CHAPTER – 5

DISCUSSION

5. Discussion

Codon usage bias (CUB) is a unique property in all domains of life which reflects an optimized combination of frequent and rare codons in order to perform appropriate translation of a certain gene, in an organism under certain conditions, and the gene is finally expressed in some tissues or organelles. During the past three decades the codon bias study has been performed in a variety of organisms ranging from lower prokaryotes to higher eukaryotes and some of the insights gained from such studies have been widely applied in biotechnology as a strategy to optimize gene expression for improving protein production rates and yields (Quax *et al*. 2015). Moreover, the codon bias strategy also supports the cell-cycle regulation since variations in tRNA expression during different states of a cell may enable differential expression of sets of genes which are adapted to different tRNA gene pools. As a result, the genes exhibit different codon usage at different phases of the cell-cycle (Frenkel-Morgenstern *et al*. 2012). The selective forces on codon bias during the period of evolution have been under the balance between mutation and natural selection (Roth *et al*. 2012).

The current study was carried out to analyze the *CAI*, *tAI*, *ENC*, *Fop*, *RSCU*, base composition for the selected human proto-oncogenes/oncogenes and tumour suppressor genes. We also found out the level at which the above mentioned genetic factors are involved in the formation of codon usage pattern. As per our mentioned objectives in this present study, we selected eighty-two proto-oncogenes/oncogenes and sixty-three tumour suppressor genes (Table 1 and 2) from *Homo sapiens* for CUB analysis. The accurate coding sequences (cds) having correct initial and termination codons were retrieved using a program in perl, developed by us.

After analyzing the coding sequences, it was found that nearly 54% of the cds selected for proto-oncogenes/oncogenes and 51% for tumour suppressor genes are rich in GC contents. The overall nucleotide composition analysis in the complete coding sequences of selected human proto-oncogene/oncogenes (Table 3) showed that the mean value of A (520.8) was the highest, followed by C (517.6), G (502.2) and T (416.6). Similarly, for the selected coding sequences of tumour suppressor genes (Table 4) the mean value of A (767.2) was the highest, followed by G (655.2), C (654.4) and T (613.2).

We also observed that the GC content at different codon positions (GC_1, GC_2, GC_3) varies among both proto-oncogene/oncogene and tumour suppressor genes (Figure 4 and 5). From the nucleotide composition analysis, it can be presumed that G/C –ending codons might be preferred in the coding sequences of proto-oncogene/oncogene and almost equal distribution of G/C and A/T –ending codons in the coding sequences of tumour suppressor genes.

The base composition variation within each coding sequence of the selected proto-oncogenes/oncogenes and tumour suppressor genes were calculated from their differences in usage for the GC, AT, keto, amino, purine, and pyrimidine bases (Table 5 and 6). It is well known that, skew values represent the base composition bias and are linked to the transcription process (Fujimori *et al*. 2005, Touchon and Rocha 2008). The skew values of our analysis showed that forty-three proto-oncogene/oncogene out of eighty-two and twenty-nine tumour suppressor genes out of sixty-three had negative GC skew values (Figure 6 and 7), indicating richness of C over G (Tillier and Collins 2000). However, in case of AT skew as well as AT3 skew,

nearly all the selected proto-oncogenes/oncogenes and tumour suppressor genes showed positive value which indicates richness of A over T (Tillier and Collins 2000). Zhang & Gerstein (2003) reported that negative keto skew was observed in human genome (Zhang and Gerstein 2003). Our analysis also revealed that the keto skew values were negative for forty-three proto-oncogene/oncogene and twenty-three tumour suppressor genes. In addition, wide variations in purine, pyrimidine and amino skew were observed in the coding sequences of selected proto-oncogenes/oncogenes as well as tumour suppressor genes, which might affect the gene expression patterns.

In order to find out the relationship between the codon usage variation and GC constraints among the selected coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes, we analyzed the correlation coefficient between codon usage and GC bias using heat map (Figure 8 and 9). In our analysis, we observed nearly all codons ending with G/C base showed positive correlation ($p<0.01$) with GC_{3s} indicating that codon usage had been influenced by GC bias and all A/T–ending codons showed negative correlation with GC bias. However, three G-ending codons AGG (arginine), TTG (leucine), CAG (glutamine) for proto-oncogenes/oncogenes and the codons AAG (lysine), AGG and TTG for tumour suppressor genes showed negative correlation with GC_{3s} and displayed non-linear upward usage profiles as a function of GC bias (Figure 10 and 11). This suggested that the usage frequency of these codons will decrease with the increase of GC bias as indicated by GC_{3s} (Palidwor *et al*. 2010).

We again performed, correlation analysis between the values of A, T, G, C and GC with A_3, T_3, G_3, C_3 and GC_3 values, respectively in order to find out the relationship

between codon usage variation and compositional constraints (Table 7 and 8) among the coding sequences of proto-oncogene/oncogene and tumour suppressor gene. Interestingly, we observed significant positive correlation ($p<0.01$) in the nucleotide composition, suggesting that nucleotide constraint may influence the codon usage pattern in both the cases of proto-oncogene/oncogenes and tumour suppressor genes. However, many investigators claimed that codon bias is related mainly to the non-random mutations caused by the global GC content of an organism, because the GC content seems determined for the complete genome, not only for the coding part of the genome (Chen *et al*. 2004, Knight *et al*. 2001).

Furthermore, a neutrality plot of GC_{12} versus GC_3 (Figure 12 and 13) (Sueoka 1988) was constructed. In neutrality plots, when there exists a significant correlation between GC_{12} and GC_3 and the slope of the regression line is close to 1, it indicates that mutation bias is the main force in shaping the codon usage. Conversely, a lack of correlation between GC_{12} and GC_3 indicates selection against mutation bias which results in a narrow distribution of GC content (Sueoka 1988). In our neutrality plot analysis we compared the values of GC_{12} and GC_{3s}, and observed a significant positive correlation ($p<0.01$) in the coding sequences of proto-oncogenes/oncogenes as well as tumour suppressor genes, which suggested that intragenomic GC mutation bias plays an important role in shaping the codon usage pattern of these genes (Dass and Sudandiradoss 2012, Liu H. *et al*. 2012, Nair *et al*. 2013). The neutrality plot (as shown in figure 12 and 13) reveals the results of equilibrium coefficient of mutation and selection. In the plot (Figure 12 and 13) some of the points are located in the diagonal distribution and the values of GC_3 are in a wide range of distribution, indicating GC_{12} and GC_3 are definitely the affects of mutation

bias model (Sueoka 1988). Accordingly, mutation bias might just play a minor role in shaping the codon bias, whereas the natural selection seems to probably dominate the codon bias. Further, we quantify the natural selection and the mutation pressure using regression coefficient. The regression coefficient of GC_{12} on GC_3 in human proto-oncogene/oncogene is 0.176, indicating the relative neutrality is 17.6 %, while the relative constraint is 82.4 % for GC_3. The linear regression coefficient of GC_{12} on GC_{3s} in human tumour suppressor gene is 0.259 indicating the relative neutrality is 25.9 %, while the relative constraint is 74.1 % for GC_3. These results suggest that natural selection played a major role while mutation pressure played a minor role in the codon usage patterns of proto-oncogenes/oncogenes as well as tumour suppressor genes in human.

We also predicted the heterogeneity of codon usage by analyzing the effective number of codons (*ENC*) and its relationship with GC_{3s} values in the coding sequences of proto-oncogene/oncogenes and tumor suppressor genes, and found a significant negative correlation ($p<0.01$) between *ENC* and GC_{3s} for these genes.

In addition, we plotted the values of *ENC* versus GC_{3s} (Figure 14 and 15) as per Wright (1990) which revealed that GC_{3s} value was a major determinant factor and that other trends independent of compositional constraints might influence the overall codon usage variation in human proto-oncogenes/oncogenes and tumour suppressor genes. We further analyzed the frequently used optimal codon (*Fop*) values for the proto-oncogenes/oncogenes and tumour suppressor genes in order to determine the occurrence of the highest and the lowest frequently used codons for the corresponding amino acid. The average percentage of the *Fop* value of codons showed that the most frequently used codons for proto-oncogenes/oncogenes were

GCC, AGA, AAC, GAC, TGC, CAG, GAG, GGC, CAC, ATC, CTG, AAG, TTC, CCC, AGC, ACC, TAC and GTG for the amino acid alanine, arginine, asparagine, aspartate, cysteine, glutamine, glutamate, glycine, histidine, isoleucine, leucine, lysine, phenylalanine, proline, serine, threonine, tyrosine and valine respectively (Table 9). The results of the Fop values in case of tumour suppressor genes were similar to that of proto-oncogenes/oncogenes, except the codon TTT, CCT and ACA encoding amino acid phenylalanine, proline and threonine respectively (Table 10).

We analyzed the relative synonymous codon usage values of 59 codons in the coding sequences of each proto-oncogene/oncogene and tumour suppressor gene excluding the codons ATG, TGG that encode for amino acids methionine and tryptophan respectively. Our analysis from the overall *RSCU* values in the selected coding sequences of proto-oncogenes/oncogenes revealed that 27 codons were most frequently used among the set of 59 codons and the most predominantly used codons were G/C–ending compared to A/T–ending types (Table 11).

However, the overall *RSCU* values in the selected coding sequences of tumour suppressor genes revealed that 29 codons were the most frequently used among the 59 codons and the most predominantly used codons were C/T–ending compared to A/G–ending (Table 12). Further, it is also revealed that the codons ending with C base was mostly favored in comparison to other bases and the codon CTG encoding leucine amino acid was mostly over-represented (highest *RSCU* value) among human proto-oncogenes/oncogenes and tumour suppressor genes. The most preferred codon ending with base C and the most over-represented codon CTG encoding leucine amino acid were also reported earlier in human serotonin receptor

and mammalian rhodopsin gene families including human (Dass and Sudandiradoss 2012, Du *et al*. 2014).

Previously, it was reported that dinucleotide bias can affect the overall codon usage patterns in a variety of organisms (Chiusano *et al*. 2000, Karlin and Burge 1995). The relative abundance of 16 dinucleotides in the coding sequences of the selected proto-oncogenes/oncogenes and tumour suppressor genes was calculated in order to assess the effect of dinucleotides on the codon usage patterns of these genes (Figure 16 and 18). In our analysis, it was evident that CpG dinucleotides were under-represented whereas GpC dinucleotides were over-represented in both the cases of proto-oncogenes/oncogenes and tumour suppressor genes. However, nearly all the codons containing dinucleotide CpA and TpG were over-represented and most of them were also used as preferred codons for their corresponding amino acid based on *RSCU* analysis in the selected genes under study. Interestingly, it was reported earlier that CpG dinucleotide might play an effective role for the over-representation of CpA and TpG dinucleotides in different organisms due to spontaneous deamination of methylated cytosine residues that are prone to mutate into thymine residues, resulting in the dinucleotide TpG and CpA on the opposite strands of DNA after replication (Bird 1980). In our analysis since most of the codons containing CpA and TpG dinucleotides were over-represented, this theory was partly applicable in case of both proto-oncogenes/oncogenes and tumour suppressor genes of human under study.

Further, correspondence analysis (COA) based on *RSCU* values was performed in order to investigate the major trends in codon usage variation among human proto-oncogene/oncogenes and tumour suppressor genes which showed that the principal

axis (f_1) accounted for 46.68% in case of proto-oncogenes/oncogenes and 56.10% in tumour suppressor genes of all variations within their gene sets (Figure 17 and 19). Significant correlation between the principal axis and each of the four major indices namely *ENC*, GC, GC3s and *CAI*, revealed that nucleotide composition and mutation bias might play a pivotal role in shaping the codon usage patterns of human proto-oncogenes/oncogenes and tumour suppressor genes. The result was similar to the findings of Dass and Sudandiradoss (2012) on human serotonin receptor gene family (Dass and Sudandiradoss 2012).

To predict the level of gene expression on the basis of extent of bias towards codon sequence, we use the codon adaptation index (*CAI*) parameter as reported earlier in lower prokaryotes to higher eukaryotes (Behura and Severson 2012, Gupta *et al*. 2004, Liu Fei *et al*. 2011, Pavlicek *et al*. 2004). We calculated the *CAI* values for proto-oncogene/oncogene and tumour suppressor gene and observed that most of the coding sequences selected from *Homo sapiens* qualify as highly expressed genes (Figure 20 and 24). We analyzed normalized AT and GC frequency at each codon site. Significant correlation was observed between gene expression as measured by *CAI* and GC content at any codon site. Among all GC_{3s} showed the highest correlation with *CAI* i.e. 0.847 and 0.839 with proto-oncogene/oncogene and tumour suppressor gene respectively (Figure 21 and 25). Moreover, in order to investigate the relationship between the codon usage variation and gene expression as measured by *CAI* among the selected coding sequences of human proto-oncogene/oncogenes and tumour suppressor genes, a heat map was generated to represent the correlation coefficient between codon usage and *CAI* values (Figure 22 and 26) which showed that nearly all G/C–ending codons are positively correlated

with *CAI* and vice versa for A/T–ending codons, indicating that gene expression increases with the increase in usage of G/C–ending codons.

Moreover, we used *tRNA* adaptation index (*tAI*) as a parameter to investigate whether translational selection plays any role in shaping the codon usage in human proto-oncogene/oncogene as well as tumour suppressor gene pool. We estimated the *tAI* value of each cds as a measure of adaptation to genomic *tRNA* pool. We carried out a correlation analysis between codon usage and *tAI* which was represented in a heat map (Figure 23A and Figure 27A). Our results showed that nearly all the codons ending with G/C base were associated with positive correlation between codon usage and *tAI* and vice versa for the A/T ending codons both for proto-oncogene/oncogene and tumour suppressor gene pool.

Then we performed a correlation analysis between the *tAI* and the effective number of codons (*ENC*) for each cds in order to investigate the contribution of translational selection to the overall codon usage among the proto-oncogene/oncogene and tumour suppressor gene pool. Since the *ENC* value is related to the amount of entropy in codon usage, the lower value of *ENC* indicates the selection for optimal translation in the coding sequences if translational selection is the driving force in shaping the codon usage among proto-oncogenes/oncogenes and tumour suppressor genes. We observed a significant negative correlation ($r=-0.545$, $p<0.01$) between *ENC* and *tAI* (Figure 23B) in case of proto-oncogene/oncogene pool but significant positive correlation ($r=0.328$, $p<0.01$) in case of tumour suppressor gene pool (Figure 27B). Moreover, the codon usage along with its *ENC* is largely influenced by the percentage of GC contents at the third nucleotide position of the codons (dos Reis *et al*. 2004) and a significant positive correlation (r=0.382 and

r=0.511, p<0.01) between *tAI* and GC3s (Figure 23C and 27C) was observed in proto-oncogene/oncogene and tumour suppressor gene pool respectively. Our results thus suggest the co-adaptation between codon usage and *tRNA* gene copy numbers among the proto-oncogenes/oncogenes and tumour suppressor genes. Besides, the expression of proto-oncogenes/oncogenes and tumour suppressor genes (*CAI*) was remarkably influenced by the genomic *tRNA* pool as shown by the significant positive correlation between *tAI* and *CAI*. Previously it was reported that translational selection may not operate in all genomes (dos Reis *et al*. 2004). Our results suggest that even in those cases where translational selection is likely to be present as indicated by the effect of GC3s and *ENC;* the contribution of translational selection may be insignificant to the overall codon bias. This might be due to the effect of mutational pressure outriding the translational selection effect on the overall codon bias in the coding sequences.

In brief, we analyzed the codon usage pattern and the key genetic factors playing a decisive role in determining the pattern of codon usage for the eighty-two proto-oncogenes/oncogenes and sixty-three tumour suppressor genes. Based on the hypothesis that gene expressivity and codon composition are strongly correlated, the codon adaptation index has been defined to provide an intuitively meaningful measure of the extent of the codon preference in a gene. In addition, we observed that nature highly preferred the C–ending codons in the coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes and the codon CTG encoding leucine amino acid was the over-represented codon (highest *RSCU* value).

The most preferred codon ending with base C and the most over-represented codon CTG encoding leucine amino acid were also reported earlier in human serotonin receptor and mammalian rhodopsin gene families (Dass and Sudandiradoss 2012, Du *et al*. 2014). According to Yang & Nielsen (2008), codon bias in mammals is mainly influenced by mutation bias and that the selection on codon bias is weak for nearly neutral synonymous mutations (Yang and Nielsen 2008). Our results also revealed that the codon usage of human proto-oncogene/oncogene and tumour suppressor gene was primarily affected by GC mutation bias since its effects were present at all codon positions and that relatively weak codon bias was observed in these genes.

Significant correlation exists between the compositional constraints influencing the codon bias and gene expression level as measured by *CAI*, in human proto-oncogenes/oncogenes as well as tumour suppressor genes, indicating that gene expression level might play a pivotal role in shaping their codon usage patterns. Further, significant positive correlation was observed between *tAI* and *CAI*, suggesting that the expression of proto-oncogenes/oncogenes and tumour suppressor genes (*CAI*) was remarkably influenced by the genomic tRNA pool. Our present findings certainly report a novel insight into the codon usage patterns in gaining the clues for the functional conservation of gene expression, translational studies and codon optimization for desired expression and the significance of the nucleotide composition in the coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes. Since, our analysis has given better insights into the codon usage; it may be useful in further understanding the molecular evolutionary relationship between the coding sequences of these genes.

Summary

The degeneracy of the genetic code and unequal usage of synonymous codons for encoding the corresponding amino acid during the translation of a gene into protein are components of a widely accepted phenomenon commonly known as codon usage bias (CUB). Literature suggests that codon bias is directly linked to the level of gene expression and several genomic factors such as mutation pressure, natural or translational selection, secondary protein structure, translational efficiency and fidelity, hydrophobicity and hydrophilicity of the protein and the external environment also play a major role in shaping the codon usage patterns of different organisms or within the genes of the same organisms (Butt *et al*. 2014). The study of codon usage bias acquires significance in molecular biology for understanding the codon usage patterns of an organism/gene during the period of evolution as well as new gene discovery, design of transgenes for increased expression, detecting lateral gene transfer based on nucleotide compositional dynamics and for analyzing the functional conservation of gene expression (Carbone *et al*. 2005, Lithwick and Margalit 2005).

In human a group of genes known as proto-oncogenes usually help in normal cellular growth and development. When they mutate (change) or there are too many copies of each gene, they become “bad” genes and permanently turned on or activated resulting in uncontrolled proliferation of cell growth which can lead to cancer (Adamson 1987). As a result of mutation when the expression level of proto-oncogenes increases they turn into oncogenes which exhibit increased production of these proteins resulting in increased cell division, decreased cell differentiation and inhibition of cell death. Thus, conversion or activation of a proto-oncogene into an

oncogene involves the *gain-of-function* mutation that has the potential to induce cancer (Blanchard 2002, Todd and Wong 1999). On the other hand, tumour suppressor gene, the “care taker of the genome”, plays an important role in the regulation of cell proliferation, differentiation by involving cell cycle control, signal transduction, angiogenesis and development of normal as well as tumour related functions (Marshall 1991). Inactivation or mutation of a tumour suppressor gene leads to a negative regulation of cell proliferation and contribute to tumour development in combination with other genetic changes (Vousden and Lu 2002).

In this study, we have analyzed the nucleotide compositional constraints and estimated the expression level of the coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes in the cell using several genetic indices namely, the codon adaptation index (*CAI*), *tRNA* adaptation index (*tAI*), frequency of optimal codon (*Fop*), relative synonymous codon usage (*RSCU*), effective number of codons (*ENC*) and compositional dynamics of the background nucleotide constraints.

In our analysis, we selected only those coding sequences which have perfect start and stop codons and devoid of any unknown bases (N) and any intercalary stop codon in the entire sequence. Finally, 82 cds sequences of human proto-oncogenes/oncogenes and 63 cds sequences of tumour suppressor gene pool that fulfill the aforementioned criteria were retrieved from the GenBank database of National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) and used for CUB analysis.

The major objective of this study is to understand the codon usage patterns and to estimate the strength of selection on the expression of human proto-oncogene/oncogenes and tumour suppressor genes in cells.

Major Findings:

- ✓ Our results showed that codon usage in human proto-oncogene/oncogene and tumour suppressor genes has been influenced by GC bias, mainly due to GC_{3s}.
- ✓ The majority of the frequently used codons were G/C ending in which C–ending codons were mostly favored compared to G–ending codons for the corresponding amino acid.
- ✓ The present study also revealed that intragenomic GC mutation bias influences the codon usage patterns in the coding sequence of human proto-oncogene/oncogene as well as tumour suppressor gene, since its effects are present at all codon positions. Relatively weak codon bias was observed in these genes.
- ✓ The linear regression coefficient of GC_{12} on GC_{3s} of human proto-oncogenes/oncogenes and tumour suppressor genes suggested that natural selection played a major role while mutation pressure played a minor role in the codon usage patterns of proto-oncogenes/oncogenes as well as tumour suppressor genes in human.
- ✓ Further, correspondence analysis (COA) based on *RSCU* values showed that the principal axis was significantly correlated with the four major indices of codon bias namely *ENC*, GC, GC3s and *CAI*, which revealed that nucleotide composition and mutation bias might play a pivotal role in shaping the codon usage patterns of human proto-oncogenes/oncogenes and tumour suppressor genes.
- ✓ In addition, we observed that nature might have preferred the over-representation (highest *RSCU* value) of the codon CTG encoding leucine amino acid in the coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes.

- ✓ The frequently used optimal codon (*Fop*) values in the coding sequences of proto-oncogenes/oncogenes were similar to that of tumour suppressor genes, except the codons TTT, CCT and ACA encoding amino acids phenylalanine, proline and threonine respectively.
- ✓ Moreover, significant correlation was observed between compositional constraints (codon usage) and gene expression level (measured by *CAI*) suggesting that expression level might play a pivotal role in the codon usage of human proto-oncogenes/oncogenes and tumour suppressor genes.
- ✓ In addition, significant positive correlation was observed between *tRNA* adaptation index (*tAI*) and codon adaptation index (*CAI*) suggesting that the expression of proto-oncogenes/oncogenes and tumour suppressor genes (*CAI*) was remarkably influenced by the genomic *tRNA* pool.

To the best of our knowledge, this is the first report on the codon usage pattern in the coding sequences of human proto-oncogene/oncogene and tumour suppressor genes. Our present findings certainly report a novel insight into the codon usage patterns in gaining the clues for the functional conservation of gene expression, translational studies and codon optimization for desired expression and the significance of the nucleotide composition in the coding sequences of human proto-oncogene/oncogenes and tumour suppressor genes.