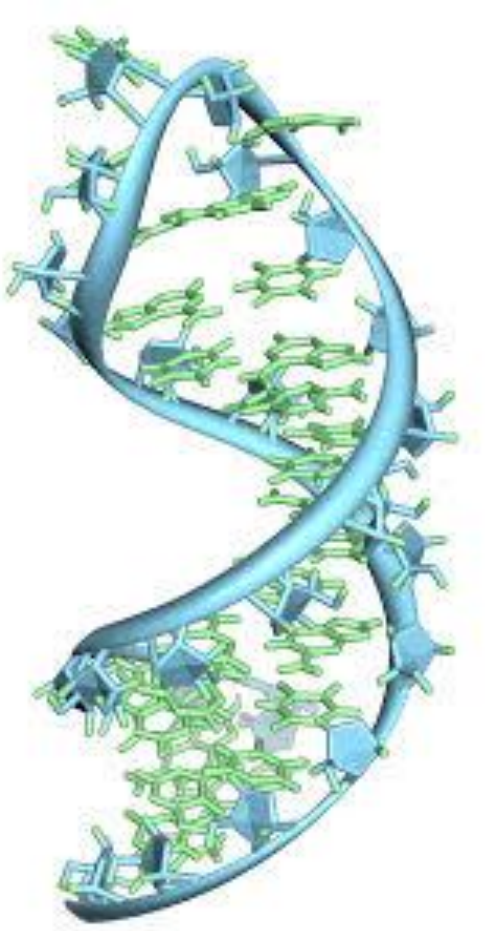
Chapter

3



MATERIALS AND METHODS

CHAPTER 3

MATERIALS AND METHODS

3.1 Sequence data retrieval

Complete coding sequences (cds) of eight different genes *viz.* hemagglutinin (HA), neuraminidase (NA), nucleoprotein (NP), matrix protein (M1 and M2), polymerase acidic protein (PA) polymerase basic proteins (PB1 and PB2), belonging to five IAV subtypes were used in the present study. All the sequences having exact multiple of 3 bases with correct start and stop codons were retrieved from GenBank database of NCBI (http://www.ncbi.nlm.nih.gov/).

3.2 Compositional analysis

Nucleobase composition is widely considered as an important feature in shaping up the codon usage profile in the genes. The critical role of nucleobase composition is apparent from the fact that majority of the indices for codon usage bias are largely rooted in the nucleobase composition of the genes. Amongst the compositional indices, GC pattern has been a very vastly significant force in the codon usage pattern. In this work, GC_3 represents the occurrence of the nucleotides G+C at the 3^{rd} codon positions of the codons exclusive of Met, Trp and the termination codons. Likewise, $GC_{1\ and}$ GC_2 stand for G+C frequency at 1^{st} and 2^{nd} codon positions respectively. GC_3 is considered as a useful indicator of the degree of base compositional bias.

3.3 Indices of codon usage study

3.3.1 Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) is an extensively used index for inspecting the synonymous codon usage pattern at both genes and genome scale.

RSCU is the proportion of the observed to the expected incidence of codons considering all the synonymous codons to be equally used (Sharp and Li, 1986). The synonymous codons are considered to be randomly and equally used when RSCU value is close to unity. The positively favoured codons have RSCU value over 1, whereas a value < 1 denotes a negative codon usage bias.

3.3.2 Effective number of codons (Nc)

The **effective number of codons (Nc)** is an index which reveals the degree of bias among the synonymous codons in a gene (Wright, 1990). Nc is calculated to acquire information on the synonymous codon usage in the target sequence. Mathematically Nc is represented as:

$$Nc = 2 + \frac{9}{\mathrm{F}_2} + \frac{1}{\mathrm{F}_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

In the above equation, the F value means the probability that the arbitrarily selected synonymous codons for a particular amino acid are identical. F_k (k = 2, 3, 4 or 6) is the mean of the F values for the amino acids with k-fold degeneracy level. The extreme values of Nc are 20-61. The lower lowest Nc value 20 is for the instance when only solitary codon is used per amino acid; whereas Nc value of 61 means that all the synonymous codons are uniformly used for each amino acid (Wright, 1990; Comeron and Aguade, 1998; Novembre, 2002). The codon usage bias is usually considered low when the Nc value exceeds 40.

3.3.3 Dinucleotide odds ratio

Dinucleotide odds ratio was calculated as the proportion of observed frequency of a dinucleotide pair to the frequencies of the constituting individual nucleotides of that

particular dinucleotide pair (Karlin and Burge, 1995). The calculation of the odds ratio was as per the following equation:

$$\rho xy = \frac{fxy}{fxfy}$$

Here, *fx* and *fy* stand for the frequencies of mononucleotides x and y respectively, whereas *fxy* represents the incidence of the dinucleotide composed by x and y. The range 0.8-1.25 is taken as the periphery values of odds ratios. Any dinucleotide with an odds ratio value less than 0.8 is considered under-represented. On the other hand, the odds ratio value over 1.25 is taken as over-representation of the dinucleotide in the coding sequence.

3.3.4 Codon adaptation index (CAI)

Codon adaptation index (CAI) is a widely used index to figure out the adaptiveness of the synonymous codons in a gene towards the codon usage of highly expressed genes. CAI is also used as a predictive measure of gene expressivity. This parameter was given by Sharp and Li (1980) while working on the codon usage bias in *E. coli* (Sharp and Li, 1987). CAI was initially proposed to offer a normalized estimate which can be used both at the gene and species level. The range of CAI values varies from 0 to 1. A CAI value of 1 is consigned to the most recurrent codon within a gene whilst the least occurred codon is assigned a value of *0.* CAI is calculated as:

$$CAI = exp\frac{1}{L}\sum_{k=1}^{L}\ln w_{c(k)}$$

Here, L is the number of codons in the gene and $w_{c(k)}$ is the relative adaptiveness (ω) of the k-th codon in the gene.

3.3.5 MELP (MILC based expression measure)

Measure independent of length and composition (MILC) is a predictive measure of gene expressivity that is not subjective to the length of gene and nucleobase composition (Supek and Vlahovicek, 2005). MILC is estimated as:

$$MILC = \frac{1}{L}\sum_{a \in A} M_a - K$$

where, L stands for the number of codons in the cds, M_a represents the goodness of fit statistical test for observed to expected usage of codons, while *K* is a correction factor. The maximum MILC value from the reference set is used for working out MELP**,** the gene expression measure, which is given by:

$$MELP = \frac{MILC^{(gene)}}{MILC^{(ref)}}$$

Here, $MILC^{(ref)}$ is the maximum MILC value of the reference set, and $MILC^{(gene)}$ is the MILC value of the concerned gene.

3.3.6 Frequency of optimal codons (Fop)

Frequency of optimal codons (Fop) is another index of codon usage bias in a gene given by Ikemura in the year 1985 as an estimate to stand for the ratio of the number of optimally used codons to the total number of synonymous codons. Fop value of a gene *g* was calculated as per Lavner and Kotler (2005) as follows:

$$Fop(g) = \frac{1}{N}\sum syn(i)n_i(g)$$

Here, $n_i(g)$ is the count of the i^{th} codon in the gene *g*, N is the total number of codons in *g* and *syn(i)* is the degeneracy level of the amino acid encoded by the i^{th} codon (Lavner and Kotlar, 2005). The Fop values calculated using the above formula is independent of amino acid composition. Thus, a gene close to random synonymous

usage of codon will have Fop value close to 1 regardless of its amino acid composition (Lavner and Kotlar, 2005).

3.3.7 Neutrality analysis

An analytical method for weighing up codon bias is the neutrality analysis. In this analysis, the average GC contents at the first two codon positions, represented by GC12, are plotted along the ordinate and GC3 values are put along the abscissa in a scatter plot. In this plot, a statistically significant correlation between GC12 and GC3, along with a regression line having slope close to 1 indicates that mutation bias could be the vital force inflicting codon usage. In contrast, selection against mutation bias may bring about a constricted allotment of GC content which is impersonated in a lack of correlation between GC12 and GC3.

3.3.8 PR2 analysis

The parity rule 2 (PR2) plot is a means to assess the intra-strand bias. The plot is useful to examine the influence of mutational pressure and selective constraint on the usage of codon in a gene (Sueoka, 1995). In this plot AT-bias *i.e.* [A3/(A3 +T3)] as the ordinate is plotted against GC-bias *i.e.* [G3/(G3+C3)] as the abscissa. Parity Rule 2 (PR2) is a logically obtained rule from Watson-Crick Model of DNA organization. According to PR2, in absence of mutation/selection bias, intra-strand composition should follow the rule A = T and G = C. Thus, the centre of the plot at which both the coordinates are equal to 0.5, is the place where A becomes equal to T and G equals C. Deviation from the rule (PR2-biases) hints asymmetric mutation/selection pressures amid the two DNA strands (Sueoka, 2002).

3.3.9 Codon pair context analysis

Codon-pair context represents the codon pairs harboured by the ribosomal A- and P-sites. All codon context analyses were executed using the Anaconda program version 2.0 (Moura *et al*., 2007). The alliance of codon-pairs is estimated using chi-square test of independence. Based on the adjusted residual values for contingency table the preferential and the rejected codon-pairs are recognized and displayed in a 64x64 color coded map. The map imparts an overall view of the codon-pair context data.