CHAPTER-3

MATERIALS AND METHODS

$$CAI = \exp\frac{1}{L}\sum_{k=1}^{L}\ln w_{c(k)}$$

$$ENC^{expected} = 2 + s + \frac{29}{s^2 + (1-s^2)}$$

$$AT_{skew} = \frac{A-T}{A+T}$$

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} g_{ij}} n_i$$

$$FOP_s(g) = \frac{1}{N}\sum_{i} n_i(g)$$

$$P_{xy} = \frac{f_{xy}}{f_y f_x}$$

$$GC_{skew} = \frac{G-C}{G+C}$$

$$AT_{skew} = \frac{A-T}{A+T}$$

3. Materials and Methodology

3.1 Availability of sequence data

The coding sequences (CDS) of the 13 mitochondrial protein-coding genes from five species of pisces, aves and mammals each were retrieved from National Center for Biotechnology Information (NCBI), USA (http://www.ncbi.nlm.nih.gov/). The accession numbers and name of different species are shown in Table 3.1.

Table 3.1 Accession numbers of coding sequences of 15 species with their respective family

Species	Family	Accession No
Toxotes chatareus	Toxotidae	AP006806
Elasma zonatum	Elassomatidae	AP006813
Jordanella floridae	Cyprinodontidae	AP006778
Platax orbicularis	*Ephippidae*	AP006825
Latimeria menadoensis	*Latimeriidae*	AP006858
Gallus gallus	*Phasianidae*	X52392
Aythya americana	*Anatidae*	AF090337
Vidua chalybeata	Viduidae	AF090341
Falco peregrinus	Falconidae	AF090338
Smithornis sharpei	Calyptomenidae	AF090340
Canis familiaris	*Canidae*	U96639
Myoxus glis	Myoxidae	AJ001562
Rattus norvegicus	*Muridae*	X148148
Dasypus novemcinctus	*Dasypodidae*	Y11832
Oryctolagus cuniculus	Leporidae	AJ001588

3.2 Compositional constraints

The compositional properties of the mitochondrial 13 protein-coding gene sequences (CDS) were calculated for the 5 different species of pisces, aves and mammals namely (i) general nucleotide composition (A, C, T and G %) and nucleotide composition in its 3rd codon position. (ii) the frequency of the occurrence of entire GC contents of CDS and GC

contents at the first (GC1), second (GC2) and third position (GC3). The analysis is based on vertebrate mitochondrial genetic code which has four termination codons (excluded from analysis) but met, trp are encoded by two codons each (met and trp included in analysis) (Perna and Kocher 1995).

3.3 Indices of synonymous codon usage bias

Some of the most relevant and widely used measures of codon usage bias analyzed in this study are discussed below.

3.3.1 Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) was calculated as the ratio of the observed frequency of a codon to the expected frequency if all the synonymous codons of a particular amino acid are used equally. If RSCU value of a codon >1 means the codon is frequently used than expected, RSCU<1 means the codon is less frequently used than expected. If RSCU= 1 it means the codon is used randomly and equally (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, it is treated as over-represented. RSCU was calculated using the formula.

$$RSCU = \frac{g_{ij}}{\sum_{j}^{ni} gij} ni$$

where, g_{ij} is the frequency of occurrence of the i^{th} codon for the j^{th} amino acid (any g_{ij} with a value of zero is arbitrarily assigned a value of 0.5) and n_i is the kind of synonymous codon.

3.3.2 Effective number of codons (ENC)

The ENC is the most commonly used parameter to measure the usage bias of synonymous codons (Wright 1990). ENC value of a gene is used to quantify the codon usage bias. The ENC value ranges from 20 (when only one codon is used for each amino acid) to 61 (when all codons are used randomly). If the calculated ENC is greater than 61 (because codon usage is more evenly distributed than expected), it is adjusted to 61. Higher ENC value

means low codon usage bias. ENC value < 35 is generally considered as the significant codon usage bias (Wright 1990). It is calculated as:

$$ENC^{expected} = 2 + s + \frac{29}{s^2 + (1 - s^2)}$$

where, *s* denotes the given GC_3% values.

3.3.3 Codon adaptation index (CAI)

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. The CAI is calculated as

$$CAI = \exp\left(\frac{1}{L} \sum_{k=1}^{L} \ln \omega k \right)$$

where, ωk is the relative adaptiveness of the kth codon and L is the number of synonymous codons in the gene.

3.4 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 distributes 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 of 60 codons for the mitochondrial protein coding genes each of pisces, aves and mammals. The major trends in codon usage variation can be determined with relative inertia, according to which the coding sequences are analyzed to investigate the major factors affecting codon usage pattern. COA was done using XLSTAT Pro software.

3.5 Neutrality plot

Mutations that mostly occur in the 3rd position of codon results in synonymous mutation, whereas mutation that occurs in 2nd or 3rd position leads 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 in three

different codon positions is not same in the presence of selection pressure (Sueoka 1988). 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).

3.6 Hierarchal Clustering

The RSCU values of codons from different species of pisces, aves and mammals were clustered by hierarchial clustering methods using XLSTAT.

3.7 Software used

A novel software developed by us using Perl script was used to calculate all the CUB parameters and nucleotide compositions (UDDIN *et al.*). Correlation analysis was used to identify the relationship between overall nucleotide composition and each base at 3rd codon position. All the statistical analyses were done using the SPSS software.