5. Discussion

Codon usage bias analysis has attracted the attention of researchers worldwide due to over expanding whole genome sequencing projects followed by the record of data in publicly available nucleotide databases. Codon usage bias is the unequal use of synonymous codons, some codons are more preferred than others. Synonymous codons encoding amino acids range from two to six, usually differing in the third position (Ikemura 1985). However, the patterns of synonymous codon usage differ among the genes of an organism and among different organisms (Behura and Severson 2012, Sharp and Li 1987). Some codons that encode an amino acid are more frequently used than others by the genome. This bias is very common in highly expressed gene. Some workers reported that codon usage matches the organism tRNA pool (dos Reis *et al.* 2004). Codon bias may have important role in cellular processes and gene product (Baba *et al.* 2006). The codon frequencies vary significantly between different organisms, between proteins expressed at high or low levels within the same organism, and sometimes even within the same operon (Gustafsson *et al.* 2004). Factors that influence the codon bias include gene expression (Gustafsson *et al.* 2004), gene length (Duret and Mouchiroud 1999), base compositional mutational bias (Jenkins and Holmes 2003) and natural selection (Akashi 1994). As the synonymous codon usage during translation is non-uniform, identifying the codon usage pattern is essential for understanding the mode of translational selection of protein-coding genes among allied species.

The present investigation highlights the codon usage patterns in a comparative manner among pisces, aves and mammals. The different species of pisces, aves and mammals analyzed in this study are important because their modes of respiration and energy consumption are different, inhabiting aquatic, aerial and terrestrial environments respectively. Codon usage bias is an essential and complex evolutionary process, and it exists in a wide variety of diverse organisms, ranging from prokaryotes to eukaryotes. Among all the theories proposed to explain the origin of codon usage bias, neutral theory and selection-mutation drift balance model are most important. According to neutral theory, mutations at synonymous coding position should be selectively neutral, thus ensuing in non-uniform codon preference. In the selection-mutation-drift model, CUB is mainly determined by a balance between weak selection, mutation pressure and genetic drift. However**,** after the completion of the whole genome sequencing of many organisms, these two theories were not sufficient in elucidating the phenomenon of CUB (Yang *et al.*

2014). Several other factors have been proposed to influence CUB and these include GC-content (Hey and Kliman 2002), gene expression level (Duret and Mouchiroud 1999), gene length (Duret and Mouchiroud 1999), RNA structure (Hartl *et al.* 1994), protein structure (Orešič *et al.* 2003), the hydrophobicity and the aromaticity of the encoded proteins (Romero *et al.* 2000), recombination rate (KLIMAN and HEY 2003), population size (Berg 1996), evolutionary age of the genes (Prat *et al.* 2009) and environmental stress (Goodarzi *et al.* 2008) etc.

5.1 Codon usage pattern

To understand the pattern of non-uniform usage of synonymous codons in 13 mitochondrial protein-coding genes, relative synonymous codon usage (RSCU) of individual codons was compared between these genes. If RSCU value of a codon >1, that codon is frequently used than expected, when RSCU value <1, the codon is less frequently used than expected. If RSCU= 1, it means that the codon is used randomly and equally with other synonymous codons (Behura and Severson 2012). If the RSCU value is <0.6, the codon is under represented and if the RSCU value of a codon is >1.6, the codon is over-represented (Behura and Severson 2012).

In this study, we analyzed the relative synonymous codon usage in pisces, aves and mammals in 13 mitochondrial protein coding genes, as well as the relationship between codon usage patterns among these genes. When clustering these biases using RSCU values of codons in a heat map, these values displayed a remarkable difference between different species, belonging to pisces, aves and mammals each. Codon usage pattern differs in pisces, aves and mammals and it was also confirmed by relative synonymous codon usage as well as correspondence analysis. In ND1 gene, out of 60 codons, GCG, CTG, CTA, TCG, TGG, TCA, GTA, CCA, CAA, CCT encoding amino acids ala, leu, leu, ser, trp, ser, val, pro, gln, and pro respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ND2 gene, ACG, CGT, AGC, CCG, TCA, AAG, and GAT encoding amino acids thr, arg, ser, pro, ser, lys, and asp respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ND3 gene, GAT, CAC, CCG, CGC, TAT and TGG encoding amino acids asp, his, pro, arg, tyr and trp respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ND4 gene, GCA, ACA, ACG, TTT, AAG, ACT, TTA and TTG encoding amino acids ala, thr, thr, phe, lys,

thr, leu and leu respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ND4L gene, CTC, GCT, AAG, ACT, CTT, TTG, GTT, TAT, GCC and GTG encoding amino acids leu, ala, lys, thr, leu, leu, val, tyr, ala and val respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ND5 gene, CTA, TCA, TGT, TCC, CCG, GTG, CAG, CGG, ACG and GGG encoding amino acids leu, ser, cys, ser, pro, val, gln, arg, thr and gly respectively were used as over-represented codons in some species but under-presented codons in most of the species. For ND6 gene, the codon TAC, CCT, GGA, GAG, TTT, TAT and GAT encoding amino acids tyr, pro, gly, glu, phe, tyr and asp respectively were used as over-represented codons in some species but under-presented codons in most of the species. In COI gene, out of 60 codons, ACG, ATC, GTT, TTA and TGT encoding amino acids thr, ile, val, leu, and cys respectively were used as over-represented codons in some species but under-presented codons in most of the species. For COII gene, AGT, GCA, GGG, TCC, TCG, TGG, TTG, CAT, CGT and CGG encoding amino acids ser, ala, gly, ser, ser, trp, leu, his, arg, and arg respectively were used as over-represented codons in some species but under-presented codons in most of the species.. In COIII gene, AAT, AGC, TGT, GCG, CAG and CCG encoding amino acids asn, ser, cys, ala, gln and pro respectively were used as over-represented codons in some species but under-presented codons in most of the species. In CYB gene, out of 60 codons, AGT, CTT, GTA, AGT, CAG encoding amino acids ser, leu, val, ser, gln respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ATP6 gene, of these 60 codons, GAT, GGT, GTC, TGG encoding amino acids asp, gly, val, trp respectively were used as over-represented codons in some species but under-presented codons in most of the species. In ATP8 gene, codons such as GAT, TAT, TCT, and CTA encoding amino acids asp, tyr, ser and leu respectively were used as over-represented codons in some species but as under-represented codons in most of the species. These results suggest that compositional properties under mutation pressure played an important role in codon usage pattern in mitochondrial protein-coding genes in pisces, aves and mammals. The compositional constraint may not be the only factor associated with codon usage patterns, because even though the overall RSCU values could indicate the codon usage pattern for the genomes, it may hide the codon usage deviation among several genes in a genome (Hassan *et al.* 2010).

5.2 Codon usage bias

Effective number of codon (ENC) is a measure of codon usage bias. ENC depends upon nucleotide composition of the gene. Its value ranges from 20 to 61. If ENC value is 20 it means only one codon is used for each amino acid and if the value is 61, it means all the synonymous codons are equally likely to code for the same amino acid. Low ENC value means high codon usage bias (Wright 1990). If the ENC value of a gene (CDS) is less than 35, it reveals significant codon usage bias (Wright 1990).

Table 5.1 ENC values of 13 mitochondrial protein-coding genes among pisces, aves and mammals

Genes	ENC value in pisces	ENC value in aves	ENC value in mammals
ND1	57.40±3.20	59.80±0.44	56±2.64
ND2	57±2.91	59.80±0.44	55±1.58
ND3	57.8±2.48	59±1.41	53.2±3.63
ND4	58±2.82	59.8±0.44	54.4±3.20
ND4l	56.6±2.79	59.2±0.83	52.6±3.04
ND5	58.2±1.64	59.80±0.44	56.2±2.58
ND6	59.6±0.89	58.4±2.07	59.4±0.89
COI	58.40±1.81	59.80±0.44	53.80±3.49
COII	56.8±1.92	59.8±0.44	53.6±3.13
COIII	57.8±2.58	59.8±0.44	55.2±4.65
CYB	59.80±0.44	59.80±0.44	58.20±1.30
ATP6	57.8±2.16	59.80±0.44	53.20±4.20
ATP8	56.8±2.88	59.2±0.83	50.2±3.70

The average ENC value of the coding sequences of different species of pisces, aves and mammals for 13 mitochondrial protein-coding genes was high which indicates weak codon usage bias in 13 mitochondrial genes and that it is apparently maintained at a stable level (**Table 5.1**). Previous studies on codon usage bias analysis among the mitochondrial genes such as MT-ATP8 (54.1±5.93) in mammals (Uddin Arif and Chakraborty 2014) and MT-ND2 gene in pisces, aves and mammals (57±2.91, 59±0.44 and 55±1.58 respectively) (Uddin Arif et al.), also reported weak codon usage bias. The same is also true in the case of ENC value of albumin superfamily that varies from 51.65 to 56.62. It was also reported that ENC values of codon usage analysis of rabbit was 51.31±5.71. A probable explanation for the low codon usage bias is that it might be advantageous for efficient replication in each cell, with potentially distinct codon preferences (Jenkins and Holmes 2003).

5.3 Expression level

Expression level is measured by codon Adaptation Index (CAI). CAI values range from 0 to 1; with higher values indicating a higher proportion of the most abundant codons (Sharp and Li 1987). CAI is a measure of the relative adaptedness of the codon usage of a gene towards the codon usage of highly expressed genes. The relative adaptiveness (ω) of each codon is the ratio of the usage of each codon, to that of the most abundant codon within the same synonymous family. In our study, the CAI value in pisces, aves and mammals among 13 mitochondrial protein-coding genes as shown in **Table 5.2.**

Table 5.2 CAI values of 13 mitochondrial protein coding genes among pisces, aves and mammals

Genes	CAI value in pisces	CAI value in aves	CAI value in mammals
ND1	0.78±0.053	0.79±0.066	0.76±0.03.
ND2	0.78±0.05	0.76±0.05	0.76±0.02
ND3	0.44±0.11	0.30±0.056	0.46±0.014
ND4	0.85±0.05	0.81±0.02	0.84±0.015
ND4l	0.35±0.036	0.33±0.077	0.36±0.104
ND5	0.88±0.034	0.86±0.021	0.89±0.023
ND6	0.53±0.076	0.53±0.065	0.51±0.048
COI	0.87±0.02	0.84±0.03	0.86±0.03
COII	0.74±0.09	0.61±0.01	0.66±0.02
COIII	0.73±0.10	0.71±0.12	0.73±0.06
CYB	0.82±0.05	0.76±0.05	0.78±0.01
ATP6	0.6624±0.056	0.6807±0.065	0.6814±0.092
ATP8	0.2839±0.20	0.2882±0.087	0.3632±0.171

The CAI value in most of the protein-coding genes in mitochondria is high which indicates high expression level. The expression levels of ND1, ND2, ND4, ND5, COI, COII, CYB and ATP6 genes, were very high. The expression level of ND6 and COIII was moderately high. The expression level of ND3, ND4L and ATP8 genes was low. The CAI value in mitochondrial genes in *B. mori* ranges from 0.5-0.7 (Wei *et al.* 2014). A possible reason for high expression level of the mitochondrial protein coding genes might be due to the need that the respiratory chain fulfils the required high energy demand needed to adapt to the aquatic, aerial and terrestrial habitats (pisces, aves and mammals). The process is also related to the species specific abundant tRNA molecules (Ikemura 1985).

5.4 Interrelationships of gene expression with codon usage bias

We performed comparison between ENC and codon adaptation index (CAI) to determine the differences between nucleotide composition and codon selection in 13 mitochondrial genes among pisces, aves and mammals. CAI is a directional measure of codon usage bias parallel to relative codon bias score unlike ENC. Thus correlation between CAI and ENC provides a good qualitative measurement between the nucleotide composition and the codon bias selection (Vicario *et al.* 2007). We found no significant correlation between ENC and CAI which suggests that there was no relationship between codon usage bias and expression level in pisces, aves and mammals among 13 mitochondrial protein-coding genes. This indicates that codon usage bias is influenced by natural selection and not by translational selection. Similar result was also found by Karlin and Mrezek (Karlin and Mrázek 1996) in humans which suggests that the codon usage bias is influenced by natural selection in *H. sapiens*, *C. elegans*, *D. melanogaster* but not by translational selection (Blier *et al.* 2001, Karlin and Mrázek 1996, Stenico *et al.* 1994).

5.5 Compositional features

The composition of guanine and cytosine (GC content) plays an important role in codon usage bias. The GC content may influence the thermostability, bendability, and the ability to convert B form of DNA to Z form of DNA. GC content is also involved in active process of transcription because it has the ability to keep the coding region in an open chromatin state (Schwartz *et al.* 2009). In has been reported that highly expressed genes may have low mutation rates due to DNA repair meachanisms (Hoeijmakers 2001). In our study, the compositional properties such as overall nucleotide composition and its composition at 3rd codon position vary among 13 mitochondrial genes (**Table 5.3**).

Table 5.3 Comparison of compositional properties in 13 mitochondrial protein-coding genes

Genes	**Composition properties in pisces, aves and mammals**
ND1	In ND1 gene, the nucleobase C was the highest in pisces and aves but nucleobase A in mammals, whereas G was the lowest in pisces, aves and mammals respectively. The nucleobase A/C at the 3rd codon position was the highest in pisces and mammals but in aves C/A was found to be the highest, whereas G was the lowest in pisces, aves and mammals.
ND2	In ND2 gene, the nucleobase C/A was the highest in pisces and aves but nucleobase A/C was the highest in mammals, whereas G was the lowest in pisces, aves and mammals respectively. The nucleobase A/C at the 3rd codon position was the highest in pisces, aves and mammals, whereas G was the lowest in pisces, aves and mammals.
ND3	In ND3 gene, the nucleobase C/T was the highest in pisces but in aves, C/A was the highest while in mammals, the nucleobase A/T was the highest whereas G was the lowest in pisces, aves and mammals respectively. However, the analysis of nucleotide composition at the 3rd position of codons suggests that the nucleobase A/C at the 3rd codon position was the highest in pisces, aves and mammals, whereas G was the lowest in pisces, aves and mammals.
ND4	In ND4 gene, the nucleobase C/A was the highest in pisces and aves but nucleobase A/T in mammals, whereas G was the lowest in pisces, aves and mammals respectively. However, the analysis of nucleotide composition at the 3rd position of codons suggests that nucleobase A/C at the 3rd codon position was the highest in pisces and mammals but in aves C/A was found to be the highest, whereas G was the lowest in pisces, aves and mammals.
ND4l	In ND4L gene, the C/T was the highest in pisces while nucleobase C/A was the highest in aves but T/A was the highest in mammals. The nucleobase G was the lowest in pisces, aves and mammals respectively. However, the analysis of nucleotide composition of ND4L at the 3rd position of codons suggests that the nucleobase A/C at the 3rd codon position was the highest in pisces and mammals but in aves C/A was found to be the highest, whereas G was the lowest in pisces, aves and mammals.

Continued

Genes	**Composition properties in pisces, aves and mammals**
ND5	In ND5 gene, the A/C was the highest in pisces while nucleobase C/A was the highest in aves but A/T was the highest in mammals. The nucleobase G was the lowest in pisces, aves and mammals respectively. But, the analysis of nucleotide composition at the 3rd position of codons suggests that the nucleobase A/C at the 3rd codon position was the highest in pisces and mammals but in aves C/A was found to be the highest, whereas G was the lowest in pisces, aves and mammals.
ND6	In ND6 gene, the A/C was the highest in pisces, aves and mammals while G was the lowest in pisces, aves and mammals respectively. Further, the analysis of nucleotide composition at the 3rd position of codons suggests that the nucleobase C/A at the 3rd codon position was the highest in pisces, aves and mammals whereas T was the lowest in pisces, aves but nucleobase G was the lowest in mammals.
COI	In COI gene, the nucleobase T and C occurred more frequently than A and G in pisces, and in aves, C and A occurred more frequently than T and G while in mammals A and T occurred more frequently than C and G. The A3 was the highest, followed by C3 and T3 in pisces and in aves, C3 was the highest, followed by A3 and T3 while in mammals, A3 was the highest followed by T3 and C3.
COII	In COII gene, the nucleobase A was the highest, followed by C and T, and with G was the lowest in pisces. In aves, C was the highest, followed by A and T, but G was the lowest while in mammals, A was the highest, followed by T and C, with the G being the lowest. The nucleobase A3 was the highest, followed by C3 and T3 in pisces. The nucleobase C3 was the highest, followed by A3 and T3 in aves while in mammals, A3 was the highest followed by C3 and T3. The nucleobase G3 was the lowest in mammals followed by aves and pisces.
COIII	In COIII gene, in pisces, nucleobase C was the highest, followed by T and A, with G was the lowest. In aves, C was the highest, followed by A and T, but G was the lowest while in mammals T was the highest followed by A, and C, G being the lowest. However nucleobase at the 3rd codon position revealed that nucleobase A was the highest followed by C and T in pisces. In aves, nucleobase C was the highest followed by A and T while in mammals, A was the highest followed by C and T. The nucleobase G was the lowest in mammals followed by pisces, and aves**.**

Continued

Genes	**Composition properties in pisces, aves and mammals**
CYB	In CYB gene of pisces, nucleobase C was the highest, followed by T and A, with G was the lowest. In aves, C was the highest, followed by A and T, with G was the lowest. The nucleobase A was the highest, followed by T and C, and with G was the lowest in mammals. The C3 was the highest, followed by A3 and T3 in pisces and aves respectively. The A3 was the highest followed by C3 and T3 in mammals. The G3 % was the lowest in mammals followed by aves and pisces**.**
ATP6	In ATP6 gene of pisces**,** nucleobase C was the highest, followed by T and A, with G was the lowest. In aves, C was the highest, followed by A and T, but G was the lowest. The A was the highest, followed by T and C, and G was the lowest in mammals. The A3 was the highest, followed by C3, T3 and G3 in pisces. The C3 was the highest followed by A3, T3 and G3 in aves. The A3 was the highest followed by C3, T3 and G3 in mammals.
ATP8	In ATP 8 gene of pisces, nucleobases, A and C occurred more frequently whereas in aves C and A occurred more frequently than nucleobases T and G respectively. In mammals, A and T occurred more frequently than C and G. The nucleotide A/C occurred most frequently at the third codon position than T/G in pisces and aves respectively while in mammals, A/T occurred more frequently than C/A.

The GC content in general was lower than AT content in 13 mitochondrial protein-coding genes *i.e.* the genes are AT rich in all 13 mitochondrial genes among pisces, aves and mammals. We found GC content was high in aves followed by pisces and mammals for all 13 protein-coding genes. Mostly, the large difference in GC contents was found between 1st and 2nd codon position and between 1st and 3rd codon position.

Nucleotide composition could be one of the most important factors in shaping the codon usage in genes as well as genomes (Jenkins and Holmes 2003). Previous studies suggested that the distribution of the nucleobase C at the 3rd codon position was the most frequent followed by G, A and T, respectively *i.e.* GC content is higher than AT content in the genome analysis of rabbit (FADIEL 2003). Wei *et al.,* reported that the AT content was higher than GC content in *B.mori* supporting our result (Wei *et al.* 2014). The genomes of *Plasmodium falciparum* (Peixoto *et al.* 2004), *Tetraphalerus bruchi*, *Trachypachus holmbergi*, *Sphaerius* sp., *Chaetosoma scaritides* , *Cyphon* sp., and *Priasilpha obscura* (Sheffield *et al.* 2008) are found to be rich in AT nucleobases. GC content was also

reported to be the lowest in mitochondrial genomes of eight nemertean species such *as Cephalothrix hongkongiensis, Cephalothrix* sp., *Lineus alborostratus, Lineus viridis, Zygeupolia rubens, Emplectonema gracile, Nectonemertes cf. mirabilis,* and *Paranemertes cf. peregrine* (Chen Haixia *et al.* 2014).

Most synonymous codons differ only at the third codon position– GC3 (guanine and cytosine at the 3rd position) and it is a good indicator of the degree of synonymous CUB (Shen *et al.* 2015). Earlier studies have revealed that genes with higher GC3 content tend to get methylated which ultimately leads to mutation compared to those with low GC3 content (Tatarinova *et al.* 2010). It was reported that GC3 acts as an isochore marker and the association between GC3 and the GC content of the flanking regions is still doubtful (Aota and Ikemura 1986). Duan *et al.* (Duan *et al.* 2015) reported that GC1 and GC2 were much more conserved than GC3 in the wobble position in 23 species of vertebrates consisting of eight mammals, four birds, one reptile, one amphibian and eight fishes (including *M. amblycephala*). In their work, GC3 was higher than GC1 in fishes and mammals, but was lower in amphibian, reptile and birds except *Taeniopygia guttata*.

5.6 Interrelationships among different compositional features

Two major evolutionary forces namely mutation pressure and natural selection are considered to shape the codon usage pattern in a species. We performed correlation analysis between general nucleotide composition and nucleotide composition at 3rd codon position to determine whether evolutionary process is driven by mutation pressure alone or by both mutation pressure and natural selection.

Highly significant correlation in some of the compositional constraint in pisces, aves and mammals among 13 protein-coding genes suggest that both natural selection and mutation pressure influenced the codon usage pattern in these genes. Furthermore, significant correlation between ENC and various GC contents, suggests that the nucleotide composition under mutation pressure and natural selection affect the synonymous codon usage in mitochondrial protein-coding genes of pisces, aves and mammals. Apart from the compositional constraint, the codon usage bias is mainly influenced by mutation pressure and natural selection.

Previous studies also reported similar result in ND2 gene (Uddin A. *et al*. 2015) and mitochondrial DNA in *B.mori* (Wei *et al.* 2014). Mutational biases are caused due to certain types of mutations, such as non-uniform DNA repair, non-random replication

errors and chemical decay of nucleotide bases (Kaufmann and Paules 1996). Mutational biases are neutral, which do not affect the protein properties and typically act on all DNA sequences of an organism. Several mutations initiate from non-random mismatch repairs following replication errors and methylation. Such strand-specific mutational biases result from differential fidelities of replication of the leading and lagging strands. Such asymmetric mutation rates of the leading and lagging strands are found in both bacteria (Lobry 1996) and eukaryotes (Pavlov and Anrep 2003). Global species differences in codon usage are typically explained by mutational biases. Earlier work on codon usage bias was carried out in vertebrates such as *Xenopus* (Romero *et al.* 2003) and *Gallus* (Rao *et al.* 2011), as well as in mammals (Chamary *et al.* 2006). However, there are a huge number of nonchordate species with big effective population sizes that could enable natural selection to act effectively on the synonymous codon usage pattern (Kober and Pogson 2013). This probably could affect the rate of translation due to matching of transfer RNA abundance and codon usage as reported by Zuckerkandl and Pauling (Zuckerkandl and Pauling 1965). It was observed that, in *Escherichia coli*, the rate and accuracy of translation are affected by the use or disuse of "major" codons (Tuller *et al.* 2010). It was also observed in synonymous codon usage in *D. melanogaster* and *C. Elegans*, a bias towards a set of preferred codons corresponds to the most abundant tRNAs in the cell and the number of tRNA gene copies in the genome (Moriyama and Powell 1998, Sharp *et al.* 1988). Protein translation, a biological process, is prone to errors. The impact of translation error on the evolution of CUB depends on its effect on protein function as well as its frequency (Shah and Gilchrist 2010). Translation errors include both missense errors and nonsense errors. Nonsense errors lead to premature termination of a growing polypeptide chain but missense errors insert wrong amino acids in the growing peptide chain. Many researchers believe that selection against missense errors results in codon usage bias so translation accuracy is maintained (Akashi 1994, 2001, Arava *et al.* 2005, Drummond and Wilke 2009, Stoletzki and Eyre-Walker 2007). Several studies revealed that the synonymous codons encoding an amino acids differ in translation error proneness. The translation error rate of a codon depends, in part, on the relative abundances of its cognate and near cognate tRNAs (Kramer and Farabaugh 2007).

Neutrality plot, a graphical plot of GC12 against GC3, depicts the roles of directional mutational pressure against natural selection. In this plot, regression coefficient of GC12 on GC3 is the equilibrium condition of mutation-selection pressure (Sueoka 1988). Mutations that mostly occur in the 3rd position of codon result in synonymous mutation,

whereas mutations that occur in 2nd or 3rd position lead to non synonymous change. Non synonymous mutations occur less frequently due to gene function. Theoretically mutation should occur randomly if there is no external pressure. The preference of bases at three different codon positions is not same under the influence of selection pressure (Sueoka 1988).

In our study we found that the regression coefficient between GC12 and GC3 was <0.5 in all 13 mitochondrial protein-coding genes except ND6 which suggest that natural selection played a major role while mutation pressure played a minor role in shaping codon usage pattern in 12 mitochondrial genes. However, in ND6 gene, mutation pressure played a major role while natural selection played a minor role in codon usage pattern. Wei *et.al* also found that natural selection played a major role and mutation pressure played a minor role in codon usage bias in mitochondrial DNA in *B.mori* (Wei *et al.* 2014). Chen also reported similar results in codon usage bias pattern in DNA and RNA virus genomes (Chen Youhua 2013)

5.7 Correspondence analysis

Correspondence analysis is a multivariate statistical method which is used to study the major trends in codon usage variation in nucleic acid sequence and to distribute the codons in axis1 and axis2 with these trends (Shields and Sharp 1987). Each CDS is represented as 60 dimensional vectors, each vector corresponding to RSCU value of each codon for 13 mitochondrial protein-coding genes among pisces, aves and mammals together. The major trends in codon usage variation can be determined with relative inertia, according to which the genes are located to investigate the major factors affecting codon usage pattern. The correspondence analysis suggests that the pattern of codon usage differs among 13 protein-coding genes and among pisces, aves and mammals (**Table 5.4**).

Further, for 13 protein-coding genes among pisces, aves and mammals, the positions of most codons are more close to axes with a concentrate distribution, indicating that the base composition for mutation bias might correlate to the codon usage bias. A few codons are in a discrete distribution, indicating that there are many other factors that exist, for example, natural selection that influenced the codon usage pattern of 13 mitochondrial genes among pisces, aves and mammals.

Table 5.4 F1 and F2 of COA among pisces, aves and mammals

Genes	F1% and F2% of COA in pisces	F1% and F2% of COA in aves	F1% and F2% of COA in mammals
ND1	40.62 and 35.16	37.55 and 25.49	38.40 and 30.65
ND2	49.38 23.07	36.80 and 32.88	37.01 and 26.55
ND3	34.63 and 26.24	36 and 28.09	37.77 and 31.80
ND4	43.66 and 31.44	52.44 and 19.33	39.17 and 27.37
ND4l	33.59 and 27.36	35.91 and 30.13	38.12 and 26.64
ND5	42.98 and 38.85	40.51 and 26.15	32.84 and 30.86
ND6	36.09 and 30.70	33.14 and 25.89	36.83 and 29.04
COI	78.74 and 11.52	72.15 and 14.57	40.19 and 34.11% .
COII	37.31 and 32.24	40.97 and 32.52	35.53 and 26.76% .
COIII	47.79 and 21.22	67.88 and 14.52	42.07 and 29.93
CYB	38.48 and 28.67	39.66 and 26.19	38.79 and 24.00
ATP6	45.22 and 26.89	36.73 and 28.84.	36.73 and 28.84
ATP8	38.28 and 24.37	38.28 and 24.37	38.28 and 24.37

Wei et *al.* studied the codon usage bias of mitochondrial DNA in *B.mori* and they found axis 1 contributed 12.07% and axis 2 contributed 8.64% of the total variation. In their work, the positions of the AT ended codons were more close to axis1 than the GC ending codons with a concentrate distribution suggesting that the compositional properties for mutation bias might correlate to the codon bias. Although the genes with different GC content showed a somewhat regular distribution, and the genes with lower GC content located more close to axis1, this indicated that GC content for mutation pressure perhaps influenced the codon usage bias. Further, a considerable number of the genes was in a discrete distribution, suggesting that natural selection might also influence the codon usage bias (Wei *et al.* 2014). Jia *et al* reported that first axis (F1) contributed 24.51% of the total variation and the second axis contributed 7.46%, of the total variation for nuclear genes in *B.mori* (Jia *et al.* 2015).

5.8 Amino acid composition

The non uniform usage of synonymous codons has the potential to tilt the relative abundance of various amino acids in proteins. This is because the base composition of codons differs in GC content (Foster and Hickey 1999). On the other hand, selection may

distort the frequencies of amino acids because amino acids with similar function may have different tRNA abundances or need different metabolic costs for their production (Barrai *et al.* 1995). In a number of species, the base composition has been shown to correlate with the amino acid content of proteins (Akashi and Gojobori 2002, Singer and Hickey 2000). The highly and the lowly expressed proteins can have different amino acid usage (Akashi and Gojobori 2002). However, gene function may stun the interpretation of variations in amino acid usages of the encoded proteins. For example, highly abundant proteins might have similar functions, so amino acid usage similarity among them could merely reflect their common peptide domains rather than selection for efficient and/or accurate translation (Cutter *et al.* 2006).

In our study, the pattern of amino acid usage also differs in 13 protein-coding genes among pisces, aves and mammals. In ND1 protein, the usage of leucine residue was the highest in the amino acid composition of pisces, aves and mammals while cysteine, histidine, aspartate residues were lower in the proteins. In ND2 protein, the usage of leucine in different species of pisces, aves and mammals was the highest. The usages of tyrosine, cysteine, histidine, glutamine, arginine, asparagine, valine, aspartate and glutamate were lower in the ND2 proteins. In ND3 protein, the usage of amino acid leucine was the highest and the usages of amino acids such as arginine, asparagine, aspartate, cysteine, glutamine, histidine, lysine, methionine, tyrosine and valine were low. In ND4 protein, the usage of amino acid leucine was the highest while amino acids such as arginine, aspartate, cysteine, glutamine, glutamate, histidine, lysine, tyrosine and valine were lower in ND4 proteins. In ND4L protein, the usage of amino acid leucine was the highest while arginine, asparagine, asparatate, glutamine, histidine, isoleucine, lysine, proline, theonine, tryptophan, tyrosine and valine were lower in usage in the amino acid sequence of ND4L protein. In ND5 protein, the usage of amino acid leucine was the highest while arginine, aspartate, cysteine, glutamate, histidine and valine were lower in the amino acid composition of protein. In ND6 protein, the amino acids such as asparagine, glutamine, histidine, lysine and proline were higher while alanine, aspartate, cysteine, glutamate, glycine, phenylalanine, tryptophan, tyrosine and valine were lower in usage in the amino acid sequence of ND6 protein. The usage of leucine residue was the highest in the amino acid composition of ND6 protein in pisces, aves and mammals while cysteine residue was the least**.** The usage of amino acids such as glutamine, arginine, lysine, glutamate and tryptophan were lower in COI protein. In COII protein, the usage of leucine residue was

the highest in all species of pisces, aves and mammals while cysteine residue was the least in COII. The usage of amino acids such as asparagine and lysine were lower in COII protein. In COIII protein, the amino acid usage in different species of pisces aves and mammals was estimated. The frequency of leucine residue was the highest in the amino acid composition of most of the species of pisces, aves and mammals while cysteine, arginine, asparagine, lysine, aspartate residue were the lower in the proteins**.** The usage of amino acids such as serine was the highest in only *S. sharpei* for COIII protein. In CYB protein, the frequency of leucine was the highest in pisces, aves and mammals while cysteine residue was the least. The other amino acids such as glutamine, arginine, aspartate, glutamate were low in usage in the amino acid sequence of CYB. In ATP6 protein, the usage of leucine residue was the highest in pisces, aves and mammals while tyrosine, cysteine, histidine, lysine, aspartate and glutamate residues were lower in the ATP6 proteins. In ATP8 protein, the usage of leucine, proline and theonine residues were higher among other amino acids in pisces, aves and mammals while tyrosine, cysteine, histidine, arginine, aspartate and glycine residues were lower.

Cutter *et al* (2006) reported that differences in codon usage for several amino acids may reflect an effect of phylogeny. In their study, all Meloidogyne species and most Spiruromorph nematodes (including *Brugia malayi*) use the leucine TTG as an optimal codon, whereas their nearest outgroup species do not use TTG as optimal codon. The GC poor genomes preferentially use isoleucine codon ATT and threonine codon ACT, unlike their nearest relatives with higher GC content. Optimal codon changes among species for alanine and threonine demonstrate the potential for both phylogeny and base composition to affect the loss, gain, and switching of optimal codon identities, even though the long phylogenetic timescale and predominance of parasitic species in their data set make any inference of ancestral states preliminary (Cutter *et al.* 2006).

Hierarchial clustering of 13 protein-coding genes among pisces, aves and mammals using RSCU values suggests that the pattern of codon usage differs among 13 protein-coding genes and among pisces, aves and mammals. In our study we found codon usage bias was low in 13 protein-coding genes. The expression level in ND1, ND2, ND4, ND5, COI, COII, CYB, and ATP6 genes was very high. The expression level of ND6, COIII was moderately high. The expression level of ND3, ND4l and ATP8 was low. Correlation between ENC and CAI suggests that no significant relationship exists between codon usage bias and gene expression level. We found GC content was high in aves followed by

pisces and mammals. Correlations among compositional constraints and correlation between ENC and various GC contents suggest that both mutation pressure and natural selection affect the codon usage pattern in 13 protein coding genes in mitochondria among pisces, aves and mammals. Neutrality plot suggests that natural selection played a major role while mutation pressure played a minor role in shaping the codon usage pattern in mitochondrial protein-coding genes except ND6 gene in pisces where mutation pressure played the dominant role while natural selection played a minor role in codon usage pattern. Correspondence analysis also suggests that the patterns of codon usage are different among genes and vary among species. The study also revealed that both natural selection and mutation pressure affect the codon usage pattern in 13 protein-coding genes of mitochondria in pisces, aves and mammals.