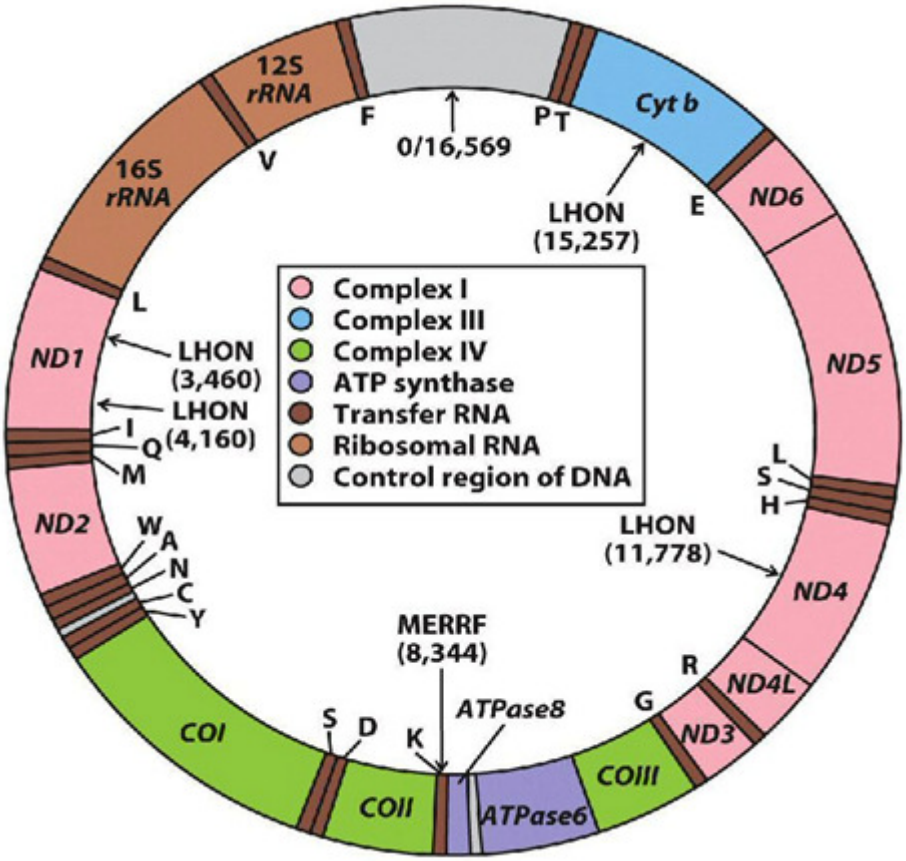
CHAPTER-1

INTRODUCTION



1. Introduction

1.1 Genetic code

A codon is a sequence of three nitrogen bases which encodes a particular amino acid. The genetic code is the set of codons which encode twenty amino acids and protein stop signals. In living cells, genetic information is encoded in DNA, which is transcribed into mRNA transcripts, and subsequently translated into proteins. Standard genetic code consists of 64 codons, out of which 61 represent 20 standard amino acids and the remaining three are non sense or stop codons (TAA, TAG and TGA). Codons that encode the same amino acid are termed as synonymous codons. Each amino acid can be encoded by just one codon (Met, Trp) or by two to six synonymous codons. The preference of codons that encode same amino acids is species specific and hence the codons occur at unequal frequencies in genes (Plotkin and Kudla 2011, Plotkin *et al*. 2006).

1.2 Properties of genetic code

The main characteristics of the genetic code were worked out during the 1960s. Its most important properties are:

- The genetic code is composed of codons (nucleotide triplets). One amino acid in the polypeptide sequence is encoded by three nucleotides in mRNA; thus each codon includes three nucleotides.
- The genetic code is non-overlapping. Each nucleotide in mRNA belongs to just one codon except in rare cases where genes overlie and a nucleotide sequence is comprehended by two dissimilar reading frames.
- The genetic code is comma-less. There are no punctuation within the coding regions of mRNA molecules. During translation, the codons are read one after the other.
- The genetic code is degenerate. All of the amino acids except Met and Trp are encoded by more than one codon.
- The genetic code is well-organized. Multiple codons for a given amino acid and the codons for amino acids with similar chemical properties are closely related, generally differing by a single nucleotide.
- The genetic code contains initiation and termination codons. Specific codons are used to initiate and terminate polypeptide sequence.

- The genetic code is universal. With the exception in mitochondrial DNA, the same codon codes for the same amino acid in every form of life, be it bacteria or human beings *i.e.* the codons have the same meaning in all living organisms, from viruses to humans.

1.3 Codon Usage Bias

Synonymous codons encoding a particular amino acid are not used with equal frequency regardless of the degeneracy of the genetic code due to a phenomenon known as codon bias (Ikemura 1981). In genes, a number of codons are used more often than other synonymous codons which create such bias. These codons are called optimal or preferred codons. CUB is a common trend in an extensive variety of organisms, including prokaryotes as well as eukaryotes (Akashi 1997, Sharp *et al.* 1993). The pattern of synonymous codon usage is a unique property of a genome (Grantham *et al*. 1980). Moreover, within the same organism, various tissues exhibit a different codon usage pattern (Plotkin *et al.* 2004). However, mutations in the wobble base *i.e*. the third codon position usually change the synonymous codons without changing the encoded amino acid thereby maintaining the primary sequence of the protein (Biro 2008).

Most protein-coding DNA sequences use synonymous codons with unequal frequencies. The first reports of random codon usage date to as early as four decades ago. Clarke (1970) and later Ikemura (1981a), proposed that codon usage adapted to match an organism's tRNA pool (Clarke and Clark 2010, Ikemura 1985). Ikemura (1981a) concluded that evolutionary forces acting on the preferences of codons results in differences in codon bias between species (Ikemura 1985). Codon usage bias can vary widely not only between organisms, but also within a genome. For example, eukaryotic genomes are identified to display heterogeneous nucleotide content generating an isochore structure. Isochores are long segments of DNA with relatively uniform GC content (Macaya *et al.* 1976). Protein-coding genes are found in isochores and as a result, it affects codon usage in genes inside isochores, although codon usage bias does not directly affect the protein sequence, it may have essential impact on the protein product and cellular processes. But the exact mechanisms operating the synonymous variation are still not well understood. It is an evolutionary relic. Its role is significant to understand the evolution of genome (Jenkins and Holmes 2003). Codon usage pattern is affected by various factors such as compositional bias (GC% and GC skew), mutation pressure, natural selection, gene length, expression level, replication, RNA stability, hydrophobicity and hydrophilicity of the

protein (Akashi 1997, Moriyama and Powell 1998, Powell and Moriyama 1997, Powell *et al.* 2003). Generally, compositional constraints under mutational pressure and natural selection have been found to be the two main evolutionary forces accounting for codon usage variation among genes (Sharp *et al*. 1993, Shields *et al*. 1988, Stenico *et al*. 1994). In some organisms, codon usage is determined by the mutation pressure and genetic drift while in others, it is due to the balance between natural selection and mutational biases (Bulmer 1991). In certain genes with extremely high content of any one of the four nucleotides, mutation pressure plays an important role in affecting the synonymous codon usage pattern (Karlin and Mrázek 1996, Sharp *et al.* 1993, Zhao *et al*. 2007, Zhong *et al.* 2007). The percentage of very high or low G or C nucleotide in the 3rd codon position in an open reading frame indicates mutational bias (Sueoka 1988).

Some earlier findings suggested that the genes which have high level of expression, the codon usage bias is due to the phenomenon of translational selection. In highly expressed genes, preferred codons are easily recognized by the abundant tRNA molecules (Bibb *et al*. 1984, McEwan and Gatherer 1999).

1.3.1 Factors for codon usage bias

1.3.1.1 Mutational pressure

The DNA compositional constraints in presence of mutation pressure and natural selection are the major factors which vary across species (Sharp *et al*. 1986, Sharp *et al*. 1993). The modifications of biochemical mechanism *i.e.* more frequent changes of certain bases than others cause mutational biases (Francino and Ochman 2001, Green *et al.* 2003). Mutation pressure is mainly responsible for codon usage pattern in some prokaryotes and in many mammals with high AT or GC contents (Sharp *et al*. 1993, Zhao *et al.* 2007). However, in *Drosophila* and in some plants, the codon usage pattern is mainly governed by translational selection (Liu *et al*. 2004). The non synonymous substitution is driven by selection because it alters the amino acids and thus affects protein's biochemical nature (Plotkin and Kudla 2011).

1.3.1.2 Selection affecting codon usage

In comparison to mutational pressure, natural selection may also affect the synonymous codon usage bias. Codon usage bias due to selection may be unambiguous to genes or even positions of codon and it can initiate more capable or accurate translation of protein or

protein folding. These patterns can be found by comparing the coding and the non-coding regions of DNA. The changes caused by neutral mutational processes, where selection acts, may arise from several sources and may vary in strength. For example, the synonymous codon usage is influenced by mutational bias in some genes but in other genes synonymous codon usage is governed by translational selection. Different types of selection operate at different levels and the codon usage patterns are either avoided or preferred at the DNA level. These can be related to packing of DNA in nucleosomes and other changing nucleotide distributions along the genome. The mRNA with more abundant nucleotides is transcribed more quickly in which selection occurred for effective transcription at the RNA level (Xia 1996). Selection can also have effect at the mRNA level, where some codon usage patterns are avoided or preferred, to persuade mRNA folding and decay. Codon usage bias also correlates well with the mRNA levels which indicate that there is a global optimization of minimizing the time the ribosomes are engaged in translation of the mRNA. Codons with positive selection have corresponding tRNAs in larger quantities and at the ribosome, probably they bind to the mRNA more rapidly (Ran and Higgs 2010)

In fast-growing organisms with huge population size, the codon usage pattern is mainly driven by selection (Green *et al*. 2003, Ikemura 1982, 1985, Sharp and Li 1987). However, the effect of natural selection in codon usage in the mammalian genome is considered to be low (Duret 2002, Sharp *et al*. 1995). This is due to small population size in many mammalian species, and the codon usage pattern is due to the effect of genetic drift (Keightley *et al*. 2005, Sharp *et al*. 1995). But with the exception, in non mammalian species highly expressed genes with high codon usage bias are under selection pressure to diminish the error in expression level (Hershberg and Petrov 2008). Essentially, the efficiency of gene expression is due to the redundancy of genetic code tuned by selective forces (Gingold and Pilpel 2011). Moreover, codon usage declines the proofreading expenses by reducing the time and energy required to discard the non-cognate tRNAs (Bulmer 1991). Use of unpreferred codons would increase proofreading expenses and would result in a net decline in the protein levels.

1.3.1.3 Gene expression level

The association between codon bias and the level of gene expression has been experimentally established in *E. coli* (Andersson and Kurland 1990). Moreover, the *in-vitro* expression proficiency has been shown to be significantly increased by using the preferred codons of the host cell in heterologous genes of cultured eukaryotic cells (Kim *et al.* 1997).

1.3.2 Parameters for CUB

Several parameters are widely used to study codon usage bias and these include effective number of codons (ENC)**,** codon bias index (CBI), frequency of optimum codons in a gene (Fop), codon adaptation index (CAI), intrinsic codon deviation index (ICDI); frequency of the synonymous codons (Freire-Picos *et al*. 1994, Tanguy *et al.* 2008, Wright 1990). Moreover, mutational response index (MRI) determines the extent to which a gene responds to mutational pressure (Malumbres *et al*. 1993). The number of amino acids (Laa) in a protein determines the number of translatable codons (Gatherer and McEwan 1997). Some of these measures used in our study are discussed below:

1.3.2.1 Base composition at silent sites

In highly expressed protein, there is a selection for optimal codons, which have pyrimidine, particularly C at the 3rd position of codon. As a result, GC content at the 3rd position of codon is often correlated with gene expression (Shields *et al.* 1988). The nucleotide composition at the 3rd position of codon correlated with the 1st principal component, which contributed a major fraction of the variance as suggested from multivariate analyses. Several studies suggested that selection acting on silent-site base composition exists (Eyre-Walker and Keightley 1999, Stenico *et al.* 1994). The G and C nucleobases are stronger than A and T nucleobases which binds more strongly to each other due to 3-hydrogen bonds. For this reason, they are likely to be more important on codon usage. The base composition at the wobble positions *i.e.* GC content at the 3rd position of codon (GC3) can be used as a measure of codon bias.

1.3.2.2 Relative Synonymous Codon Usage (RSCU)

Relative synonymous codon usage (RSCU) is the observed frequency of a codon to the expected frequency if all synonymous codons of a particular amino acid are used evenly. RSCU value of a codon >1.0 indicates that the corresponding codon is used more frequently than the expected frequency whereas the RSCU value < 1.0 indicates that the particular codon is used less frequently. Besides, the RSCU value > 1.6 is treated as over represented codon while RSCU value < 0.6 is treated as under- represented codon (Behura and Severson 2012, Sharp and Li 1986).

1.3.2.3 Effective Number of Codons (ENC)

The effective number of codons (ENC) is the commonly used parameter to measure the usage bias of synonymous codons (Wright 1990). The ENC value ranges from 20 (when only one codon is used for each amino acid) to 61 (when all codons are used randomly). Higher ENC value means low codon usage bias and vice-versa. ENC values < 35 are generally considered as the significant codon usage bias.

1.3.2.4 Codon adaptation index (CAI)

The codon adaptation index (CAI) is a very extensively used parameter to measure the codon usage bias and the gene expression level. Its value ranges from 0 to 1; with high value indicating a higher proportion of the most abundant codons coupled with high expression level and vice-versa. CAI is a measure of the relative adaptedness of the codon usage of a gene to the codon usage of the highly expressed genes (Sharp and Li 1987). The relative adaptiveness (ω) of each codon is the codon usage of each codon, to that of the most abundant codon within the same synonymous family.

1.3.3 Application of codon usage bias

Analysis of codon usage pattern has an immense importance in understanding genome evolution (Sharp and Matassi 1994). It also has significant impact in better understanding of molecular biology and evolution (Yang *et al.* 2014), design of transgenes (Yang *et al.* 2014), new gene discovery (Yang *et al*. 2014), determining the origin of species (Ahn *et al*. 2006), design of primers (Zheng *et al*. 2007), heterologous gene expression (Kane 1995), prediction of expression level (Gupta *et al.* 2004) and the prediction of gene function (Lin *et al.* 2002).

1.3.3.1 Cotranslational Protein Folding

Protein synthesis occurs at the peptidyl transferase site of the ribosome. It is a 25 nm diameter structure found within the core of roughly spherical ribosome. When the ribosome reads through the mature mRNA sequence, the newly synthesized and the continuously growing polypeptide chain pass through a 10 nm long tunnel. This tunnel, being too narrow, does not permit the growing polypeptide chain to acquire tertiary structure but permits the alpha-helical conformation (Woolhead et al. 2004) (**Figure 1.1 a**). When the N-terminus of the growing polypeptide reaches a length of 30-40 aa, it emerges on the surface of the ribosome and has the ability to begin folding before the C-terminus end appears. The amino acid sequence of the protein provides all the necessary information to specify its native 3-dimensional structure (developed by Anfinsen) (Anfinsen 1972). The native structure of a protein need not represent a unique thermodynamic energy minimum. In some proteins, the native structure may represent a kinetically trapped state which can acquire more than one native structure (Baker and Agard 1994, Burmann *et al*. 2012, Luo *et al.* 2004, Sinclair *et al.* 1994). Usually the larger proteins with more complex structural properties might depend on cotranslational folding to avoid wrong folding and protein aggregation. Synonymous codon substitution/ usage can modulate the translation rate and provide more time to the N-terminus to fold correctly. Synonymous rare codon substitutions play a crucial role in protein folding. Rare codons slow down the translation rate and thereby reduce the probability of protein misfolding. Replacing the rare codons of mRNA with more common synonymous codons results in faster translation but reduces the specific activity of the enzyme-protein (Sun *et al.* 2001).

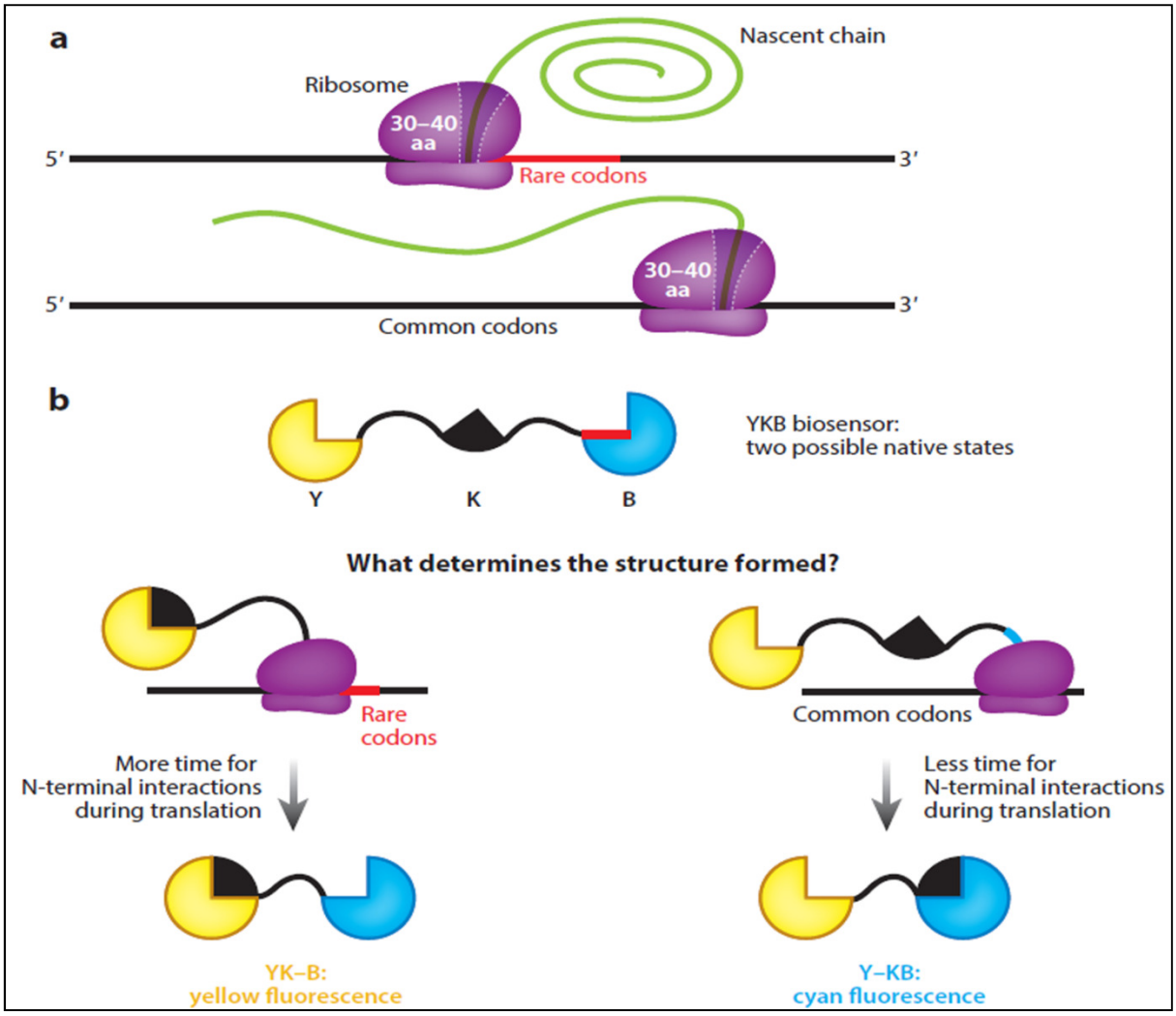


Figure 1.1(*a*) Rare codons are hypothesized to introduce translational pauses that modulate cotranslational protein folding. **(*b*)** The YKB biosensor is a split fluorescence system designed to detect changes in translation rates due to synonymous codon usage (adopted from Chaney and Clark, 2015)

Zhou *et.al* (Zhou *et al*. 2013) showed that synonymous codon substitutions can impact the physiology of the organism. The protease susceptibility of the FRQ protein, a fungal circadian clock regulator, was found to change as a result of rare codon substitution with more common synonymous codons.

The presence of rare codons in the mRNA may induce translational pauses during protein synthesis by the ribosomes. Rare codons assist in proper cotranslational folding (Fedyunin *et al.* 2012). Even increasing the cellular tRNA levels is expected to raise the translational rate of codons and reduce the translational pauses. These examples amply provide compelling a evidence that the synonymous codon usage has a significant role in proper cotranslational folding of some proteins. But the magnitude to which synonymous codon

usage affects the cotranslational folding remains still unknown. Hence, it is a dire need to study the codon usage in a more systematic way in order to develop a predictive model for better understanding of the effect of codon usage on cotranslational protein folding. But the development of such a predictive model appears to be an uphill task because the factors that affect the translation rate are likely to differ between organisms.

1.3.3.2 Cotranslational Interactions with Other Cellular Components

The nascent chains of polypeptide may form cotranslational interactions with other cellular components which undergo covalent modifications during translation (Jha and Komar 2011, Kramer Günter *et al*. 2009). For example, the cotranslational interactions between N-terminal nascent chain signal sequences and the signal recognition particle, which primarily arrest translation elongation to make sure subsequent cotranslational translocation of the nascent polypeptide chain once it appears at the endoplasmic reticulum translocon (Walter and Blobel 1981, Wickner and Schekman 2005, Wiedmann *et al*. 1986). Since the $5'$ ends of coding sequences of secreted proteins are more abundant in rare codons compared to those of cytosolic proteins, it has been assumed that they might contribute to the increase of membrane targeting and secretion efficiency (Clarke and Clark 2010). The translational pauses due to rare-codon are assumed to allow additional time for other cotranslational interactions as well (Clarke IV and Clark 2008). For example, with the reduction of translation rate in actin by rare-codon, the probability of the cotranslational arginylation increases (Zhang Fangliang *et al*. 2010). It is a covalent modification that regulates actin activity and checks aggregation of actin filaments (Karakozova *et al*. 2006).

1.3.3.3 Functional Regulation of Expression Level

The codon usage can considerably alter protein expression levels in addition to affecting translation rates and cotranslational processes. In fact, because of the well-known negative effects of rare codons on gene expression level, it is a common practice while expressing a human gene in *E. coli* or another expression host, to substitute the arginine codons AGA and AGG. These codons are common in *Homo sapiens* but very rare in *E. Coli* and hence these codons in *E.coli* are changed with more common versions to increase expression level. In highly expressed genes, coding sequences are abundant in common codons. This finding supports the hypothesis that selection favours the common codons, which allow

most rare codons to gather in lowly expressed genes, which are thought to be under weaker selection (Sharp and Li 1987). However, this view was contradicted by elucidating the functional implication of expression level regulation by rare codons in certain sequences. Xu *et.al*. (Xu *et al*. 2013) tested the effects of codon usage of circadian clock proteins KaiB and KaiC, in cyanobacterium (*Synechococcus elongatus*) whose wild-type coding sequences contain many rare codons. He replaced these rare codons with common synonymous codons which increased the expression level of protein. The increased expression level disrupted the circadian growth rhythms of the organism (Xu *et al.* 2013). This result suggested that synonymous codons may perhaps be under selection to control the protein expression level and are not only a by-product of mutational drift in lowly expressed genes.

1.3.3.4 Viral Codon Usage

Viruses provide excellent example of genes that adapt to expression conditions. Viral genes rely on host translational machinery for expression which has evolved different mechanisms to utilize the host codon usage biases. The high expression viral coding sequences characteristically have codon usage biases similar to those of their host organisms (Bahir *et al*. 2009). However, some viral genes are rich in codons that are rare in their hosts' genomes, which provide a mechanism to decrease the expression of viral protein and reduce the host immune system responses (Mueller *et al*. 2006, Tindle 2002). A number of viruses also encodes their own tRNA (often a tRNA that is rare in the host) which allows more efficient translation of viral transcripts (Dreher 2010).

1.3.3.5 Codon Usage and Human Health

The mechanisms by which synonymous codon substitutions alter protein biogenesis played a key role for understanding the relationship between synonymous codon usage and human health. Many disease-associated SNPs are nonsynonymous, but some synonymous SNPs have been identified in diseases (Sauna and Kimchi-Sarfaty 2011). For example, synonymous mutation of Ile507 ATC→ATT in CFTR gene alters the stability of mRNA and intensifies the effects of the cystic fibrosis mutation. In multi drug resistance 1 (MDR1), three disease associated SNPs (two synonymous and one nonsynonymous) in a gene encoding a trans-membrane efflux pump correlate with patient responses to chemotherapy (Fung *et al.* 2014, Kimchi-Sarfaty *et al.* 2007). In comparison with wild-

type MDR1, the SNPs produced an efflux pump with an altered structure (assessed by protease susceptibility and antibody binding) and altered the drug-pumping activity. Further, in addition to changing co translational folding, the pathogenic synonymous SNPs can change the splicing motifs that are important to alternative splicing (Daidone *et al.* 2011, Du et al. 1998, Faa *et al*. 2010). Supek *et.al.* (Supek *et al*. 2014) reported that in human cancers, synonymous mutations have frequently influenced the mutations and hypothesized that some of these synonymous SNPs could affect the protein folding, because in the beginning of α-helical regions in oncogenes, the SNPs are more abundant.

1.3.3.6 Codon Usage in Horizontally Transferred Genes

Some genes were found to have remarkable codon usage because they were horizontally transferred between different species or strains. Horizontal transfer frequently involves plasmids, and in bacteria, horizontally transferred genes are associated with virulence or antibiotic resistance. But this phenomenon can also occur in eukaryotes and the transferred genes can be integrated into the genome (Gyles and Boerlin 2013, Syvanen 2012). Genes that have undergone horizontal transfer often have remarkable GC content and codon usage compared with their host genome, and researchers have used this remarkable codon usage to design algorithms to recognize the putative horizontally transferred genes (Garcia-Vallvé *et al.* 2003).

1.3.3.7. Heterologous Protein Expression

Similar to naturally-occurring horizontal gene transfer, genes are also commonly transferred between species in the laboratory. This is because the synonymous codon usage can be functionally significant, and the rare and common codons distribution within the same coding sequence can be hugely different. Variations in codon usage frequencies between organisms can have consequences for heterologous protein expression (**Figure 1.2**). This result has led to a number of algorithms that regulate the codon usage of a sequence for expression in a particular organism (Fuglsang 2003, Grote *et al*. 2005, Gustafsson *et al*. 2004).

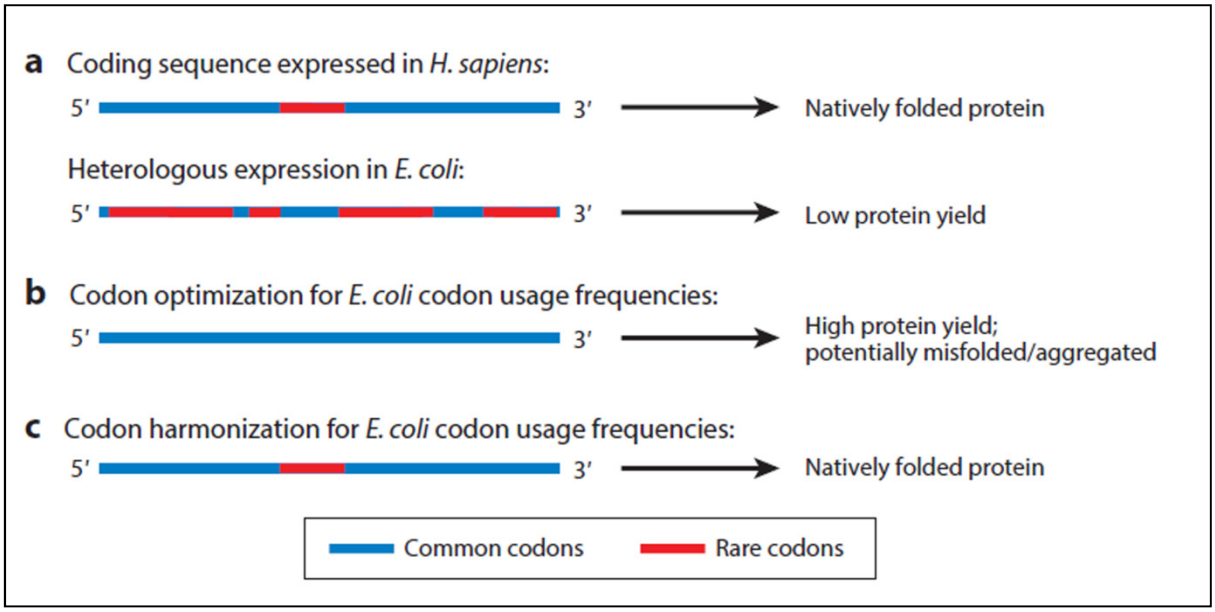


Fig. 1.2 Hypothetical example of heterologous expression of a human coding sequence in *Escherichia coli*. (*a*) Because many codons that are common in *Homo sapiens* are rare in *E. coli*, a human coding sequence often contains. more rare codons when expressed in *E. coli*. This can lower protein expression levels. (*b*) To solve this problem, coding sequences are frequently optimized for the codon usage of the heterologous expression host by selecting only synonymous codons that are common in the host. This approach often results in increased expression levels, but much of the resulting protein product may be misfolded. (*c*) Codon harmonization is based on the hypothesis that the original pattern of rare and common codon usage, under native expression conditions, promotes proper folding. Instead of choosing all common synonymous codons, codons that are rare in *H. sapiens* are replaced by codons that are rare in *E. coli*, and codons that are common in *H. sapiens* are replaced by codons that are common in *E. coli* (adopted from Chaney and Clark, 2015 (Chaney *et al*. 2015).

1.4. Mitochondrial DNA

The mitochondrial DNA is a covalently closed, double stranded structure with nearly 16.6 kb size, which encodes 2 rRNA, 22 tRNAs and 13 polypeptides (Chen Jin-Qiang *et al.* 2009) **(Figure 1.3)**. Each polypeptide encoded by mitochondrial genome is a subunit of one of five respiratory complexes in the electron transport chain (ETC) localized in the inner membrane of mitochondria (Braun *et al.* 1992). The mitochondrial DNA is maternally inherited and harbours higher rates of mutation (Taylor and Turnbull 2005). The lack of introns in the mitochondrial genes and histones in packaging of mitochondrial

genome makes mitochondrial DNA more prone to mutation due to the presence of reactive oxygen species (ROS) generated by oxidative phosphorylation in the mitochondria (Kunkel and Loeb 1981, Matsukage et al. 1975, Modica-Napolitano and Singh 2004, Shay and Werbin 1992, Singh *et al.* 2001, Torri and Englund 1995). Mutation rate in mitochondrial DNA is tenfold higher than nuclear DNA (Shoubridge 2000, Wilson and Roof 1997). Mitochondrial DNA is an excellent tool for evolutionary study due to its small size and relatively conserved gene content and high mutation rate (Clark *et al.* 2007).

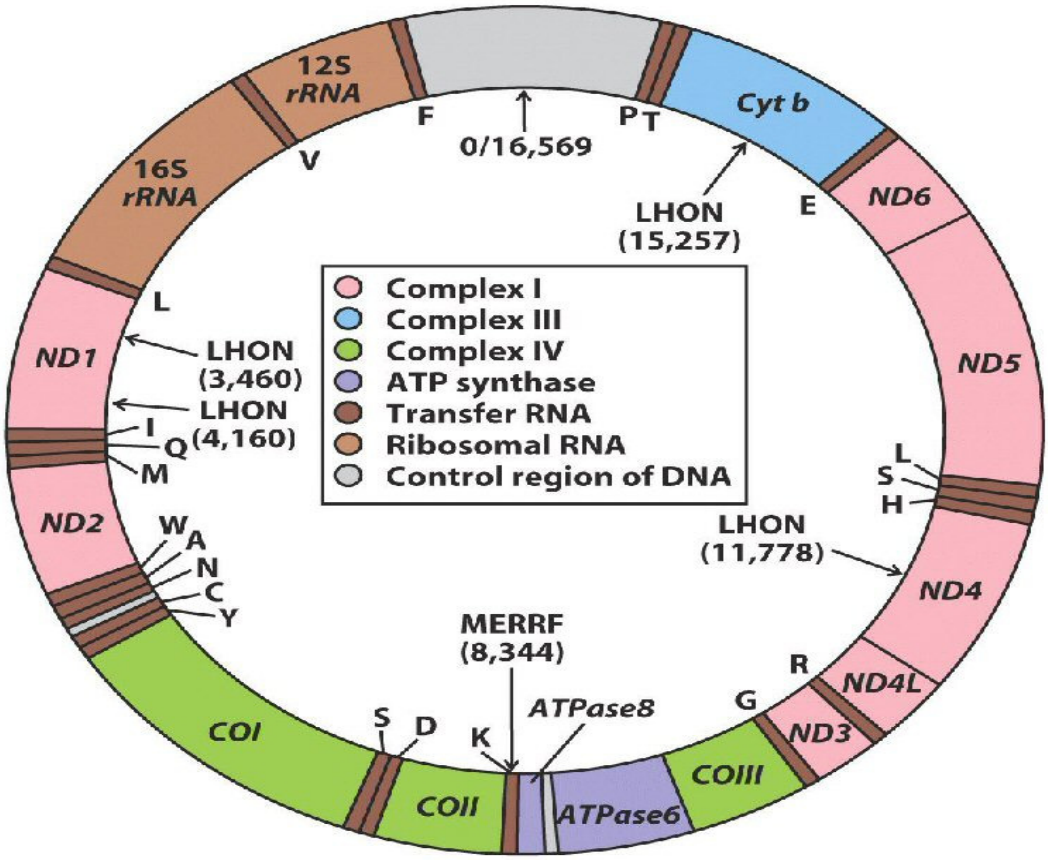


Figure 1.3 Structure of mitochondrial DNA

1.4.1 Mitochondrial genetic code

Unlike nuclear genetic code, mitochondrial genetic code in vertebrates is composed of 60 sense codons that represent 20 standard amino acids, and the remaining four codons that represent the termination signals are TAA, TAG, AGA and AGG (Knight *et al.* 2001) **(Figure 1.4)**. Six codons in animal mitochondria have been known as nonuniversal genetic triplets, which differed in the process of animal evolution. They are TGA (though it is a stop codon in the universal genetic code but encodes Trp in all animal mitochondria), ATA (encodes Ile to Met in most metazoan mitochondria), AAA (encodes Lys to Asn in

echinoderm and some platyhelminths mitochondria), AGA/AGG (encodes Arg to Ser in most invertebrate, Arg to Gly in tunicate, and Arg to stop codon in vertebrate mitochondria), and TAA (stop codon to Tyr in a planaria and a nematode mitochondria, but convincing proof is lacking in this case) (Watanabe and Yokobori 2011).

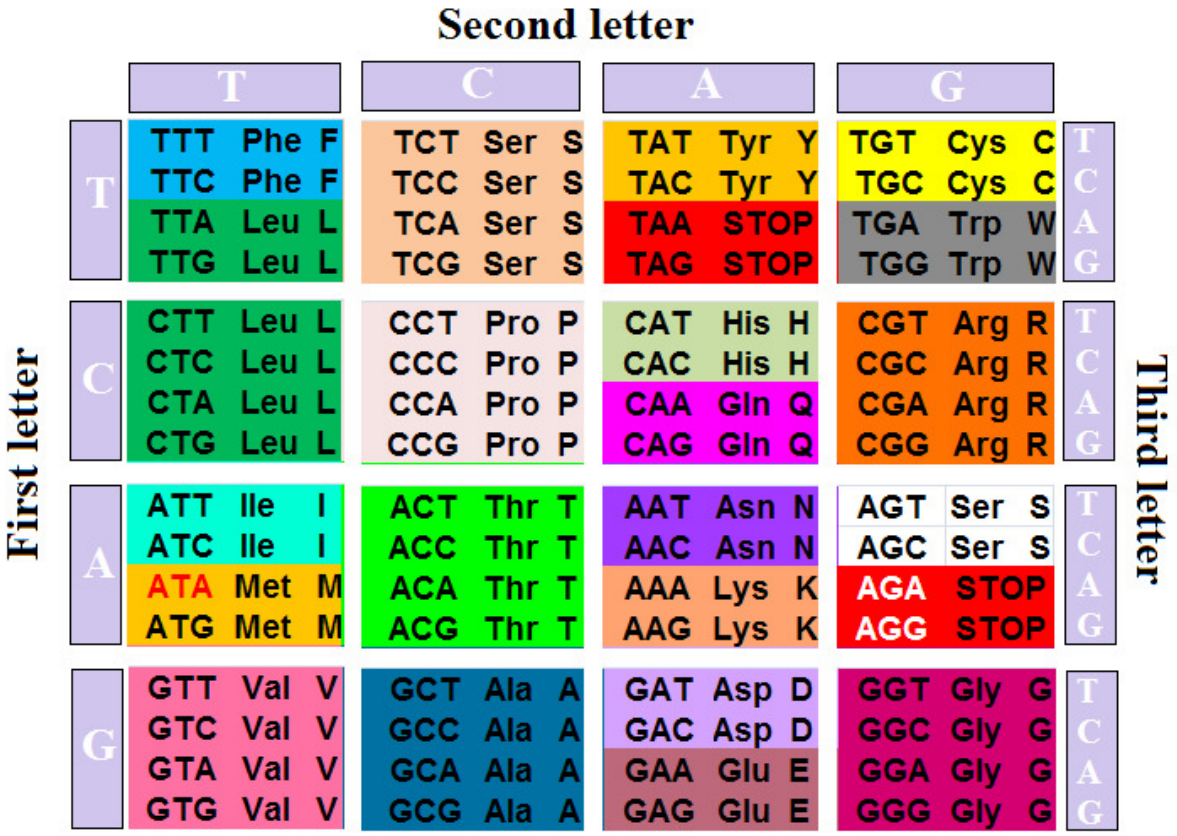


Figure 1.4 Mitochondrial genetic code

1.4.2 Evolution of Mitochondria

The origin of the eukaryotic cell is still a matter of continuing study and different theories have been proposed. But the origin of mitochondria from alpha protobacteria appears to be universally accepted. During the evolution of the organelle the original bacterial genome has become gradually smaller as most of its genetic material has been transferred to the nucleus. As we know, the evolution patterns of mitochondrial DNA and genomic DNA are not the same. At present, there are two main theories regarding the endosymbiotic origin of mitochondria. One of the theories suggests that the eukaryote engulfed the mitochondrion (Embley and Martin 2006), whereas the other hypothesizes that the prokaryote host acquired the mitochondrion. As somewhat independent organelle inside the cell, early studies used to believe that mitochondrial DNA acquired neutral evolution (Wise *et al.* 1998). However, later result has denied those findings, and it is now supposed that

mitochondrial DNA suffers from positive and negative selection (Rand and Kann 1998, Rand *et al*. 1994). A plenty of proof supports the existence of coevolution and co-adaption between the nuclear and mitochondrial genomes (Dowling *et al.* 2008, Gemmell *et al.* 2004, Rand *et al*. 2004). However, different substitution rates have been established between them, where the mitochondrial DNA shows a higher rate of nucleotide substitution than that of the nuclear DNA (Pesole *et al.* 1999).

1.4.3 Functional Role of Mitochondrial Genomes

The oxidative phosphorylation is the most important biochemical process located in mitochondria by which aerobic eukaryotic cells synthesize ATP with molecular oxygen as electron terminal acceptor. Mitochondrial DNA encodes for the three distinct classes of genes involved in this process: ribosomal RNA, transfer RNA, and protein- coding genes. Genes for small and large rRNA (12 S and 16 S in Mammals, for example) are found widely; in fact, genes for tRNA vary greatly in terms of numbers, although a set of 22–27 tRNAs is common in many eukaryotic groups.

Protein-coding genes are categorized into two groups *viz*: ribosomal protein and bioenergetic. The former are concerned with synthesis of ribosomal subunit and mostly occur in protist and plant mitochondrial genomes (Adams Keith L and Palmer 2003). The latter are universal and encode for the protein subunits of the respiratory chain (RC), the system of multienzymatic complex which generates the proton gradient essential for ATP synthesis (or heat generation). All respiring organisms always have a minimal set of bioenergetic genes (CYB and COI genes are mainly conserved), but, in all of them, the mitochondrial bioenergetic gene pool is inadequate to encode for all the RC subunits (Adams Keith L and Palmer 2003).

In Metazoa, mitochondrial gene content is quite stable. There are 37 mitochondrial genes encoding for two ribosomal RNAs, 22 for tRNAs, and 13 for the RC subunits. The complex I of RC contains 7 of the 13 mitochondrially encoded proteins (ND1, ND2, ND3, ND4, ND4L, ND5, ND6): ND2, ND4, and ND5 appear to work as electron transporters, while ND1 and ND2 play an important structural role between the membrane-embedded and peripheral arms of the complex (Fonseca *et al.* 2008). The complex II consists of entirely nuclear-encoded proteins. The CYB protein, having catalytic activity (cytochrome c reduction), is the only mitochondrially encoded subunit of Complex III. In the CO protein of complex IV, in which COI protein catalyzes electron transfer to the ultimate

acceptor, molecular oxygen; COII and COIII also belong to the catalytic core of the complex, in which nuclear subunits are mostly located externally. For ATP synthesis in complex V, ATP6 is a key component of the proton channel (FO component) and ATP8 seems to be a regulator of complex assembly (Fonseca *et al.* 2008). The nucleus regulates the genes for the production of all other RC proteins: about 39 for Complex I, 4 for Complex II, 10 for Complex III, 10 for Complex IV, and 15 for Complex V (Scarpulla 2008). Conversely, the role of the nucleus to mitochondrial functionality is not incomplete because 80-odd proteins (in mammals) are directly involved in oxidative phosphorylation. It has been reported that more than 1,500 genes regulate the varying aspects of mitochondrial activity (Wallace 2005), such as DNA replication and repair, gene expression and its modulation, complex assembly, etc.

1.4.4 Base composition in Mitochondria

Base composition is one of the major sources of variation of metazoan mitochondrial DNA (mtDNA) genomes. It is observed that the AT/GC content is rather variable between species and always lower in third codon positions. This is likely due to the fact that the depletion of G in the coding strand is always more marked at the level of the third codon positions of H-strand protein coding genes. The strong compositional bias, in particular the trend to avoid G at the third codon position (Brown *et al*. 1986, Perez-Reyes 1998), might also explain several deviations of the vertebrate mt code with respect to the universal one. Figure 1.4 shows genetic code deviation events that occurred during evolution of metazoa. At least ten independent genetic code changes might have occurred in the different lineages. Most ancient changes were UGA (Stop! Trp) and AUA (Ile! Met) that probably occurred before the divergence between yeast and metazoa. In vertebrate mitochondria, the most used initiation codon is AUA instead of AUG.

1.5. Chordates

Chordates (phylum Chordata) are deuterostome coelomates whose nearest relatives in the animal kingdom are the echinoderms, the only other deuterostomes. However, unlike echinoderms, chordates are characterized by a notochord, jointed appendages, and segmentation. Four features characterize the chordates and have played an important role in the evolution of the phylum:

- A single, hollow nerve cord runs just beneath the dorsal surface of the animal. In vertebrates, the dorsal nerve cord differentiates into the brain and spinal cord.
- A flexible rod, the notochord, forms on the dorsal side of the primitive gut in the early embryo and is present at some developmental stage in all chordates. The notochord is located just below the nerve cord. The notochord may persist throughout the life cycle of some chordates or be displaced during embryonic development, as in most vertebrates, by the vertebral column that forms around the nerve cord.
- Pharyngeal slits connect the pharynx, a muscular tube that links the mouth cavity and the oesophagus with the outside. In terrestrial vertebrates, the slits do not actually connect to the outside and are better termed pharyngeal pouches. Pharyngeal pouches are present in the embryos of all vertebrates. They become slits, open to the outside in animals with gills, but disappear in those lacking gills. The presence of these structures in all vertebrate embryos provides evidence of their aquatic ancestry.
- Chordates have a postanal tail that extends beyond the anus, at least during their embryonic development. Nearly all other animals have a terminal anus.

Phylum Chordate is divided into three Sub-Phyla *viz*: Sub-Phylum: 1. Urochordata 2. Cephalochordata 3. Vertebrata. The first two sub phyla are called lower chordates or protochordates. They are usually called Acrania group. The vertebrata sub-phylum is called Craniata.

1.5.1 Reason for selection of pisces, aves and mammals

The mitochondrial electron transport chain plays a vital role in fulfilling energy requirements of an organism. Analysis of codon usage patterns in protein-coding genes is of great interest to understand how the energy requirement of pisces, aves and mammals, influences the codon usage pattern against rapid environmental changes during the course of evolution. These three groups namely pisces, aves and mammals live in three different environments and so their mode of respiration and energy requirement are also different (Ellington 2001). Therefore, the study of synonymous codon usage helps in understanding the factors influencing gene evolution. Some previous studies on codon usage pattern of nuclear and mitochondrial genomes have been found in some invertebrates and vertebrates

(Karlin and Mrázek 1996, Wei *et al.* 2014). However, no work was reported for mitochondrial protein-coding genes among pisces, aves and mammals. Since all the protein-coding mitochondrial genes play a crucial role in the electron transport chain, the study of their codon usage will be interesting. The different species of pisces, aves and mammals are (**Figure 1.5 a-o**).

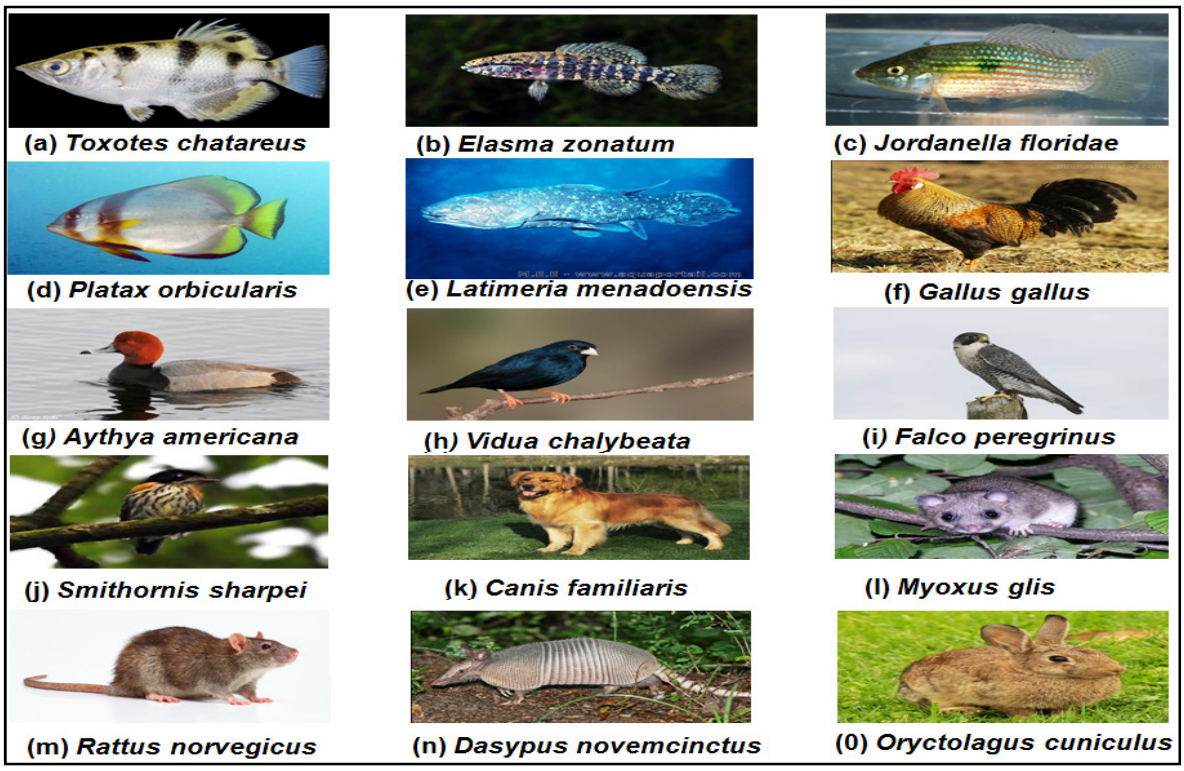


Figure 1.5. Different species of chordates under study. (a-e) pisces; (f-j) aves; (k-o) mammals. Courtesy from URL as follows :

(a) http://ffish.asia/?p=none&o=ss&id=623
(b) http://jonahsaquarium.com/JonahSite/picezonatum05.htm
(c) http://www.seriouslyfish.com/species/jordanella-floridae/
(d) http://fins.actwin.com/species/index.php?t=9&i=132
(e) http://www.aquaportail.com/fiche-poisson-3096-latimeria-menadoensis.html
(f) http://animaldiversity.org/accounts/Gallus_gallus/
(g) http://avise-birds.bio.uci.edu/anseriformes/anatidae/aythya_americana/index.html
(h) http://www.biodiversityexplorer.org/birds/viduidae/vidua_chalybeata.htm
(i) http://www.ontfin.com/Word/peregrine-falcon/
(j) http://www.taenos.com/en/itis/smithornis-sharpei-sharpei/Smithornis%20sharpei%20sharpei/
(k) http://www.allposters.com/-sp/Golden-Retriever-Canis-Familiaris-Illinois-USA-Posters_i2636278_.htm
(l) http://www.uniprot.org/taxonomy/41261
(m) http://www.controledepraga.com.br/?p=166
(n) http://www.flickriver.com/photos/jackhailman/5204323677/
(o) http://18o16o.deviantart.com/art/Rabbit-Oryctolagus-cuniculus-273005790

1.6 Statement of the problem

Codon usage is an important characteristic of a gene, which directly influences its expression and helps us understand the genetic and evolutionary relationship of an organism. Mitochondrial genome is a suitable tool to study the evolutionary relationship due to its small size, relatively conserved gene content, maternal inheritance pattern and high mutation rate. Herein, we investigated the mitochondrial protein-coding genes to understand the codon usage pattern among 15 species each of pisces, aves and mammals belonging to aquatic, aerial and terrestrial environments, respectively. Understanding the synonymous codon usage patterns in mitochondrial protein-coding genes among pisces, aves and mammals would improve our knowledge on the distribution of codons, their variation and certainly elucidate the factors influencing the codon usage pattern in these three groups.

1.6.1 Rationale of the study

- The evolutionary relationship among the chordates was not studied properly and requires much deeper and accurate molecular studies.
- Mitochondrial DNA is maternally inherited and the mutation rate is high.

1.6.2 Novelty of the Study

To the best of our knowledge, this is the first work on the analysis of codon usage bias and the comparison of mitochondrial protein-coding genes in pisces, aves and mammals as no work was reported yet. This study might help to study the evolution, molecular biology and also precursor study for heterologous gene expression and design of transgene. Further, this study would also shed light on the factors influencing the codon usage pattern of 13 mitochondrial protein-coding genes in pisces, aves and mammals.

1.7 Objectives

- To analyze the codon usage pattern of mitochondrial genes involved in respiratory process in different chordate species (Pisces, Aves and Mammals)
- To analyse the codon usage bias of the genes using different parameters (ENC, RSCU)
- To predict the level of gene expression using CAI
- To study the interrelationships of gene expression (CAI) with codon usage bias parameter (ENC)
- To analyse the compositional features of these mitochondrial genes and to study the interrelationships among different compositional features
- To compare the codon usage patterns of the mitochondrial genes among the chordates