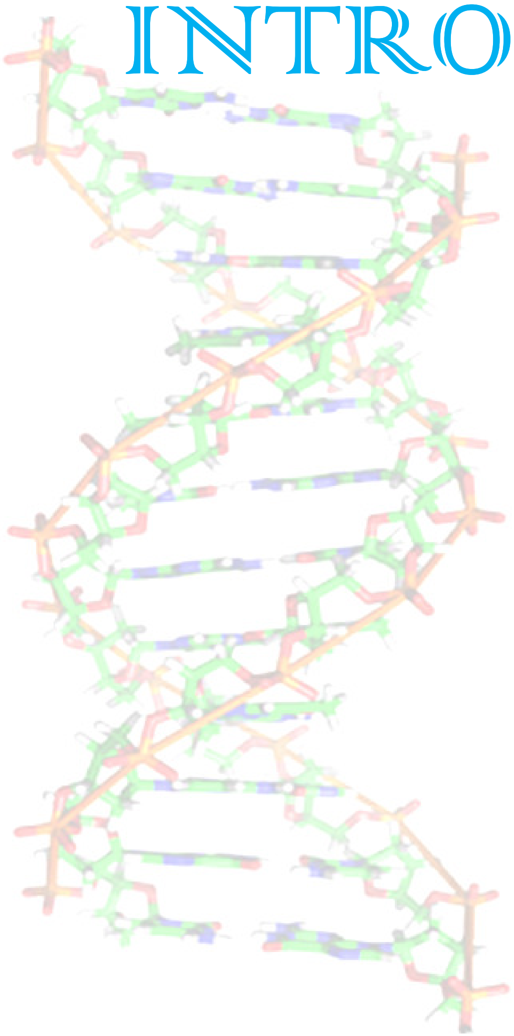
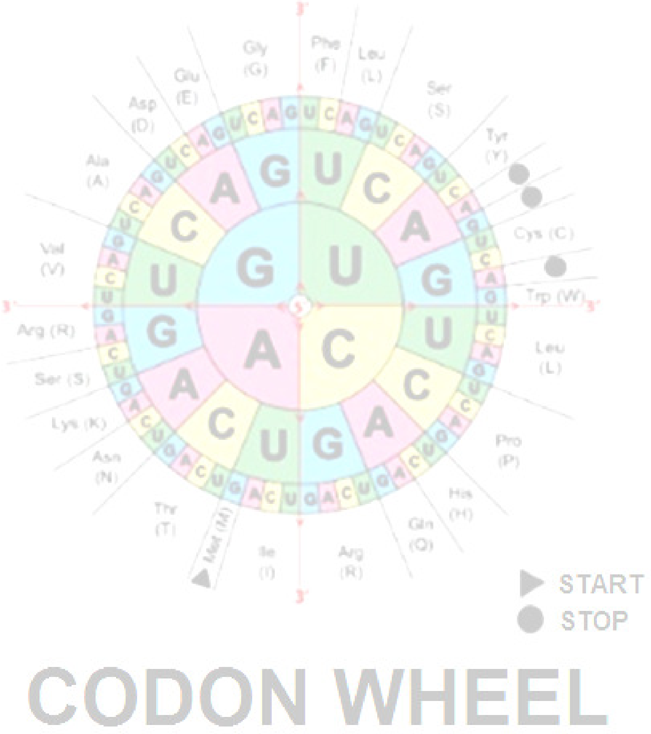
CHAPTER – 1

INTRODUCTION

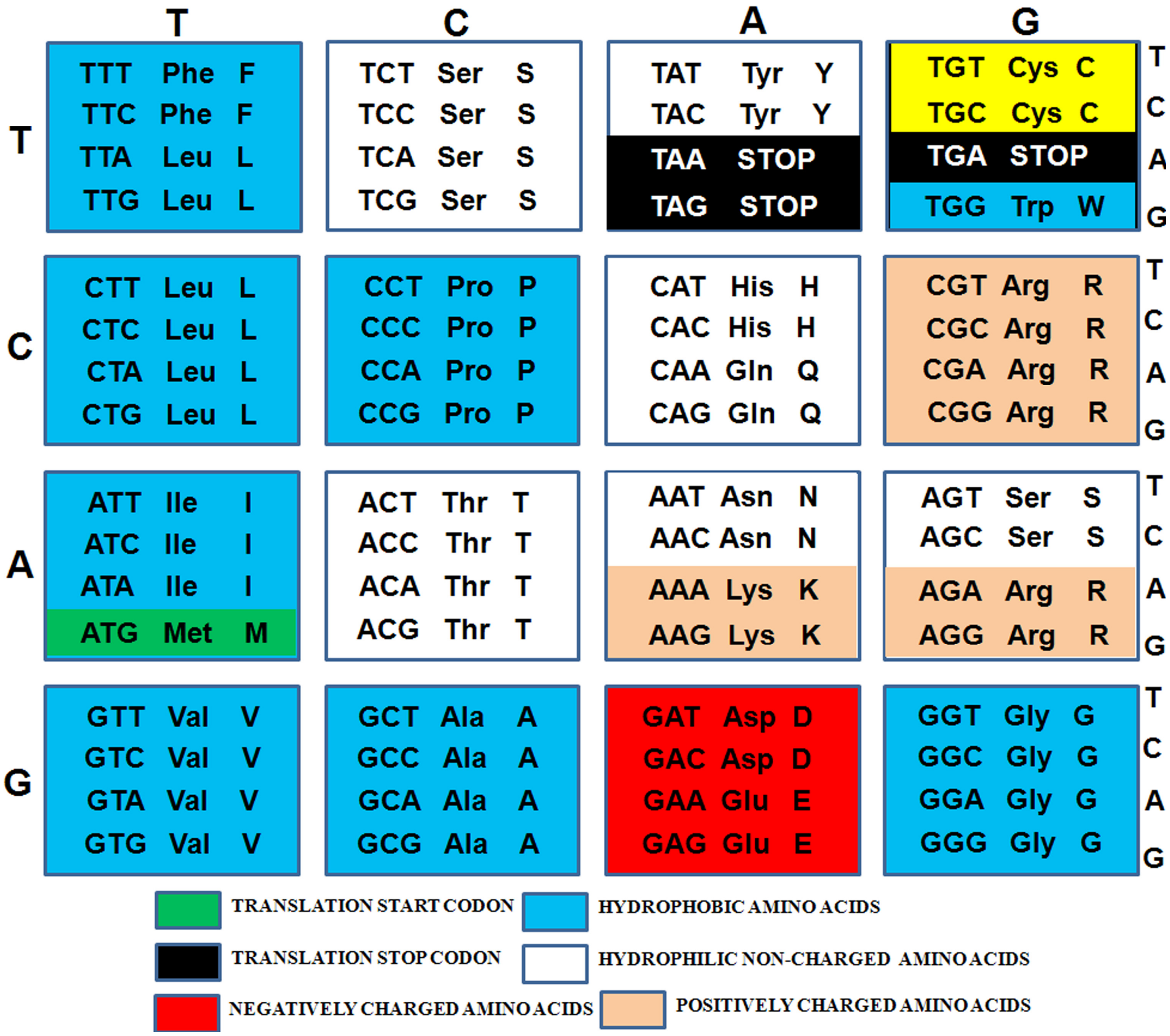




1. Introduction

1.1. Standard genetic code and synonymous codons

A codon refers to a set of three nitrogenous bases which encodes an amino acid that builds up a protein. Nature has gifted the genetic code that provides the basic instructions and information which are transferred from DNA to protein through transcription and translation. There are 61 codons that recognize only 20 standard amino acids found commonly in protein sequences together with three termination signals (UAA, UAG & UGA) (Figure 1). Most of these amino acids (building blocks of protein) are encoded by more than one codon (i.e., a triplet of nucleotides) with the exception of methionine and tryptophan. It indicates a vital step in modulating the proficiency of protein synthesis (Butt *et al*. 2014). This redundancy in the genetic code may have evolved to preserve protein structural information within the nucleotide content (Zull and Smith 1990). The alternative codons that encode for the same amino acid are usually represented as 'synonymous' codons. In nature, two amino acids namely methionine and tryptophan each are encoded by one codon, nine amino acids (phenylalanine, tyrosine, histidine, glutamine, asparagine, lysine, aspartate, glutamate and cysteine) are encoded by two synonymous codons, one amino acid isoleucine is coded by three synonymous codons, five amino acids (alanine, glycine, proline, threonine and valine) are encoded by four synonymous codons, and three amino acids (leucine, serine and arginine) are encoded by six synonymous codons. Figure 1 explains the concept of standard genetic code including codons and their corresponding amino acids.



	T	C	A	G	
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C	T
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C	C
	TTA Leu L	TCA Ser S	TAA STOP	TGA STOP	A
	TTG Leu L	TCG Ser S	TAG STOP	TGG Trp W	G
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R	T
	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R	C
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R	A
	CTG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R	G
A	ATT Ile I	ACT Thr T	AAT Asn N	AGT Ser S	T
	ATC Ile I	ACC Thr T	AAC Asn N	AGC Ser S	C
	ATA Ile I	ACA Thr T	AAA Lys K	AGA Arg R	A
	ATG Met M	ACG Thr T	AAG Lys K	AGG Arg R	G
G	GTT Val V	GCT Ala A	GAT Asp D	GGT Gly G	T
	GTC Val V	GCC Ala A	GAC Asp D	GGC Gly G	C
	GTA Val V	GCA Ala A	GAA Glu E	GGA Gly G	A
	GTG Val V	GCG Ala A	GAG Glu E	GGG Gly G	G

🟩 TRANSLATION START CODON
🟦 HYDROPHOBIC AMINO ACIDS
⬛ TRANSLATION STOP CODON
⬜ HYDROPHILIC NON-CHARGED AMINO ACIDS
🟥 NEGATIVELY CHARGED AMINO ACIDS
🟧 POSITIVELY CHARGED AMINO ACIDS

Figure 1: Standard genetic code along with 20 amino acids and three stop codons

1.2. Codon usage bias (CUB)

Literature suggests that, synonymous codons are not used at equal frequencies in most sequenced genomes and show species specific deviation, which is referred to as codon bias or more preferably codon usage bias (CUB) (Grantham *et al*. 1981, Marin *et al*. 1989). The inherent property of redundancy of amino acid led to the understanding of synonymous mutations. Synonymous mutations that do not change the amino acid in protein do not bring any change in the resulting protein sequence and therefore have no effect on cellular function, organism fitness or evolution (Plotkin and Kudla 2011). Variation in the use of synonymous codons can be found at different levels, between genomes, between genes in the same genome, and within a single gene (Hooper and Berg 2000, Lavner and Kotlar 2005). In highly expressed genes, strong codon bias can be observed exhibiting greater use of a synonymous codon (Henry and Sharp 2007). Such preferred codons are called optimized codons. Expression level of a gene can be predicted by determining the property of codon usage bias (Hiraoka *et al*. 2009).

Previously, several studies were conducted on synonymous codon usage bias in a wide variety of organisms including prokaryotes and eukaryotes (Ahn *et al*. 2009, Ermolaeva 2001, Gu *et al*. 2004, Jenkins and Holmes 2003, Lavner and Kotlar 2005, Liu H. *et al*. 2012, McInerney 1998), but till date in many organisms the codon usage patterns have been interpreted for diverse reasons.

Many genomic factors such as gene length, GC-content, recombination rate, gene expression level, or modulation in the genetic code are associated with CUB in different organisms (Duret 2000, Karlin and Mrazek 1996, Palidwor *et al*. 2010, Roymondal *et al*. 2009, Urrutia and Hurst 2003). In general, compositional constraints under natural selection or mutation pressure are considered as major factors in the codon usage variation among different organisms (Duret and Mouchiroud 1999, Li 1987, Nair *et al*. 2013, Xu *et al*. 2011). The selection associated with translational efficiency/accuracy is often termed as ‘translation selection’. Moreover, studies revealed that mutation pressure, natural or translational selection, secondary protein structure, translational efficiency and fidelity, replication and selective transcription, hydrophobicity and hydrophilicity of the protein and the external environment play major roles in the codon usage pattern of organisms (Butt *et al*. 2014). That is why codon usage bias among different organisms or within the genes of the same organism has attracted much attention of the molecular biologists and various works on the subject have been published in recent years. In unicellular and multicellular organism it was observed that, preferred synonymous codons/optimal codons with abundant *tRNA* gene copy number rise with gene expression level within the genome. It supports selection on high codon bias which has been confirmed by positive correlation between optimal codons and *tRNA* abundance (Akashi 1995, Duret 2000, Ikemura 1981a). Urrutia and Hurst (2003) reported weak correlation between gene expression level and codon usage bias within human genome though not related with *tRNA* abundance (Urrutia and Hurst 2003). However, Comeron (2004) observed that in human genome, highly expressed genes have preference towards codon bias favoring codons with the most abundant *tRNA* gene copy number compared to less highly expressed genes (Comeron 2004).

Lavner and Kotlar (2005) suggested that there are three possible ways in which selection may act on codon bias in the human genome: (1) Increasing translation efficiency in highly expressed genes; (2) Regulating translation efficiency of some proteins that can be a disadvantage at high levels; and (3) Improving translation efficiency and reducing the rate of amino acid misincorporation in the production of biosynthetically expensive proteins (Lavner and Kotlar 2005).

1.3. Factors affecting codon usage bias

1.3.1. *Selection for optimized translation*

Codon usage is biased towards the use of *tRNA*s that are abundant in the *tRNA* pool or towards codons that can bind their complementary *tRNA* more efficiently compared to other synonymous codons (Kurland 1991). Translational efficiency can be achieved by the use of more abundant *tRNA*s, and by avoiding different types. The *tRNA*s which can exhibit accurate codon:anticodon interaction are used more often than those that are ineffective (Rocha 2004).

1.3.2*. Gene Expression*

Gene expression is the process by which DNA is transcribed to mRNA, which is then translated to protein. Protein abundance in a cell depends upon the expression of genes (Klumpp *et al*. 2012). It was revealed from correlation analysis between gene expression and codon bias that strong bias leads to high levels of gene expression (dos Reis et al. 2003). However, the study of human genome revealed that some highly expressed genes and low expressed genes, are positively correlated to strong codon bias (Gouy and Gautier 1982). Thus, high codon bias is not a true indicator of highly expressed genes. In genes that are translated often and at high magnitudes, codon bias appears to be especially high because the cost of a missense error is elevated to filter out the defective proteins in the cells. In highly expressed genes selection usually acts on codon bias to increase the elongation rate by favoring the optimal codons. But in lowly expressed genes selection operates to reduce the elongation rate by favoring the non-optimal codons (Comeron 2004).

Another study in humans examined the relationship between gene expression level and gene expression breadth and codon bias and showed that codon usage is more strongly related to the breadth of expression than to maximum expression level (Kotlar and Lavner 2006).

1.3.3*. Location within genes*

The degree of codon bias in a gene varies along the direction of translation (Qin *et al*. 2004). The preference of low usage codons in the 3´ end of mRNA results in slowing of ribosome in the upstream region indicating as if entire gene consists of less preferred codons, thus lowering translational efficiency (Zhang S. *et al*. 1994). Therefore, strong codon bias in the 3´ end increases the speed of translation and prevents ribosomal clustering.

Second, the abundance of optimal codons may increase along the length of a gene sequence in order to prevent nonsense errors that would become increasingly expensive. In *Escherichia coli* this pattern of increasing codon bias is stronger in longer genes than in shorter genes, and codon bias is positively correlated with gene length (Qin *et al*. 2004). This suggests that as a gene becomes longer, and more energy is required for translation, it is increasingly important to prevent nonsense errors at the 3´ end of a gene that would terminate translation prematurely and make the peptide synthesized up to that point useless.

1.3.4. *Rate of evolution*

Studies in *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Escherichia coli* and *Salmonella typhimurium* have revealed a significant negative correlation between codon usage bias and the rate of nucleotide substitution at silent sites (Powell and Moriyama 1997, Sharp and Li 1987). From the study on *Escherichia coli* and *Salmonella typhimurium*, it was found that, highly expressed genes have high codon bias and low rates of synonymous substitution (Sharp and Li 1987).

Codon preferences reflect a balance between mutational biases and natural selection for translational optimization, and as mentioned before, optimal codons help to increase translation efficiency and accuracy (Akashi 2001). Since optimal codons are preferred by selection, and a synonymous substitution to a non-optimal codon would actually decrease fitness. Moreover, selection among synonymous codons constrains the rate of silent substitution in some genes (Sharp and Li 1989).

1.3.5. ***Secondary structure***

The secondary structural constraints of DNA also play an active role in determining the codon preferences of genes. This holds true in the genomes of most organisms, including those of chickens, *Drosophila melanogaster*, *Caenorhabditis elegans* etc. It was suggested that structural constraints such as, flexibility capabilities and folding manner of DNA and RNA influence codon usage bias than translational constraints do (Karlin and Mrazek 1996). Transcription process is highly constrained owing to the structural properties of the DNA to bend and be flexible. These structural properties are influenced by base sequence and length, which may reflect or influence codon bias, and which often correlate to gene expression levels. DNA that cannot condense tightly into wrapped chromatin (euchromatin) is more accessible to RNA polymerase and thus more highly expressed. Studies also revealed that the choice of codons on the protein coding sequence of mRNA influences folding of protein. This folding stability can go on to affect translation accuracy and efficiency.

Additionally, this suggests that mRNA folding stability might be important in regulating gene expression by influencing codon bias in highly and lowly expressed genes. Studies show that the stability of mRNA folded structure works to discriminate between the highly and the lowly expressed genes coding for irregular portions of protein secondary structure on the basis of amino acid usage of *Saccharomyces cerevisiae* (Hiraoka *et al*. 2009).

1.3.6. ***Nucleotide composition***

Shaping of codon usage depends also on the preference at nucleotide level, especially preference towards GC content. Different organisms show different biasness towards GC percent in their genome. The genome of the ciliate *Oxytricha trifallax* has a GC content of only 39%, with a preference for synonymous codons containing A or T, while the aspen tree, *Populus tremula*, prefers codons ending in G or C (Hiraoka *et al*. 2009). In eukaryotes various compositional constraints have been shown to exist. GC content is correlated to a number of factors such as, codon usage bias, gene length, gene density, replication timing, and methylation. Hypotheses have been proposed that nucleotide patterns may be determined by selection, mutational bias, or recombination, since there is an association between recombination and GC-rich chromosomal regions (Nekrutenko and Li 2000).

1.3.7. *Protein Length*

When genes of similar expression levels are compared, protein length and codon usage bias are positively correlated in both *Saccharomyces cerevisiae* and *Escherichia coli* (Eyre-Walker A. 1996, Nekrutenko and Li 2000). The efficiency of translating an mRNA into protein is proportional to its length, so there is greater pressure for the selection of the most accurate codons in longer cds (genes) to avoid missense errors, explaining the positive correlation between gene length and codon bias. It has also been argued that selection may act to decrease the length of highly expressed genes, especially in eukaryotes, explaining the negative correlation between gene length and codon bias (Nekrutenko and Li 2000).

1.3.8. *Environment*

Environmental conditions, including the types of tissues in which genes are expressed and the specific cellular conditions within these tissues also play a role in influencing codon preferences. From the study of genes expressed in multiple human tissues it was found that codon usage differs for sets of genes expressed in different tissues and is directly affected by the actual amount of *tRNA* molecules in each tissue (Kotlar and Lavner 2006). A second study in human tissues found that varying abundances of *tRNA* isoacceptors are found in different tissues, suggesting a relationship between *tRNA* abundance and codon usage in different tissues. The conditions under which a gene is replicated also appear to affect codon preferences. For example, it has been shown that an overrepresentation of rare codons is seen in genes expressed under starvation conditions. This suggests that during the evolution of genomes, different conditions, providing different restrictions on gene expression have influenced codon preferences. It appears that, in vivo, intracellular factors contribute to the final formation of proteins with influence from ribosomal traffic, chaperones, stress proteins, and foldases. These diverse factors also correlate to varying codon preferences (Moriyama and Powell 1998).

1.3.9. ***Time***

Another important determinant of codon bias is the time and speed of expression. Fast-growing bacteria have more abundant, less diverse *tRNA*s, leading to higher codon bias in highly expressed genes (Rocha 2004). Highly expressed protein tends to have high *CAI* value of coding sequence (Carbone *et al*. 2005). In slow-growing organisms with low codon bias, *CAI* is a less effective indicator of highly expressed genes (Willenbrock *et al*. 2006).

1.3.10. ***Neutral alternatives***

It has been argued that neutral processes such as gene conversion and mutational bias can affect codon usage bias. For example, transcription process is mutagenic, thus those genes which are frequently transcribed (*i.e*. highly expressed) has strong codon bias as a side effect. However, studies in *Drosophila* and *C. elegans* genomes showed that this transcription-coupled mutational process could not explain the observed codon bias in these species and that synonymous codon usage in these organisms is shaped by natural selection (Duret and Mouchiroud 1999).

Another neutral process, biased gene conversion, is sometimes invoked to explain the correlation between codon bias and protein sequence evolution (Kellis *et al*. 2004). Study of gene duplication in yeast genome suggested that gene conversion is a minor determinant in the evolution of proteins, while codon bias and functional constraints play a major role in evolutionary rate of proteins (Lin *et al*. 2006).

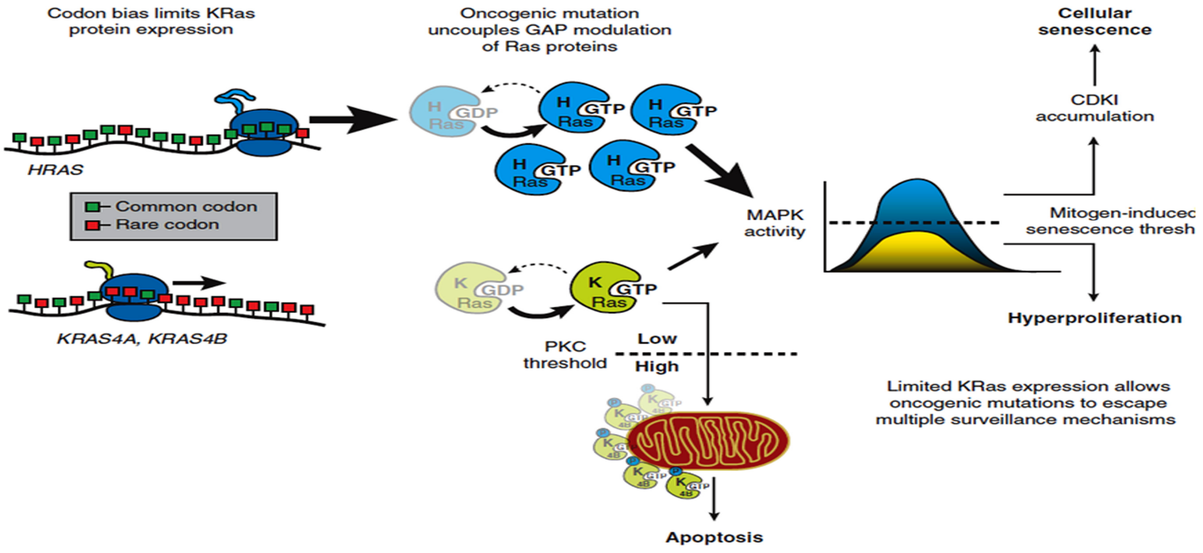
1.4. Cancer: proto-oncogenes/oncogenes

Cancer is one of the frequent and complex diseases occurring in the multiple organs per system, multiple systems per organ, or both, in the body that leads to major morbidity and mortality cases in human. A group of genes known as proto-oncogenes causes normal cell to become cancerous when they are mutated (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. Therefore, oncogenes are known as the mutant version of proto-oncogenes that function autonomously without a requirement for normal growth promoting signals. A proto-oncogene is converted or activated to oncogene either by changing the structure of the gene or by changing the regulation of the gene expression. There are several genetic mechanisms associated with oncogene activation, which are as follows:

- Point mutations, deletions/insertions that lead to a hyperactive gene product
- Point mutations, deletions/insertions in the promoter region of a proto-oncogene that guide to increased transcription
- Gene amplification events which result in extra chromosomal copies of a proto-oncogene
- Chromosomal translocation events that relocate a proto-oncogene to a new chromosomal site leading to higher expression

- Chromosomal translocations that lead to a fusion between a proto-oncogene and a second gene, which produces a fusion protein with oncogenic activity (Chial 2008).

Thus, conversion or activation of a proto-oncogene into an oncogene involves *gain-of-function* mutation that has the potential to induce cancer (Blanchard 2002, Todd and Wong 1999). Many genomic analyses have been done on oncogenes but till date very little is known about the codon usage patterns and the factors that influence them. Bodemann and White (2013) reported that codon bias may select for the prevalence of *KRas* mutations in human cancers (Bodemann and White 2013).


Codon bias limits KRas protein expression
Oncogenic mutation uncouples GAP modulation of Ras proteins
Cellular senescence
CDKI accumulation
HRAS
🟩 Common codon
🟥 Rare codon
KRAS4A, KRAS4B
H Ras GDP
H Ras GTP
K Ras GDP
K Ras GTP
MAPK activity
Mitogen-inducec senescence thresh
Hyperproliferation
PKC threshold
Low
High
Limited KRas expression allows oncogenic mutations to escape multiple surveillance mechanisms
Apoptosis

[Adapted from (Bodemann and White 2013)]

Figure 2: **Codon bias may select for the prevalence of KRas mutations in human cancers.** Rare codons limit KRas protein translation compared with HRas, which favours the accumulation of the latter. Paradoxically, the otherwise subordinate KRas protein may permit mutations at the *KRAS* locus to escape tumour suppressor surveillance mechanisms. Codon-biased mutant KRas protein is thus free to drive neoplasia *en route* to tumorigenesis (Bodemann and White 2013).

1.5. Tumour suppressor genes

Tumour suppressor gene, *i.e*. 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 contributes to tumour development in combination with other genetic changes (Vousden and Lu 2002).

The tumour suppressor gene, *TP53* is known as the guardian of the genome as it plays an important role in DNA damage repair, cell cycle progression and initiation of programmed cell death. Our cells are blessed with a good amount of xenobiotic metabolizing enzymes (XMEs) that play a vital role in biotransformation of foreign compounds which comprise of drugs, food additives, environmental pollutants, chemical carcinogens, pesticides, herbicides and even natural plant compounds (Gonzalez 2004). The reactions characterized by those enzymes are generally divided into two groups called Phase I and Phase II enzymatic reactions. Whenever any xenobiotic enters the body, they are acted upon by Phase I enzymes {Cytochrome P450s family (CYPs), epoxide hydrolase etc.}, which catalyze the addition of functional groups (such as –OH,-SH,-NH2, etc.) into the lipophilic xenobiotics by oxidation-reduction reactions. This addition converts them into more hydrophilic compounds helping in easy excretion (Ghosh *et al*. 2012). The creation of such reactive centre allows phase II enzymes {such as Glutathione S-transferases(GSTs) etc.} to introduce a hydrophilic moiety (such as glutathione etc.) into the molecule resulting in the production of its water-soluble form which is easily excreted out through urine, feces, breath and sweat (Kumar *et al*. 2009). However, the failure to biotransformation of xenobiotics guides to adduct formation with DNA, RNA or cell protein which results in serious cell damage and every type of DNA damage is first reported to the tumour suppressor TP53 protein and its pathway (Figure 3) (Mazumder *et al*. 2014).

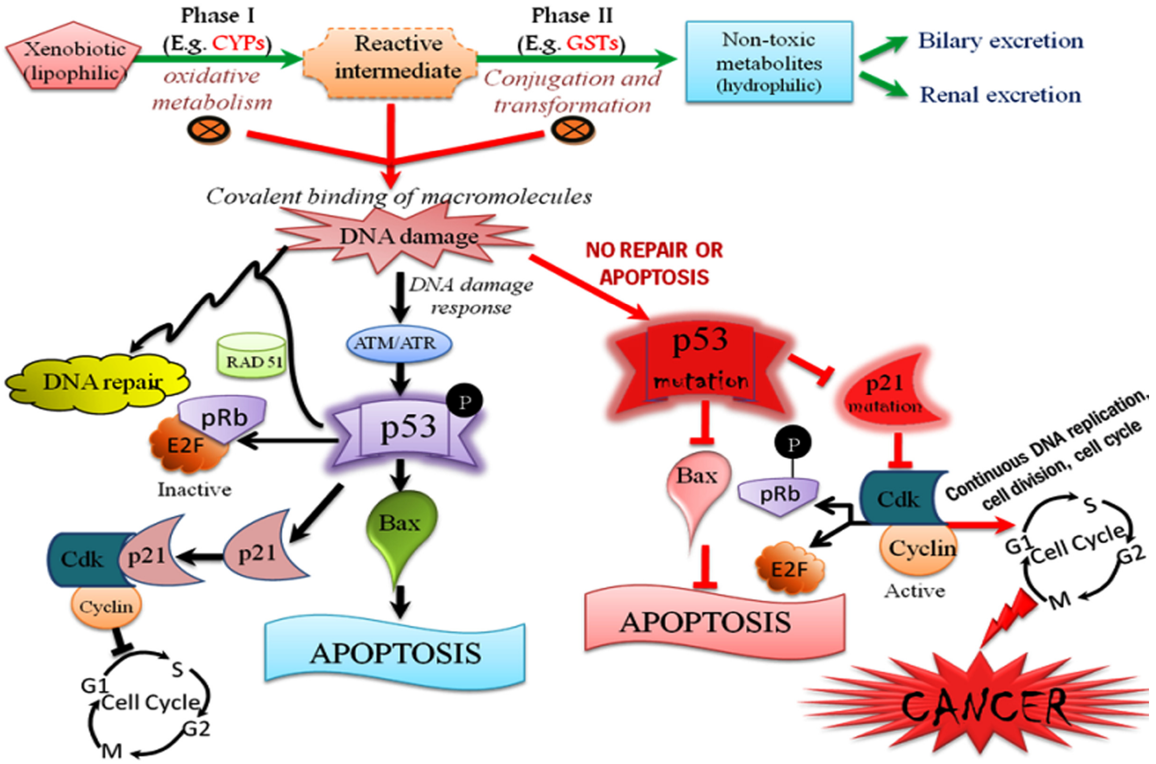


Figure 3: **A schematic representation of tumour suppressor gene *p53* pathway and its mechanism of action during DNA damage repair and cell cycle progression**. In the case of DNA repair failure, TP53 induces a wide variety of genes that participate in TP53-mediated cell death (apoptosis) either by extrinsic (the death receptor) or intrinsic (the mitochondrial) pathways. TP53 mutation inhibits the activation of p21 (CDKN1A) causing pRb (RBL1) phosphorylation and subsequent release of E2F1, which inhibits cell cycle arrest leading to uncontrolled cell proliferation (Mazumder *et al*. 2014).

1.6. The importance of codon usage bias study

The study of codon usage bias acquires significance in biology not only in the context of understanding the process of evolution at molecular level but also in designing transgenes for increased expression, discovering new gene (Carbone *et al*. 2003) based on nucleotide compositional dynamics, detecting lateral gene transfer and for analyzing the functional conservation of gene expression (Lithwick and Margalit 2005). Codon usage bias may be superimposed on the effect of natural selection. The amount of protein produced from the mRNA transcript may vary significantly since the translational properties of alternate synonymous codons are not equivalent (Miyasaka 2002). Several studies have further shown that codon usage bias is associated with highly expressed genes as some codons are used more often than others in the coding sequences (Sueoka 1988).

1.7. Rationale of the present study

The present study was carried out in order to analyze the codon usage bias and codon context patterns among the nucleotide coding sequences of human proto-oncogenes/oncogenes and tumour suppressor genes by 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 for the background nucleotide constraints.

The major objective of this study is to understand the key genetic factors playing crucial roles in determining the codon usage patterns of these genes in human.

1.7.1. *Objectives*

1. To investigate the compositional constraints of human proto-oncogenes, oncogenes and tumour suppressor genes,
2. To analyze and compare the overall GC, GC_1, GC_2 and GC_3 contents
3. To compare the codon usage pattern of these genes
4. To study the interrelationships of codon usage bias and compositional constraints
5. To predict the level of expression of these genes using *CAI*
6. To analyze the relationship of *CAI* with codon usage
7. To analyze the relationship of *tAI* with codon usage
8. To make a comparative analysis of codon usage in *TP53* and *GATA2* gene across mammals including human