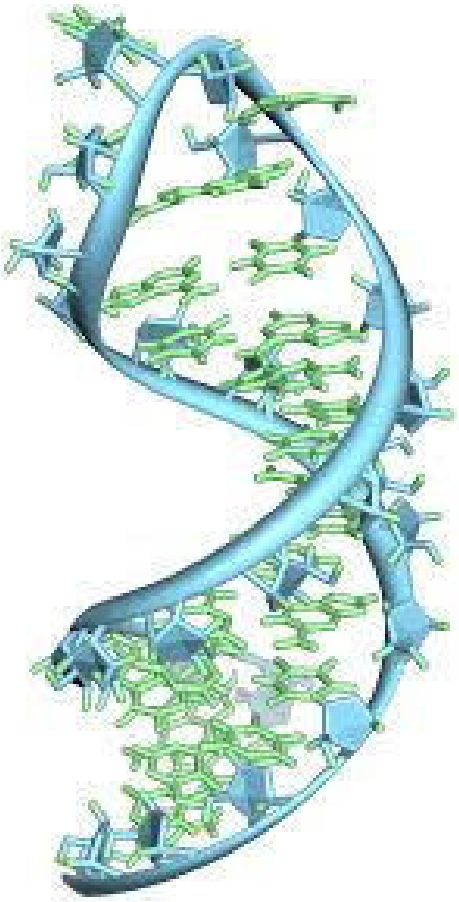
Chapter

5



DISCUSSION

CHAPTER 5

DISCUSSION

Until 1970s it was believed that variations at molecular level do not affect fitness of an organism following the controversial 'neutral evolution' hypothesis by M. Kimura (Kimura, 1968). However, after the reports of unequal codon usage by Clarke in 1970, molecular biologists started looking the whole phenomenon in a holistic manner (Clarke, 1970). In 1980 Richard Grantham put forth the 'genome hypothesis' which states that the different genes in an organism have their own coding strategies (Grantham *et al*., 1980a; Grantham *et al*., 1980b). With the profound increase in genomic data more organisms were brought under the purview of the codon usage studies and it was evident that codon usage in most of the organisms is not uniform. This phenomenon, known as 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 genome 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).

Codon usage bias has been reported in almost all kind of organisms including bacteria, viruses, and higher animals as well as plants (Moriyama and Powell, 1998; De Amicis and Marchetti, 2000; Duret, 2002; Gu *et al*., 2004b; Bodilis and Barray, 2006; Butt *et al*., 2014). The codon usage is generally attributed to the balance between natural selection and mutation pressure (Sharp *et al*., 1993; Shackelton *et al*., 2006). Mutation, selection, and random drift constitute the major evolutionary forces that shape codon usage bias among species (Sharp and Matassi, 1994; Akashi *et al*., 1998; Rocha, 2004; Vicario *et al*., 2007). Mutation bias has been reported as 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 envisaged 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)

5.1 Codon usage bias and molecular evolution

From the evolutionary perspective of the viruses, it is essential to understand the degree and causes of biases in codon usage and nucleotide composition, especially the interaction between the virus and immune response imposed by the host (Shackelton *et al*., 2006). Besides, codon usage bias studies have enabled to unravel striking molecular patterns which in turn have opened avenues for novel hypotheses about protein synthesis and cellular fitness. Quantification of the compositional dynamics in the initiation and elongation regions proved to be an important tool in obtaining information about natural synonymous discrepancy. This has tremendous influence in the design of transgenes in applied perspective (Plotkin and Kudla, 2011).

The signal peptides are indispensable for sorting and export of proteins through the secretary pathways. It has been shown that these peptides constitute key structural features which facilitate the export of proteins to the periplasm. Studies involving codon usage bias in a signal peptide has shown there is high preference of non-optimal codons in the coding sequences of the signal peptide (Zalucki *et al*., 2009).

Several viruses are known for *de novo* assembly of protein-coding genes by substituting ancestral reading frames with novel ones. This reading frame overprinting mechanism is believed to play a crucial role in pathogenicity of the viruses. Study of compositional dynamics in these *de novo* proteins would significantly improve the

knowledgebase of host-pathogen co-evolution along with the functional and structural aspects of the viral proteins. However, identification of the *de novo* frames is a daunting task. Codon usage bias studies come as a great aid in this context. A comparative analysis of codon usage in the overlapping genes with known ancestral frames has shown to be useful in detecting *de novo* reading frames (Pavesi *et al*., 2013). In fact, these investigators have shown that codon usage bias has been successful in resolving *de novo* origin of reading frame which was not been possible by phylogenetic analysis in *Deltaretrovirus* and *Alphanodavirus*.

Use of codon-optimised genes in the generation of attenuated virus for vaccine development has revolutionised the vaccinology field. This approach is called "synthetic attenuated virus engineering" or in short, SAVE. Using the reverse vaccinology approach, synthetic attenuated poliovirus was developed by inserting rare codons which were absent in its wild type. (Coleman *et al*., 2008b). Codon-optimised hemagglutinin gene was used for generation of DNA vaccines against avian H5-type influenza A viruses (Jiang *et al*., 2007).

Further, codon usage bias studies have been particularly useful in understanding heterologous gene expression. Coupling codon usage studies with modern techniques like mass spectrometry and fluorescence microscopy have enabled investigators to measure the endogenous protein levels. Recent developments in the study of ribosomal occupancy quantification and measurement of protein elongation dynamics, along with the aforesaid techniques have been extraordinary in thorough understanding of the basic cellular processes, with implications for understanding codon usage bias (Plotkin and Kudla, 2011).

It is however noteworthy that, principles regarding the heterologous usage of codons and its relationship with protein expression may vary with that of endogenous proteins. The codon optimization studies typically involve optimization of codon usage to the cellular abundance of tRNA in ideal conditions, ignoring other elements of bias. However, recently, more holistic approaches are being employed relating codon usage to the role of global usage of nucleobase, local folding of mRNA and bias in codon pair context (Han *et al*., 2004; Kudla *et al*., 2006; Nackley *et al*., 2006; Coleman *et al*., 2008a).

5.2 Codon usage in influenza A virus subtypes

The present study investigated the codon usage profiles of five subtypes of human influenza A virus (IAV) covering 8 major genes of the virus. The results were suggestive of a weak codon usage bias prevailing in these genes which was reflected by higher Nc (>40) values. The overall Nc values of genes were in the range of 47.51-58.48 considering all the subtypes together. Previously many authors had reported lower codon usage bias in IAV (Zhou *et al*., 2005; Dawood *et al*., 2009; Garten *et al*., 2009; Ahn and Son, 2010; Goni *et al*., 2012; Chen *et al*., 2016). In fact, lower codon usage bias has been found in many RNA viruses (Jenkins and Holmes, 2003; Greenbaum *et al*., 2008; Wong *et al*., 2010; Chen, 2013). Jenkins *et al* (2003) had reported an average Nc value of 50.9 in human RNA viruses including IAV (Jenkins and Holmes, 2003).

Relative synonymous codon usage (RSCU) is a universal tool used in most of the investigations involving codon usage bias (CUB) analysis. It presents an idea of the trends in non-uniform usage of codons. Generally, a codon with RSCU value less than 0.6 indicates under-representation of that codon, whereas RSCU value more than 1.6

reveals over-representation of the codon in the gene. In between these two extreme values, the codons are said to be abundantly and more or less uniformly used (Behura and Severson, 2013). RSCU analysis in our study revealed a differential picture of codon choice across the IAV subtypes. While there were some similarities of codon usage between the subtypes H1N1 and H1N2, the other three subtypes namely H2N2, H3N2 and H5N1 presented mostly different choices of codons. Differences were also observed between the genes within a subtype. The amino acid leucine, for instance, displayed as many as four different codons as preferred one for different genes within H1N1 subtype. Similar kind of observations was also recorded for other subtypes as well, however, the degree of preference varied across the subtypes. Broadly, AGA, CCT, ACA and AGT were some of the most preferentially used codons in these IAV subtypes.

The use of rare codons serves as translational pause sites or slowing down the translational process of a gene which is a very crucial factor during protein folding (Sun *et al*., 2001). We performed rare codon analysis in the IAV subtypes which revealed that the codons containing CpG dinucleotide *i.e.* CGN and NCG type codons were used at a very less frequency. Codons CGC, TCG, CGT were not often used, whereas some other codons like CCG, ACG, CGA, CGG and GCG were also suppressed to a great extent. However, they varied in usage magnitude across the subtypes.

To inspect whether the synonymous codon usage in the genes was statistically different we employed a G-test for equal usage of synonymous codons of 18 amino acids among the genes of all the subtypes. G-test follows a chi-square distribution. Synonymous codons of an amino acid significantly differ in usage at 5% if G value is

less than $p<0.05$ for that amino acid. Most of the G-values in our analysis showed p value less than 0.01 meaning that the usage of synonymous codons encoding the specific amino acid is statistically different.

Frequency of optimal codons (Fop) analysis revealed that the usage of synonymous codons in the IAV genes is not uniform. Fop can be defined as the ratio of the number of optimal codons used in a gene to the total number of synonymous codons in that gene. Fop values will be close to unity if the synonymous codon usage is random. The observed results suggested that the average Fop values are in the range of 0.22-0.40. This means that the synonymous codons were not randomly used and that a certain level of codon bias existed in these genes. This observation was validated by a significant positive correlation ($r = 0.350$, $p < 0.01$) between Fop and the codon usage bias parameter Nc.

5.3 Compositional features of the IAV genes among the subtypes

The IAV genes were analyzed for nucleobase composition which revealed that the purines A and G were most abundant nucleobases. Different observation was recorded for the nucleobase composition at the 3rd codon position which followed the usage order of A3>T3>C3>G3. The 1st and 2nd codon positions however followed the general trend of nucleobase usage. Ironically, H5N1 subtype differed among the subtypes in the usage of nucleobase at synonymous codon position, especially in case of M1 gene where G was the most preferred base at 3rd codon position. The IAV genes in all the subtypes were low in overall GC content (Mean ± SD = 44.5 ± 1.85). M1 gene recorded the highest GC content across all the subtypes. The subtypes H3N2 and H5N1 showed somewhat similar GC content, both overall GC as well as GC at 3rd codon position.

Dinucleotide bias had been reported to influence the codon usage patterns in many viruses (Karlin *et al*., 1994). RSCU and dinucleotide analyses exposed a preference of A/T ending codons and significant repression of codons containing the dinucleotide CpG. This notable depletion of codons with dinucleotide CpG reiterated the previous finding of low CpG usage in this single stranded RNA virus (Karlin *et al*., 1994; Greenbaum *et al*., 2008; Cheng *et al*., 2013). The CpG avoidance is linked to the evolutionary selective pressure as reported in many RNA viruses in previous studies (Rabadan *et al*., 2006; Zhong *et al*., 2007). The IAV strains evolving in avian hosts following the 1918 H1N1 pandemic were thought to be under tremendous selective constraint to cut down their CpG content (Wong *et al*., 2010; Goni *et al*., 2012). The CpG deficiency was also expected to be a means of immunological escape as unmethylated CpGs are recognized as signals of the invading pathogen by the innate immune system of the host (Karlin *et al*., 1994; Shackelton *et al*., 2006; Greenbaum *et al*., 2008).

The overall usage of amino acid was not much different among the IAV subtypes. Leu was the most abundant amino acid (8.4%) followed by Ser (7.8%) and Glu (7.7%); whereas tryptophan (1.5%), histidine (1.9%) and cysteine (2.0%) were some of the least frequent amino acids. We observed a slight deviation in M1 gene where Ala was the most favoured amino acid with an increased 10.5% of usage. It is noteworthy here that H5N1 did not follow this trend with Ala usage of 6.6%.

5.4 Role of compositional constraint in codon usage

A scatter plot of average GC content in the first two codon positions (GC12) against the GC3 content, popularly known as the neutrality plot, is used as an interpreter of the mutation/selection equilibrium in codon usage bias analysis (Sueoka, 1988). We

observed statistically significant positive correlation (Karl Pearson) between GC12 and GC3 contents of genes in all the subtypes except H1N1**.** The slopes however varied in magnitude. The general association of base composition with codon usage showed that mutational pressure might be the more pronounced constraint than other selective forces inflicting codon usage bias in most of the IAV genes. Prevalence of mutational pressure in codon usage had been reported in many RNA viruses including IAV (Guo *et al*., 1992; Greenbaum *et al*., 2008; Ehrhardt *et al*., 2010; Goni *et al*., 2012). The codon usage of H1N1 subtype however seems to be under selective forces as neutrality analysis suggested that the relative effect of mutational pressure is very weak. However, with the large size of RNA virus population, the effect of mutational pressure is too overpowering to the effect of selection to make a mark (Jenkins and Holmes, 2003). Nonetheless, there could be other factors responsible for the variations occurring in the IAV codon usage profile.

5.5 Comparative account of the codon usage patterns across the IAV subtypes

To check whether the disproportionate codon choices are restricted to the genes with higher degree of bias, we employed a Parity Rule 2 (PR2) analysis and examined the strand compositional bias in the IAV subtypes (Sueoka, 1995). We looked for PR2 bias fingerprints taking 58 synonymous codons leaving aside the stop codons along with the codons for Met, Trp and ATA codon of Ile to avoid asymmetry in data. The overall PR2 signatures did not show much deviation among the subtypes. The purines (A and G) were more favoured over the pyrimidines (C and T) at the synonymous sites in all the subtypes. However, there was visible bias in PR2 among subtypes when we took degeneracy level into consideration. The 4-fold codons in all subtypes except H5N1 showed a preference for A/C and to some extent G as well. H5N1

clearly favoured the usage of G/A at 3rd codon position of the 4-fold codons. The 2-fold codons predominantly preferred T/C in all the subtypes.

5.6 Trends in codon usage variation across the subtypes

To resolve the trend of codon usage variation among the IAV genes we employed correspondence analysis on RSCU values. The two central axes plotted in a two-dimensional scatter plot elucidated 55.7% of the total variations. The M1 and M2 genes from all the subtypes were seen forming two separate clusters leaving aside H5N1. The rest of the genes formed a separate cluster.

The deviation of H5N1 from the rest, however, is not limited to amino acid usage only. We observed preference of A/G at third position in case of H5N1 whereas the rest of the subtypes preferred A/T. There was difference in GC content as well for H5N1. The reason behind this could be due to the fact that contrasting to the other subtypes H5N1 is largely caused by zoonosis. The unusual difference could also be traced from reports of Korteweg and Gu (2008) which hinted that H5N1 broke the avian-human species barrier for the first time recently in 1997 (Korteweg and Gu, 2008; Chen *et al*., 2016). Being primarily a poultry disease its genetic make-up is more personalized to the avian hosts while the rest of the subtypes have been co-evolving with the human hosts for a longer span of time. Nonetheless, occasional human to human transmissions had also been reported (Ungchusak *et al*., 2005; Neumann *et al*., 2007; Chen *et al*., 2016).

5.7 Codon context analysis among the IAV subtypes

Codon pair context is a very important but often ignored feature of genetic framework in organisms that has been thought to modulate the precision of mRNA decoding (Moura *et al*., 2005). It represents the codon pairs occupying the A- and P-sites of the

ribosome during translation process. All codon context analyses were executed using the software Anaconda 2.0 (Moura *et al*., 2007). The association of codon-pairs was calculated using a chi-square test of independence. Based on the residual values for contingency table identified preferential and the rejected codon-pairs are displayed in a 64x64 color coded matrix plot. The plot gives an overall picture of the codon-context data. It has been envisaged that the codon-pair context is prone to the selection forces as that it plays crucial part in translation speed and accuracy of the mRNA decoding fidelity (Boycheva *et al*., 2003; Ogle and Ramakrishnan, 2005).

The analyses of codon context in our study did not offer much variation among the various subtypes. The plot suggested the presence of dissimilar preferences for codon contexts of different make-up in the genes. Contexts of NNC-GNN and NNT-ANN type were used at very low frequency. Amino acid pairs like Arg-Gly, Glu-Lys, Ser-Gly, Ser-Ser were preferred across the subtypes, however, with disparity in magnitudes. The comparison of subtypes one-to-one for codon context patterns presented largely similar patterns in all the cases.

5.8 Prediction of expression of the IAV genes

The relationship of codon usage with the gene expression level is well documented (Sharp and Li, 1986; Lobry and Gautier, 1994; Supek and Vlahovicek, 2005). Codon adaptation index and MELP were two indices used in this study to predict the gene expressivity level in the IAV subtypes. Both the parameters revealed high expressivity of most of the IAV genes with the exception of M2 gene. Both the indices however showed varying magnitude of predicted gene expressivity. Both CAI and MELP showed significant negative correlations with codon usage (Nc) thus establishing links of gene expressivity with codon usage.

5.9 Codon usage and basic biochemical properties of IAV proteins

It had been shown previously that the biochemical properties of amino acids composing the proteins have links with the underlying codon usage make-up of the genes encoding them (Lobry and Gautier, 1994; Zhong *et al*., 2007; Chen, 2013). We estimated Grand average of hydropathy index (Gravy), Hydrophilicity (Hydro), Aromaticity (Aroma) and Isoelectric point (pI) among the basic biochemical properties of the amino acids. We performed correlation analysis of these parameters with the codon usage indices. Significant inverse correlation was observed between pI and axis1 from correspondence analysis ($r = -0.427$, $p<0.01$) and statistically significant positive correlation between axis 1 and Gravy ($r = 0.457$, $p<0.01$). Gravy showed statistically significant inverse correlation with Hydro and pI.

5.10 Future directions of the present study

The influenza A virus is unique in its ability to cause recurrent epidemics and pandemics globally. Many studies have been carried out regarding codon usage bias in this virus till date, and it is unambiguous that the findings of these studies fell short in elucidating completely, the genotypic basis of the fitness of this virion.

Future studies should be focussed on the supplementary whole-genome analysis that is vital to comprehend fully the evolutionary machinery and epidemiological dynamics of this virus. The significance of predicting the emergence of novel influenza A virus strains for subsequent yearly vaccine development cannot be underrated. Finding the antigenically novel strains emerging as a result of reassortment among existing viral lineages is of major significance for candidate vaccine strain selection.

As the influenza virus uses the host cell machinery for establishing itself in the host, the role of protein export mechanisms becomes crucial. It has been experimentally shown that codon-usage bias plays a key role in the cellular export as discussed previously. Several questions still need to be addressed regarding the precise role of codon bias in the export of proteins. Further understanding of the role of codon-usage signals in protein folding and optimization of protein export would significantly help the production of recombinant proteins in various expression systems.

The role of codon optimization in vaccine development has been showed in many pathogenic viruses previously. In the present era of reverse vaccinology, synthetic DNA vaccine could be a possible answer to this notoriously divergent virus. Codon optimization would be crucial in this context.