). According to their illustration, genomic GC-content was a poor index of viral codon usage and genomic composition. They opined that nucleobase constitution was the sole major determinant for tilted codon usage in these viruses. Additionally, dinucleotide composition at 2^{nd} and 3^{rd} codon positions also had some effect on the usage of synonymous codons in these viruses.

Butt *et al.* (2014) analysed the genome-wide codon usage in chikungunya viruses (CHIKV) (Butt *et al*., 2014). They reported a G/C and A-ended codon preference in CHIKV. According to the authors, the codon usage of CHIKV was the result of interplay between the antagonistic and coincidental relationship with the host. Apart from compositional bias, CHIKV genomes were also under slight selective constraint arising from the host and environment.

Wang *et al.* (2014) demonstrated the codon usage profiles of swine origin H1N1 viral strains and studied their adaptation to the hosts (Wang *et al*., 2014). Phylogenetic analysis on the hemagglutinin epitopes suggested that the swine-origin H1N1 were gradually adapting to the human host conditions. According to the authors, a positive selection in the epitopic region of the hemagglutinin protein facilitated this adaptation. Studies of the amino acids from the epitopic region revealed a connection between the swine-origin H1N1 and the Spanish flu viral strains.

Christina *et al.* (2016) studied Zika viruses in view of their codon usage profiles (Cristina *et al*., 2016). The codon usage of the viral strains did not differ significantly. However, variations were noted between strains from human hosts and those from mosquito vectors. There was high preference towards A-ending codons. GC-compositional pressure as well as dinucleotide usage bias was also noted in the Zika viruses. Thus mutational pressure was found to be more pronounced in these viruses.

The analyses of codon usage in the surface glycoproteins of nuclear polyhedrosis virus (NPV) were carried out by Zhou and his co-workers (Zhao *et al*., 2016). The study revealed weak bias prevailing in these genes encoding the glycoproteins. Neutrality analysis suggested natural selection to be the prime selective force operating in these genes. The impact of mutational pressure was less pronounced and cluster analysis presented two main clusters of the 18 NPV species used in the study.

Butt and co-workers (2017) reported codon usage bias in Zika virus and concluded that the evolution of the virus is host-specific as well as vector-specific (Butt *et al*., 2016). In the coding sequences of Zika virus they noticed several genotype-specific and commonly occurring codon usage traits. Natural selection was found to be more pronounced in these viruses. The interplay of codon adaptation-deoptimization were

speculated to be the reason behind Zika viruses' successful adaptation to different hosts and vectors.

Heather *et al.* (2017) reported strong role of purifying selection on the codon usage of *Drosophila melanogaster* (Machado *et al*., 2017). Using polymorphism data and population genetic studies they demonstrated that purifying selection on usage of preferred codons varied from weak to strong in these genomes. The results showed that codon usage scenario in these genes are attributed to the distribution of selection coefficients at various sites.