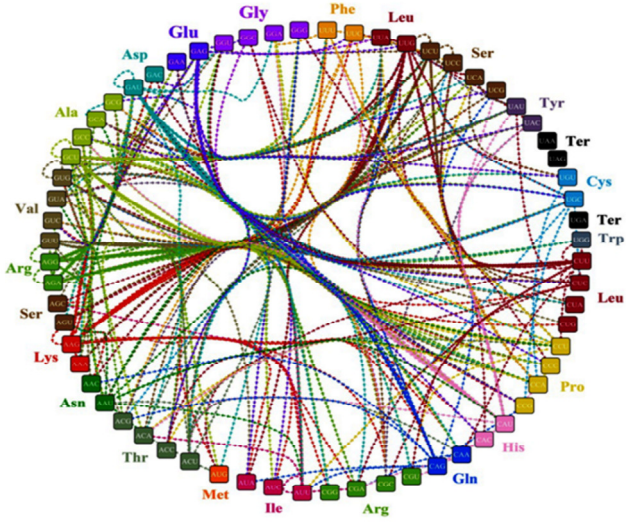
CHAPTER – 2

REVIEW OF LITERATURE



2. Review of Literature

Codon degeneracy phenomenon which means more than one codon encodes for the same amino acid was discovered in 1965 (Nirenberg *et al*. 1965). The alternative codons that encode the same amino acid are represented as synonymous codons. Researchers discovered that synonymous codons are not used at equal frequencies for encoding the same amino acid in the protein-coding DNA sequence. The first reports of unequal usage of synonymous codon can be traced back to as early as four decades ago. Clarke (1970) and later Ikemura (1981), and Akashi (1994), suggested that an organism's codon usage is influenced by its *tRNA* pool (Akashi 1994, Clarke 1970, Ikemura 1981). Observed differences in codon bias between species are a result of different evolutionary forces acting on the choice of codons (Ikemura 1981). Codon usage can differ widely not only between organisms, but also within an organism.

The codon usage patterns have been analyzed since the outstanding efforts of the first molecular sequence databases (Grantham *et al*. 1981). The result of Grantham and his co-workers demonstrated that species specific genes share similar patterns of synonymous codon usage frequency as stated by the "genome hypothesis" (Grantham *et al*. 1980, Grantham *et al*. 1981). Therefore, scanning the codon usage patterns of all the genes in an organism may obscure the underlying heterogeneity (Aota *et al*. 1988) and hence it is better to identify the trends of codon usage patterns within the genes of a species or between closely related species. Codon bias may result from mutational biases alone. Many mutations originate in genome from non-random mismatch repair following replication errors and methylation.

Such asymmetric mutation rates of the leading and lagging strands are found in both bacteria and eukaryotes (Fijalkowska *et al*. 1998, Kunkel *et al*. 2003, Lobry 1996, McLean *et al*. 1998, Pavlov *et al*. 2003).

At the RNA level, selection for the effective transcription has been proposed by Xia (1996), in which mRNA with more abundant nucleotides are transcribed more quickly (Xia 1996). In these cases, the codons are enriched in common nucleotides. Selection can also take place at the mRNA level, where some patterns are avoided or preferred, to influence the mRNA folding and decay.

It has also been found that codon bias also correlates well with mRNA levels. This is an indication of the fact that there is a global optimization of minimizing the time where the ribosomes are engaged in translation of the mRNA. Codons evolving under positive selection have corresponding *tRNA*s in larger quantities and possibly bind to the mRNA at the ribosome more rapidly (Ran and Higgs 2010).

Organisms show a tendency to select codons which will facilitate maximal translational efficiency. Selection for optimal translation is more effective in organisms with large effective population sizes. Indeed, strong codon bias was reported by Bulmer (1987), in the genomes of *E. coli* and yeast, which have large population sizes (Bulmer 1987). Codon bias in different gene regions appears to be under different selective constraints, due to the early phase of translation (Karlin *et al*. 1998). The correlation between several different indices and experimental data, such as mRNA expression levels or protein concentration, has been examined in many studies since the prediction of gene expression levels is the aim of many

researchers (Coghlan and Wolfe 2000, Comeron and Aguade 1998, Supek and Vlahovicek 2005, Suzuki *et al*. 2008, Tuller *et al*. 2007).

Reports of codon usage bias among different organisms or within the genes of the same organisms have been published across the globe. Some of which are described below.

In 2011, Botzman and Margalit conducted a study on global variation in codon usage bias in prokaryotes in association to their life styles and reported that codon bias is more in the highly expressed genes. It has a global effect on cell fitness and also revealed that facultative organisms, mesophiles and pathogenic bacteria have increased codon usage bias which helps them adjust to changing environment (Botzman and Margalit 2011).

A vast majority of codon usage studies has been conducted on viruses (Wong *et al*. 2010), bacteria (Bailly-Bechet *et al*. 2006), yeast (Freire-Picos *et al*. 1994), *Caenorhabditis elegans* (Stenico *et al*. 1994) and *Arabidopsis thaliana* (Chiapello *et al*. 1998). But, only a few studies have been done on the vertebrate species such as mammals (Eyre-Walker 1991, Sharp et al. 1995, Sueoka and Kawanishi 2000) and rodents (Smith and Hurst 1999).

Wright (1990) evaluated a genetic parameter, popularly known as effective number of codon (*ENC*) which provides useful estimation of the absolute codon usage bias for a certain gene and is independent of gene length as well as amino acid compositions. Further, codon usage patterns across genes can be examined by plotting the values of *ENC* versus G+C contents at synonymous sites and the plots were constructed for *Homo sapiens*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Bacillus subtilis*, *Dictyostelium discoideum*, and *Drosophila melanogaster* (Wright 1990).

Eyre-Walker (1991) developed a new statistical tool and applied to 62 mammalian sequences and detected significant codon usage biases in human, rats and mice. They reported that the codon biases in all the three species appeared due to the first base pair in the codon which influences the mutation pattern at the third position of the codon (Eyre-Walker 1991).

Sharp and Li (1986) proposed relative synonymous codon usage (*RSCU*) as a parameter which is useful to study the overall synonymous codon usage variation among genes (Sharp and Li 1986).

In 1996, Karlin and Mrazek anticipated that major factors on codon usage in human genes result from residue preferences and diresidue associations in proteins coupled to biases on the DNA level, related to replication and repair processes and/or DNA structural requirements (Karlin and Mrazek 1996).

Sharp and Li (1987) evaluated an useful quantitative measure, codon adaptation index (*CAI*) to identify the extent of codon usage bias towards the codons that are known to be favored in highly expressed genes (Sharp and Li 1987).

Sueoka (1988) reported an analytical method to measure the codon usage patterns where the GC contents at different codon positions were analyzed. The average of GC contents at the first and second codon positions (GC_{12}) and GC_3 were used for neutrality plot analysis. 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 reveals that mutation bias is the main evolutionary 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).

Xie and Ding (1998) conducted a study based on the analysis of synonymous codon usage pattern in association with protein three-dimensional structure and reported that there is a correlation between synonymous codon usage and protein secondary structure in mammals (Xie and Ding 1998).

Similarly, McWeeney and Valdes (1999) performed a study on codon usage bias and base composition in MHC genes in humans and common chimpanzees. Their results revealed that codon usage patterns provide us with a better understanding of both the evolutionary history of these genes and the evolution of synonymous codon usage in genes under natural selection (McWeeney and Valdes 1999).

Sueoka and Kawanishi (2000) carried out a study based on the analysis of 14,026 human genes to investigate the codon usage bias in association with GC content at the third codon position. They revealed that the directional mutation pressure, rather than the directional selection pressure, is mainly responsible for the heterogeneity of the GC content of the third codon position (Sueoka and Kawanishi 2000).

Singer and Hickey (2000) performed an analysis on the nucleotide contents of several completely sequenced genomes and reported that background nucleotide GC contents affect the amino acid composition of proteins (Singer and Hickey 2000).

Chiusano (2000) reported that second codon positions of genes and dinucleotide bias can influence the overall codon usage patterns in a variety of organisms (Chiusano *et al*. 2000).

Knight *et al*., (2001) evaluated a simple model based on mutation and selection which explains the patterns of codon usage and amino acid usage along with the GC composition within and across the genomes of bacteria, archaea and eukaryotes. They showed that the codon and the amino acid usage are consistent with

evolutionary forces acting on nucleotides, rather than on codons or amino acids, although both mutation and selection play vital roles (Knight *et al*. 2001).

Zeeberg (2002) reported that synonymous codon choice is utilized in human for control of protein folding rather than for regulation of translation. Exonic GC of human mRNA sequences as well as A, C, G, and T in codon position 3 were linearly correlated with genomic GC. A Shannon Information Theoretic measure of bias for synonymous codon usage was developed. When applied to human or mouse sequences, this measure was nonlinearly correlated with genomic, exonic, and third codon position A, C, G, and T. Information values between orthologous mouse and human sequences were linearly correlated: mouse = 0.092 + 0.55 human. Mouse genes were consistently placed in genomic regions whose GC content was closer to 50% than was the GC content of the human ortholog. Since the (nonlinear) information versus percent GC curve had a minimum at 50% GC and monotonically increased with increasing distance from 50% GC, this phenomenon directly resulted in the low slope of 0.55. This appeared to be a manifestation of an evolutionary strategy for placement of genes in regions of the genome with a GC content that related synonymous codon bias and protein folding (Zeeberg 2002).

Louie *et al*., (2003) performed an analysis on the nucleotide composition across an averaged representation of all known human genes and observed that the frequency of nucleotide usage varies across different exons and introns of human genes. They reported that such variation may arise from differential mutational pressures and from the presence of specific regulatory motifs, such as transcription and splicing factor binding sites. Moreover, they observed significant GC and TA biases (excess of G over C and T over A) in the noncoding region of genes (Louie *et al*. 2003).

Duan *et al.,* (2003) observed a strong synonymous codon usage bias in the human *DRD2* gene and suggested the role of selection on synonymous positions. It was revealed by the relative independence of the G+C content of the third codon positions from the isochoric G+C frequencies. Investigation of functional effects of the six known naturally occurring synonymous changes (C132T, G423A, T765C, C939T, C957T, and G1101A) in the human *DRD2* gene, it was reported that some synonymous mutations in the human *DRD2* had functional effects and suggest a novel genetic mechanism. The 957T, rather than being 'silent', altered the predicted mRNA folding and led to a decrease in mRNA stability and translation, and dramatically changed dopamine-induced up-regulation of *DRD2* expression. The 1101A did not show an effect by itself but annulled the above effects of 957T in the compound clone 957T/1101A, demonstrating that combinations of synonymous mutations had functional consequences drastically different from those of each isolated mutation. But C957T was found to be in linkage disequilibrium in a European-American population with the -141C Ins/Del and TaqI 'A' variants, which had been reported to be associated with schizophrenia and alcoholism, respectively (Duan J. *et al*. 2003).

dos Reis *et al*., (2003) reported that strong bias leads to high levels of gene expression confirmed by highly significant correlation between gene expression and codon bias (dos Reis *et al*. 2003).

Comeron (2004) observed that in human genome, the highly expressed genes have preference towards codon bias favoring the codons with most abundant tRNA gene copy number compared to the less highly expressed genes (Comeron 2004).

Plotkin *et al.,* (2004) reported that synonymous codon usage bias may play an important role in the differentiation and regulation of tissue-specific gene products in humans. Their results revealed that codon usage pattern in the expression of brain specific-gene differs from liver- specific genes, uterus genes differ from testis genes; and ovary genes differ from vulva genes (Plotkin *et al*. 2004).

Moreover, Semon *et al*., (2006), confirmed that synonymous codon usage differs significantly between the tissues although the effect was very weak. The variability was directly linked to isochore-scale (>100 kb) variability of GC-content that affects both coding and introns or intergenic regions. Their results revealed that the variations of synonymous codon usage between tissue-specific genes expressed in different tissues are due to regional variations of substitution patterns and not to translational selection (Semon *et al*. 2006).

In 2004, Rocha analyzed the selection-mutation-drift theory of codon usage which plays a major role in the theory of molecular evolution by explaining the co-evolution of codon usage bias and *tRNA* content in the framework of translation optimization. From the analysis of the *tRNA* gene pool of 102 bacterial species, they found that as minimal generation times get shorter, the genomes contain more *tRNA* genes, but fewer anticodon species. Surprisingly, despite the wide G+C variation of bacterial genomes these anticodons are the same in most genomes. This suggested an optimization of the translation machinery to use a small subset of optimal codons and anticodons in fast-growing bacteria and in highly expressed genes. As a result, the overrepresented codons in highly expressed genes tend to be the same in very different genomes to match the same most-frequent anticodons. Co-evolution of *tRNA* gene composition and codon usage bias in genomes seen from *tRNA*'s point of

view agrees with the selection-mutation-drift theory. It also provided new evidence that a selective force for the optimization of the translation machinery is the maximization of growth (Rocha 2004).

For better understanding the relationship between the codon frequency bias and *tRNA* abundance in multicellular organisms dos Reis *et al*., (2004) developed the *tRNA* adaptation index (*tAI*) and applied it in 126 fully sequenced genomes ranging from archaea to eukaryotes. This metric is based on the copy number of *tRNA* genes, assumed to be correlated with *tRNA* abundance in cells and also accounts for the binding efficiency of codon-anticodon. They proposed that the co-evolution of genome size and *tRNA* genes explains the observed patterns in translational selection in all living organisms for better understanding of the codon usage across prokaryotes to eukaryotes (dos Reis *et al*. 2004).

Xia (2005), reported that in vertebrate mitochondria, codon usage is maintained by strand-specific mutation bias and the biased codon usage drives the evolution of *tRNA* anticodons (Xia 2005).

Zhou *et al*., (2005), suggested that Guanine (G) and Cytosine (C) contents at third position of codon (GC_3) are good indicators of the degree of base composition bias (Zhou T. *et al*. 2005).

According to Kotler and Lavner (2006), whole genome analysis of the human genome showed that codon bias was related to selection, but in a more intricate way compared to the mechanisms of selection in other organisms. It showed that different selection forces might have shaped codon bias for different amino acids. For a group of amino acids, selection acted to enhance translation efficiency in highly expressed genes by preferring major codons, and acted to reduce translation rate in lowly

expressed genes by preferring non-major ones. For the other group, this included heavier and more complex amino acids, other mechanisms, such as reducing misincorporation rate of expensive amino acids, may be in action (Kotlar and Lavner 2006). Dittmar *et al*., (2006) reported that *tRNA* expression abundance in humans varies widely among different tissues demonstrated by microarray analysis. They further reported that the *tRNA* abundance could be statistically correlated to codon usage of highly expressed genes specific for those tissues (Dittmar *et al*. 2006).

Buchan *et al*., (2006) demonstrated that nucleotides neighboring a particular codon are distributed in a nonrandom manner, termed codon-pair bias which might influence the translational efficiency (Buchan *et al*. 2006).

Stoletzki and Eyre-Walker (2007), reported that codon bias increases along the length of genes, most likely because of relatively higher resource costs for mistranslation of larger proteins (Stoletzki and Eyre-Walker 2007).

Zhou Y. *et al*., (2007), conducted a study on human oncogene through analyzing the correlations between tissue-specific oncogene expressions and sequence compositional features in human tissues. They found significant correlations that provide new evidence to support the existence of translational selection on synonymous codon usage pattern, at least in human oncogenes in certain tissues and some have important implications for diagnosis of cancers (Zhou Y. *et al*. 2007).

Yang and Nielsen (2008), reported that codon bias in mammals is mainly influenced by mutation bias and the selection on codon bias is weak for nearly neutral synonymous mutations (Yang Z. and Nielsen 2008).

Jia and Higgs (2008), conducted an analysis based on the frequencies of synonymous codon usage in animal mitochondrial genome particularly mammals

and fish. In their analysis they observed that the frequencies of bases in codons encoding an amino acid with 4-fold degeneracy are affected by the mutations depending on the adjacent bases. This causes strong correlation between the neighboring bases (Jia and Higgs 2008).

Coleman *et al*., (2008) reported that several viral genomes hold codon pair bias, which generally matches that of their host. Alteration of this codon pair usage in virulence related genes of viruses would be a novel strategy with potential applications to produce vaccines with attenuated viruses (Coleman *et al*. 2008).

Ahn *et al*., (2009) performed a comparative analysis of the synonymous codon usage patterns between the nucleocapsid and spike genes of coronaviruses (CoVs), and C-type lectin domain (CTLD) genes of human and mouse on the codon basis. Their results showed that the nucleocapsid genes of CoVs were affected from the synonymous codon usage bias than spike genes, and the CTLDs of human and mouse partially overlapped with the nucleocapsid genes of CoVs. Besides, they observed that CTLDs which showed the similar relative synonymous codon usage (*RSCU*) patterns with CoVs were commonly derived from the human chromosome 12, and mouse chromosome 6 and 12, indicating that there might be a specific genomic region or chromosomes, which show a more similar synonymous codon usage pattern with viral genes. They reported that the results may contribute to develop codon-optimization process in DNA vaccines (Ahn *et al*. 2009).

In 2009, Najafabadi *et al*., claimed that gene regulation mechanism is a factor involved in the codon usage of organisms by confirming that co-regulated genes share the similar patterns of codon usages (Najafabadi *et al*. 2009).

Yang *et al*., (2010) reported that in the coding sequences of genes involved in Alzheimer's disease, GC bias was present and translationally optimal codons ended in G or C. Thus, besides compositional constraints, translational selection was also considered as a factor in shaping the codon usage. They also reported that codon usage bias plays an important role in understanding the etiology of central nervous system neurodegenerative diseases, especially Alzheimers disease as well as the genetic factors and may help in the possible cures for these diseases (Yang J. *et al*. 2010).

Cannarozzi *et al*., (2010) reported a bias of clustered synonymous codons called codon co-occurrence bias that is recognized by the same *tRNA* molecules, while studying all coding sequences of *Saccharomyces cerevisiae*. The effect of these co-occurrence bias involves both frequent and rare codons and is most prominent in the highly expressed genes that must be rapidly induced, such as those involved in stress response (Cannarozzi *et al*. 2010).

Zhao and Chen (2011) conducted a study on codon usage roles in human papilloma virus and reported that the HPV genomes comprise a strong codon usage bias to 18 codons, with 14 showing T at the third position amongst degenerately encoded amino acids. The codon usage pattern in HPV genome plays an important role, which regulates low or non-translational expression of the viral capsid genes and results in very weak protein expression of oncogenes in a wide range of mammalian cells. Codon modification has been proved to be a powerful technology to overcome the translational blockage and weak expression of both HPV capsid genes and oncogenes in different expression systems. Furthermore, keratinocytes are the host cells of HPV infection; the codon usage in HPV capsid genes matches available

aminoacyl-*tRNA*s in differentiated keratinocytes to modulate their protein expression. HPV DNA vaccines with codon optimization have been shown to have higher immunogenicity and induce both strong cellular and humoral responses in animal models, which may be a promising form of therapeutic HPV vaccines (Zhao and Chen 2011).

Similarly, Norkiene and Gedvilaite (2011) estimated the affect of codon bias on heterologous production of human papillomavirus type 16 (HPV-16) major structural protein L1 in yeast by expressing five variants of codon-modified open reading frames (OFRs) encoding HPV-16 L1 protein and they observed a significant positive correlation between the gene's expression level and the extent of its codon bias towards the preferred codon usage. Their results revealed that the HPV-16 L1 protein expression in yeast can be optimized by adjusting codon composition towards the most preferred codon adaptation, and this effect most probably is dependent on the improved translational elongation (Norkiene and Gedvilaite 2012).

In 2011, Rao *et al*., performed a study on codon usage in *Gallus gallus* genome and reported that GC mutation bias plays an effective role in shaping the codon usage patterns in the *Gallus gallus* genome (Rao *et al*. 2011).

Schmid and Flegel (2011) developed an event based model to measure the risk of acquiring nonsense mutations in the coding sequences and genomes of vertebrates including human. In their analysis of codon usage, they observed that codon bias in the genomes of high GC content is connected with a low risk of acquiring nonsense mutations (Schmid and Flegel 2011).

Dass and Sudandiradoss (2012) reported that nucleotide compositional mutation bias is one of the major factors influencing the codon usage patterns in serotonin receptor

gene family across mammalian species. In addition they observed that, the C–ending codons were most frequently used in comparison to the G–ending codons in serotonin receptor gene family (Dass and Sudandiradoss 2012).

Novoa *et al*., (2012) demonstrated that codon frequency bias can be better correlated with *tRNA* gene frequencies in all kingdoms of life, if two major, domain-specific tRNA modification types are taken into account (Novoa *et al*. 2012).

Doherty and McInerney (2013) performed a comparative analysis among the genes of 23 vertebrates and reported that despite variation in mutational bias, translational selection plays an effective role in shaping codon usage patterns in vertebrates (Doherty and McInerney 2013).

Nabiyouni and her co-workers (2013) performed a large scale bioinformatics analysis of nucleotide composition of coding and non-coding sequences in vertebrates and other taxa and reported that the last common vertebrate ancestor had a GC-rich genome (~65% GC). Their data suggested that the whole-genome mutational bias is the major driving force for generating codon bias and when the bias becomes prominent, it initiates to affect translation and can result in positive selection for optimal codons which in turn significantly modulates the usage of codon preferences (Nabiyouni *et al*. 2013).

Lampson *et al*., (2013) conducted an analysis in the codon usage bias in *KRAS* oncogene and reported that synonymous nucleotide differences affecting codon usage account for differences between *HRas* and *KRas* expression and function and may represent a broader regulation strategy in cell signaling (Lampson *et al*. 2013).

Pechmann and Frydman (2013), reported that universal codon usage patterns has been linked to the folding patterns of the encoded polypeptides (Pechmann and

Frydman 2013). Moreover, Gingold *et al*., (2014) conducted a study comprising human tissue of both cancerous as well as normal patients and suggested that codon context patterns and codon bias affect the differential regulation of protein expression (Gingold *et al*. 2014).

Ingolia (2014) reported that there exists no correlation between frequently used codons and high translation elongation rate as confirmed by earlier studies of ribosome profiling data (Ingolia 2014).

Wei *et al*., (2014) conducted a study based on the analysis of codon usage bias of mitochondrial genome in *Bombyx mori*. The results of their analysis revealed that mitochondrial DNA (mtDNA) of *Bombyx mori* has higher level of codon bias in comparison to that of genomic DNA. Further, they reported that natural selection plays a major role while compositional constraints for mutation bias play a minor role in shaping the high codon bias in the mtDNA of *Bombyx mori* (Wei *et al*. 2014).

Mirsafian *et al*., (2014) conducted a comparative analysis of synonymous codon usage bias pattern in human albumin superfamily, namely, albumin, α-fetoprotein, afamin, and vitamin D-binding protein. They reported that albumin superfamily members have less bias in codon usage preferences and are not subjected to mutational selection pressure (Mirsafian *et al*. 2014).

Recently in 2015, Duan and his co-workers performed an analysis of the codon usage patterns of an important fish species, namely Blunt snout bream (*Megalobrama amblycephala*) because of its delicacy and economic value in China. In their analysis, they reported that GC contents, codon usage and codon context patterns comparison among 23 vertebrates showed species specific variation (Duan X. *et al*. 2015).

Uddin *et al*., (2015) performed an analysis of codon usage bias in mitochondrial *ND2* gene among pisces, aves and mammals. The results of their analysis suggested that codon usage in mitochondrial *ND2* gene is not associated with gene expression level, besides it was reported that both mutation pressure and natural selection affect the codon usage pattern in mitochondrial *ND2* gene (Uddin *et al*. 2015).