


In Silico Analyses of Burial Codon Bias Among the Species of Dipterocarpaceae Through Molecular and Phylogenetic Data

Raju Biswas¹, Anindya Sundar Panja² and Rajib Bandopadhyay¹ 

¹UGC-Center of Advanced Study, Department of Botany, The University of Burdwan, Bardhaman, India. ²Department of Biotechnology, Oriental Institute of Science and Technology, Vidyasagar University, Midnapore, India.

Evolutionary Bioinformatics
Volume 15: 1–12
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176934319834888



ABSTRACT

INTRODUCTION: DNA barcode, a molecular marker, is used to distinguish among the closely related species, and it can be applied across a broad range of taxa to understand ecology and evolution. MaturaseK gene (*matK*) and rubisco bisphosphate carboxylase/oxygenase form I gene (*rbcL*) of the chloroplast are highly conserved in a plant system, which are used as core barcode. This present endeavor entails the comprehensive examination of the under threat plant species based on success of discrimination on DNA barcode under selection pressure.

RESULT: The family Dipterocarpaceae comprising of 15 genera is under threat due to some factors, namely, deforestation, habitat alteration, poor seed, pollen dispersal, etc. Species of this family was grouped into 6 clusters for *matK* and 5 clusters and 2 sub-clusters for *rbcL* in the phylogenetic tree by using neighbor-joining method. Cluster I to cluster VI of *matK* and cluster I to cluster V of *rbcL* genes were analyzed by various codon and substitution bias tools. Mutational pressure guided the codon bias which was favored by the avoidance of higher GC content and significant negative correlation between GC12 and GC3 (in sub-cluster I of cluster I [$0.03 < P$], cluster I [$0.00001 < P$], and cluster II [$0.01 < P$] of *rbcL*, and cluster IV [$0.013 < P$] of *matK*). After refining the results, it could be speculated that the lower null expectation values ($R=0.5$ or <0.5) were less divergent from the evolutionary perspective. Apart from that, the higher null expectation values ($R > 0.85$) also showed the same result, which possibly could be due to the negative impact of very high and low transition rate than transversion.

CONCLUSION: Through the analysis of inter-generic, inter/intra-specific variation and phylogenetic data, it was found that both selection and mutation played an important role in synonymous codon choice in these genes, but they acted inconsistently on the genes, both *matK* and *rbcL*. In vitro stable proteins of both *matK* and *rbcL* were selected through natural selection rather than mutational selection. *matK* gene had higher individual discrimination and barcode success compared with *rbcL*. These discriminatory approaches may describe the problem related to the extinction of plant species. Hence, it becomes very imperative to identify and detect the under threat plant species in advance.

KEYWORDS: codon bias, Dipterocarpaceae, DNA barcode, MaturaseK gene (*matK*), phylogeny, rubisco bisphosphate carboxylase/oxygenase form I gene (*rbcL*), transition/transversion

RECEIVED: January 24, 2019. ACCEPTED: February 7, 2019.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Rajib Bandopadhyay, UGC-Center of Advanced Study, Department of Botany, The University of Burdwan, Golapbag, Bardhaman 713104, West Bengal, India. Email: rajibindia@gmail.com

Introduction

Phylogenetic analysis is the big deal in biology, because it provides basic information of background of an organism especially about the status and modes of their existence. The phylogenetic analysis of Dipterocarpaceae has yet not been extensively studied. A very few major phylogenetic analysis on this family has been reported based on DNA barcode, *rbcL* and *matK* gene; *rbcL*, *trnH-psbA*, and *matK* from Ketambe Research Station Quang Ninh (Vietnam) and Guangxi (China) in *Hopea chinensis* (Dipterocarpaceae), respectively. It was found that 2 of the individual DNA barcodes, *matK* and *rbcL*, performed best.^{1,2} DNA barcoding has been successfully initiated in animals by using mitochondrial gene, cytochrome oxidase 1 (CO1), having more than 95% accuracy level of species identification of major animal clade.³ But it is found challenging to

identify a standard barcode system in the plants, because of low mutation rate in CO1 and rapid structural changes of mitochondrial genome.^{4–6} In plant systematic research, both *matK* and *rbcL* genes from the chloroplast genome (cpDNA) appear to be a valuable gene by providing a high phylogenetic signal and a fairly conservative level of evolution, respectively.^{7,8} The genes *matK* and *rbcL* encode the maturase enzyme subunit K and a large subunit of ribulose 1,5-bisphosphate carboxylase (Rubisco), respectively. The plant-working group, Consortium for the Barcode of Life (CBOL), recognizes 2-plastid barcoding region *rbcL* + *matK* called as core barcode and an additional marker as required.⁹ *matK* is one of the most rapidly evolving coding regions of the plastid genome having high species discrimination power, and it seems to be closely analogous to the animal barcode.¹⁰ No other 2-markers or multi-markers



combination of plastid barcode (*rpoC1* + *rpoB* + *matK* or *rpoC1* + *matK* + *trnH-psbA*¹¹; *rbcL* + *trnH-psbA*¹²) and (*atpF-H* + *psbK-I*³) except 2-plastid markers (*rbcL* + *matK*) provide appreciably better species resolution. Coding nature of both these (*rbcL* + *matK*) regions is not only useful for automatic checking of editing/assembly error and ascertaining of the existence of pseudogenes and proper sequence alignment, but it also facilitates the analysis pertaining to the comparative diversity among the taxa and evolutionary divergence based on substitution bias. Apart from *matK* and *rbcL* genes, there are other different barcodes (subunit of plant RNA polymerase [*rpoA*, *rpoB*, *rpoC1*, *rpoC2*], intergenic spacer region [*trnH-psbA*], ATP synthase subunit gene's spacer [*atpF-H*]) available for phylogenetic analysis.

The parameter for plant barcodes success

First, geographical constraints generally make a high level of distinctive species discrimination.^{14,15} In contrast, the species diversity decreases as one moves toward dense populations which lead to shared barcodes among the coexisting species.^{16,17} Second, sufficient time is required for speciation driven by mutation or drift to form a set of genetic constituent which isolates conspecific individual together and separate them from other species. Barcode sequence represents the deficiency of proper species discrimination, due to slow rate mutation (*Araucaria*), woody species with long-generation time, and also individuals radiated recently and rapidly in *Inga*.¹⁷ Third, polyploid speciation shows inconsistency between barcode sequences and taxon concept¹⁸ where hybridization and/or polyploids are frequent. Both multiple allopolyploids and independent origin allopolyploid species having an identical plastid sequence share a common ancestor. According to the evolutionary clock, autopolyploid species does not share their plastid haplotype with diploid progenitor over evolutionary timescales unlike the initial stage of their origin. Finally, those with limited seed dispersal are predicted to have less DNA barcode discrimination success and it takes more time to reach monophyly in barcode regions, which is a natural occurrence than the connected populations having regular gene flows.¹⁹

In the study of molecular evolution of individual genes, it is important to know the synonymous codon usages. Synonymous codon usages are not randomly used^{20,21} which have been influenced by factors such as CpG islands,²² gene length,²³ gene expression,²⁴ protein secondary structure,²⁵ gene density,^{26,27} and so on. Two important models, ie, both mutational bias and natural selection, determine independently the codon usages variations. It is necessary to consider more codon usage patterns because no such unified theory for codon usages has been established. The chloroplast genome is found to be the most effective in the study of plant molecular evolution due to its small size, simple structure, and high copy number, which is closely similar to a bacterial genome. Recently, many more

chloroplast genomes have been sequenced with the help of advanced DNA sequencing techniques (<http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=Plastid#pageTop>). The codon usage pattern of chloroplast genes (like *matK*, *rbcL*) shows very close similarity to those in *Escherichia coli*.²⁸ It has been reported that the codon bias in the chloroplast genome of *Euglena gracilis* is found to be determined by 2 asymmetric DNA strands.²⁹ The choice of codon usage in the chloroplast genome of grass species is influenced by context dependent mutation,³⁰ but in some cases, it is found that selection may influence the codon usage pattern of certain chloroplast genes.³¹

Transition rather than transversion, which is favored by natural selection, causes the biochemical advantage.³²⁻³⁴ The introductions of new alleles through the transition are several folds higher than that of transversion, and therefore, nucleotide transition is common in molecular evolution.³⁵ This pattern of amino acid replacement often supports the effect of selection on the ground that the transition is more conserved in their effect on protein supported by reviewing more than 8 published reports.³⁶ The selective hypothesis has proposed that conservative effect on biochemical factors by transition mutation over transversion is correlated to the pattern of evolutionary divergence.³⁷ One obvious question is that the changes fixed in the evolution of an organism are favored toward survival effect because natural selection encourages positive adaptive changes whether it happens through transitions or transversion. Every nucleotide side (eg, G) may experience one type of transition (G to A) at a rate X and 2 types of transversion (G to C, G to T) at rate Y. The aggregate rate ratio of transition to transversion has a null expectation of $R = X/2Y = 0.5 = 50\%$ ($>R$ indicates that it is being fit or divergent in accordance with value and $<R$ shows opposite results to earlier) leading to transition bias relative to a null model of equal rates. This study emphasizes on 2 major objectives. The primary objective is to analyze codon bias of *matK* and *rbcL* genes among the individuals of this family and second one is to find out evolution-based phylogenetic tree.

Materials and Methods

Data collection

The family Dipterocarpaceae comprises of 15 genera (www.theplantlist.org/1.1/browse/A/Dipterocarpaceae/). The entire coding region of *matK*, *rbcL* gene sequences, and available amino acid sequences (*matK* and *rbcL*) of this family were retrieved from the taxonomy database of National Centre for Biotechnology Information (www.ncbi.nlm.nih.gov/nucleotide/?term=Dipterocarpaceae) (dated 12 December, 2017). In all, 76 *matK* and 33 *rbcL* gene sequences belonging to 28 and 16 species, respectively, were down loaded from GenBank and documented in the supplementary table, ie, Supplemental Tables S1 and S2, respectively.

Indicator of codon usage

Codon usage pattern in the core barcode region was analyzed by using codon W 1.4.2. Relative synonymous codon usage (RSCU) is the ratio of the observed frequency of synonymous codons for a particular amino acid to the expected frequency.³⁸ Thus, RSCU values close to 1 indicate the lack of bias for codon usage where as the value >1 or <1 means preference and avoidance of that particular codon. The effective number of codon (ENC) is used to show the extent of codon bias of a gene and to quantify the absolute codon usage bias of a coding sequence.³⁹ The values of ENC always remain in-between 20 (a gene with extreme codon bias uses only one codon per amino acid) and 61 (a gene with no codon bias uses all the synonymous codon).⁴⁰ In general, 35 or less and 50 or higher ENC values of a gene are considered to have a strong and low codon bias, respectively.⁴¹ The expected ENC values from GC3s under no selection in accordance with null hypothesis have been calculated according to equation (1), where S = GC3s.

$$ENC = 2 + S + \left\{ \frac{29}{S^2 + (1 - S)^2} \right\} \quad (1)$$

The relationship between nucleotide content and codon usage by NC-plot is investigated to reveal the relationship among them. Wright³⁹ has suggested that NC-plot (ENC plotted against GC3s) is used to explain the pattern of synonymous codon usage. The codon choices of a gene influenced by a (G + C) mutation constrain usually lie on or just below the curve of the predicted value.³⁹ Codon adaptation index (CAI), a measurement of the expression of the gene, is used to estimate the extent of bias toward codon and its values range between 0 to 1.0, where in the CAI, a higher value means a stronger codon usage and a higher expression level.⁴²

Correspondence analysis by using codon W 1.4.2

Correspondence analysis (COA) is an ordination technique that identifies the major trends in the variation of the data where genes with their degree of variation are arranged along the continuous axes. It represents continuous variation accurately. The first axis captures most of the variation of genes and each of the subsequent axes shows a diminishing variation.

Chemical properties

The physiochemical properties like molecular weight (MW), theoretical isoelectric point (PI), percentage of positive and negative charged amino acid instability index, grand average of hydropathicity (GRAVY), etc were determined by using ProtParam (<http://web.expasy.org/protparam/>) (Supplemental Tables S3 and S4). The instability index provides an estimation of stability of a protein in vitro.

Sequence analysis

Gene sequences (*matK* and *rbcL*) were retrieved from the GenBank, which made a dataset of all the plant species of the family Dipterocarpaceae to find out the interspecific variation. Another dataset comparing the various genus of the family Dipterocarpaceae like *Anisoptera*, *Dipterocarpus*, *Shorea*, *Hopea*, *Parashorea*, etc was made to find out intergeneric variation.

Phylogenetic analysis

The nucleotide frequency and transition/transversion bias was computed by Molecular Evolutionary Genetics Analysis (MEGA 7.0).⁴³ DVADIST program from PHYLIP was used to analyze the distance between the clades. The phylogenetic data were validated by re-sampling sequence data using bootstrap, performed by NJ-plot in Phylogeny Inference Package (PHYLIP).⁴⁴ Clustering of individuals was made on the basis of their position on the phylogenetic tree (clusters I, II, III, IV, V, and VI for *matK* and clusters I, II, III, IV, and V and sub-clusters I and II of cluster I for *rbcL*). Few clades were found in each cluster and their numbering was given accordingly from top to bottom.

Results

Advancement of molecular biology and DNA sequencing of the genome of various organisms rapidly provide valuable information regarding their genetic makeup and function. In this study, changes in the nucleotide sequence of *matK* and *rbcL* genes of Dipterocarpaceae were analyzed to know the inter-generic, inter-specific, and intra-specific variation.

Phylogenetic analysis

Monophyly individuals of all the clusters showed poor barcode (*matK*) success (Figure 1) which may be caused by slow mutation, poor seed dispersal, woody plant with long-generation time and polyploidy (*Hopea hainanensis* and *Shorea robusta*), etc. Therefore, long time was taken to reach monophyly. Individuals of clade 1 of clusters II, III, and VI and clade 5 of cluster IV (Figure 1) may radiate recently due to slow mutation rate and dense population. Genetic drift may lead to conspecific individuals together from other as indicated by their branch length value in the clade 1 of cluster I and clade 2 of both the clusters II and IV. Either rapid mutation or geographical constrains in clade 2 of clusters I and III, in clade 1 of cluster V, and in clades 1 and 4 of cluster IV possibly lead to speciation causing high barcode success in their barcode region (*matK*). Figure 2 shows that most individuals of all the clusters except few clades and monophyly groups had lack of proper divergence of *rbcL* gene to differentiate them as indicated by branch length. *Dipterocarpus intricatus* is the most diversified in their barcode region. Most of the individual of clusters I, II, IV, and VI of *matK*, and sub-clusters I and II and clusters IV and V of *rbcL* were less evolutionary divergent and may lead to extinction.

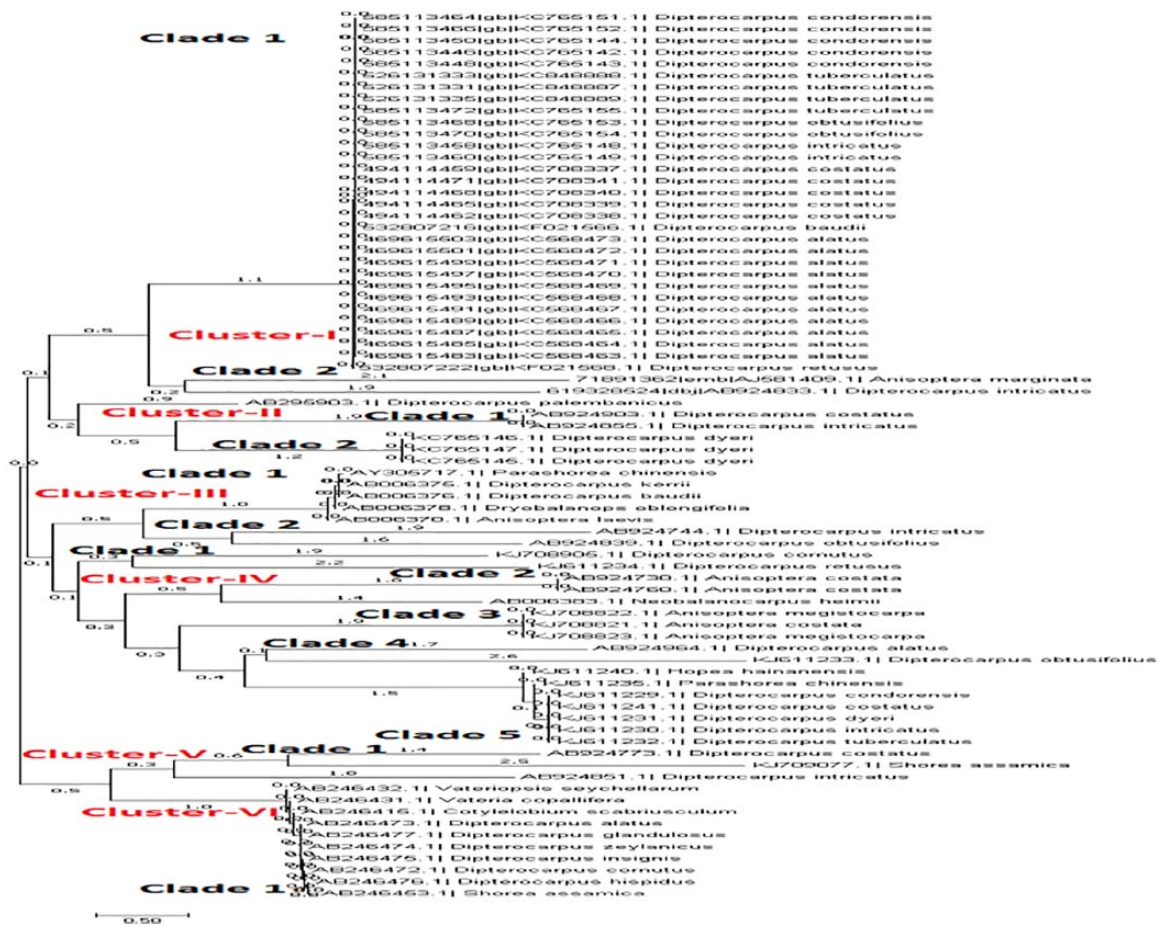


Figure 1. Evolutionary relationship of taxa based on *matK*. The evolutionary history was inferred by neighbor-joining method. Optimal tree with sum of the branch length=41.7.

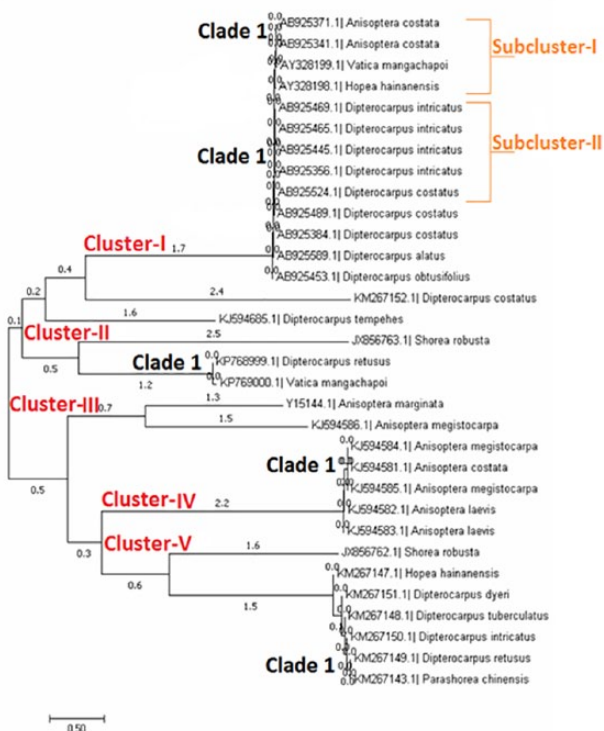


Figure 2. Evolutionary relationship of taxa based on *rbcL*. The evolutionary history was inferred by neighbor-joining method. Optimal tree with sum of the branch length=21.11.

Codon usage

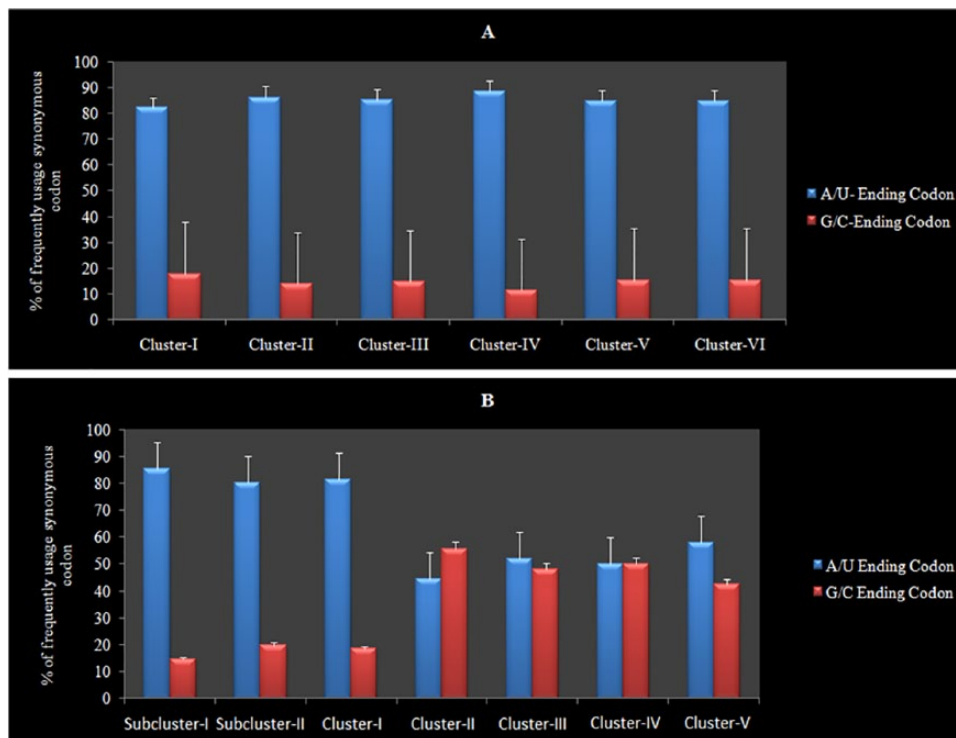
Average GC content among all the clusters ranged from 32.80% to 34.4% for *matK* and 42.2% to 44.6% for *rbcL* gene sequences (Table 1). So both these *matK* and *rbcL* genes were (A + T) rich, and overall codon usage was biased toward A- and U-ending codons (both Figure 3A and B) except clusters II (G- and C-ending bias) and IV of *rbcL* (where AU- and GC-ending codons occur at equal proportion) (Figure 3B), suggesting compositional constraints are important factors in shaping the codon usage variation in these genes.

Synonymous codon usage pattern

Almost all the clusters of *matK* (Figure 4A) showed similar kind of codon usage pattern in most frequently used codon. But clusters III and VI used very similar types of codon to encode same series of amino acids, and RSCU values of these codons were almost same. On the contrary, codon usage pattern among the clusters of *rbcL* (Figure 4B) was not uniform, but sub-clusters I and II and cluster I showed a similar fashion of codon usage pattern and had almost similar RSCU values. In broad sense, Figure 4A showed that *matK* gene in all the individuals of this family underwent the same degree of selection pressures unlike *rbcL* gene. But in strict sense, individuals

Table 1. GC content in the all the clusters of *matK* and *rbcL*.

CLUSTERS (<i>matK</i>)	AVERAGE GC%	CLUSTERS (<i>rbcL</i>)	AVERAGE GC%
Cluster I	33.0	Sub-cluster I	43.1
Cluster II	33.7	Sub-cluster II	42.2
Cluster III	33.2	Cluster I	42.5
Cluster IV	33.3	Cluster II	41.3
Cluster V	34.4	Cluster III	43.3
Cluster VI	32.8	Cluster IV	43.1
		Cluster V	44.6

**Figure 3.** Percentage of frequently synonymous codon usage among the clusters: (A) *matK* and (B) *rbcL*.

within clusters III and IV of *matK* (Figure 4A) and sub-clusters I and II and cluster I of *rbcL* (Figure 4B) were reviewed under the same scale of selection pressures.

Nucleotide content in 3 codon positions of gene

The GC content in 3 codon positions (GC1, GC2, and GC3) of *matK* gene among all the clusters was 35.5% to 39.7% (mean: 37.93%, standard deviation: 1.59%), 29.2% to 49.8% (mean: 35.33%, standard deviation: 7.82%), 30.5% to 34.9% (mean: 32.71%, standard deviation; 1.96%), respectively, and for *rbcL*, GC1 was 28.6% to 55.5% (mean: 40.95%, standard deviation: 12.90%), GC2 was 37.9% to 61.1% (mean: 48.75%, standard deviation: 8.86%), and GC3 was 28.9% to 55.3% (mean: 39.4%, standard deviation: 9.6%). Figure 5B showed nearly close GC1,

GC2, and GC3 values among the sub-cluster I, sub-cluster II, and cluster I.

According to neutrality analysis, it is found that mutational pressure presumably influences the codon bias if the correlation between GC12 and GC3 is statistically significant and the slope of the regression line is close to 1. Conversely, a narrow distribution of GC content and the nonsignificant correlation between GC12 and GC3 are caused by selection.^{45,46} In vitro stable protein encoding individuals within each cluster of both *matK* and *rbcL* (in vitro stable protein-encoded individuals [IN-VSPEI]) (Supplemental Tables S3 and S4) were considered to calculate GC (GC12 and GC3), but all individuals (in vitro stable and unstable protein encoding individuals) in the clusters were included for average GC calculation. Neutrality plot analysis (GC12 vs GC3) was performed in all the clusters

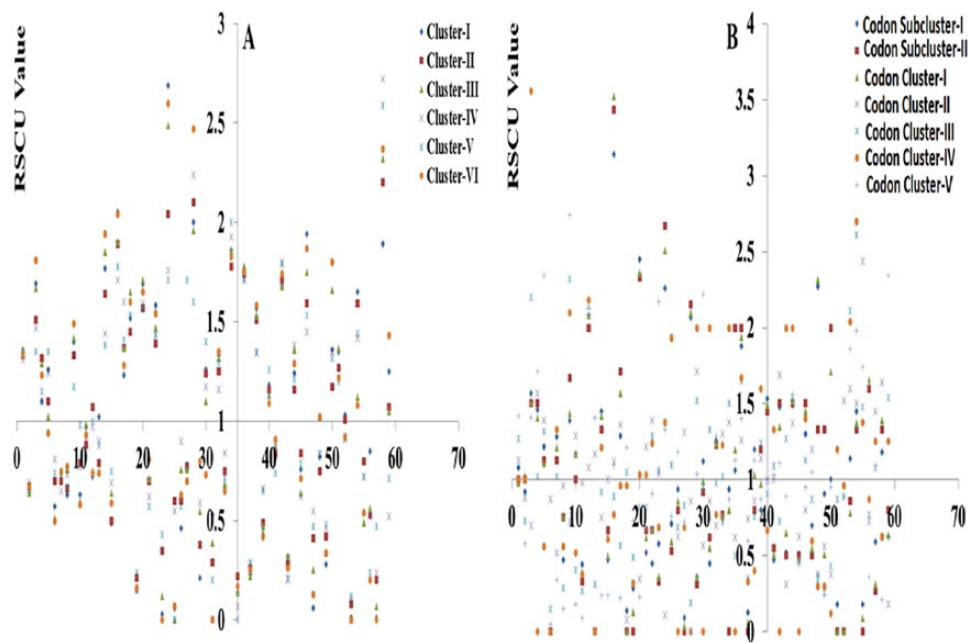


Figure 4. Codon usage pattern among the group members of this family based on RSCU value: (A) *matK* and (B) *rbcL*. RSCU: relative synonymous codon usage.

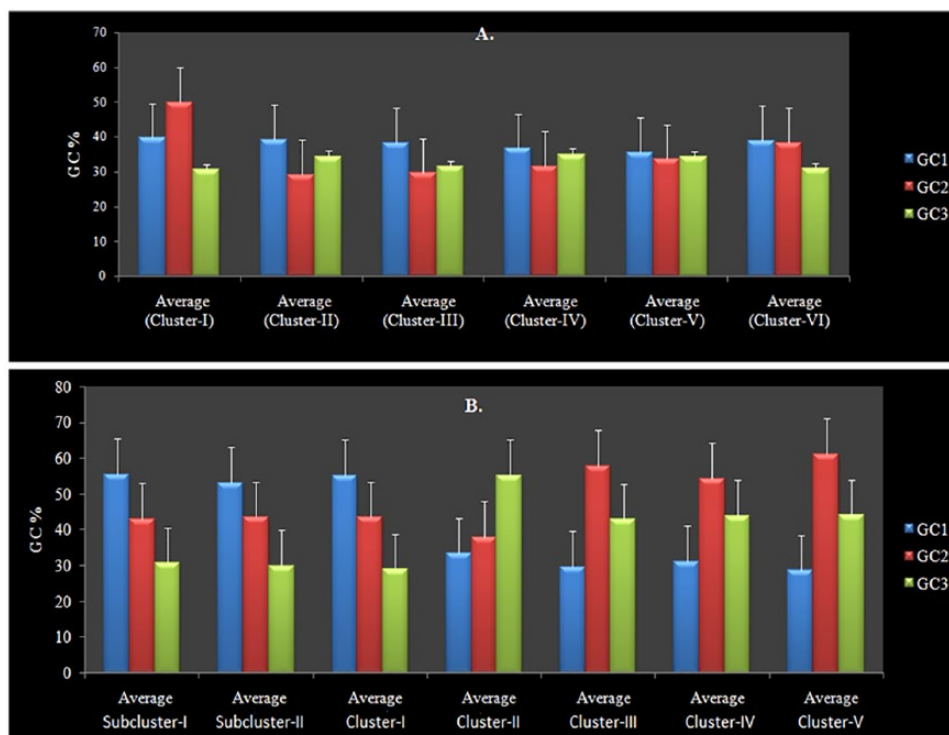


Figure 5. GC1, GC2, and GC3 among all the clusters: (A) *matK* and (B) *rbcL*.

(both *matK* and *rbcL* genes) to analyze relationships among the 3 codon positions. It was found that both *matK* and *rbcL* genes had a narrow range of GC content as mentioned earlier and all the clusters of *matK* gene showed no significant correlation between GC12 and GC3 except cluster IV (Table 2), suggesting that selection played a most prominent role in codon usage bias. In case of *rbcL*, there was significant correlation among

the clusters and the individuals within the each cluster except the clusters IV, V, and sub-cluster II (where *r* value calculation is not possible because GC12 [48.3%] and GC3 [29.9%] are same among the individuals) and V (having 2 individuals hindering from significant correlation test) (Table 2), showing that mutations rather selection are playing major role in shaping codon usage bias.

Table 2. Correlation coefficient (*r*), *P* value, and regression square of *matK* and *rbcL*, respectively.

	CORRELATION COEFFICIENT (<i>R</i>)	<i>P</i> VALUE AT 0.05% PROBABILITY LEVEL	REGRESSION SQUARE
Cluster (<i>matK</i>)			
Among the clusters	0.61	0.19	0.38
Cluster I	−0.59	0.59	0.03
Cluster II	0.65	0.54	0.42
Cluster III	0.84	0.16	0.71
Cluster IV	−0.62	0.013	0.39
Cluster V	0.52	0.48	0.27
Cluster VI	0.47	0.17	0.22
Clusters (<i>rbcL</i>)			
Among the clusters	−0.96	0.0006	0.93
Sub-cluster I	−0.97	0.03	0.95
Sub-cluster II	NA	NA	NA
Cluster I	−1.0	0.00001	1.0
Cluster II	−0.99	0.01	0.99
Cluster III	NA	NA	NA
Cluster IV	−0.85	0.15	0.73
Cluster V	NA	NA	NA

NA, not applicable, because sample size is 2 which is invalid for correlation.

Relation between ENC and GC3s

ENC and GC3s values for *matK* calculated in the same way of the previous point 3.4 (GC content—GC12 and GC3) varied from 47.47% to 57.67% (mean: 51.10%, standard deviation: 2.57%) and 0.280% to 0.333% (mean: 2.57%, standard deviation: 0.022%), respectively. For *rbcL*, it was from 45.62% to 61% (mean: 49.23%; standard deviation: 3.2%) and 0.282% to 0.553% (mean: 0.375%, standard deviation: 0.102%), respectively. Wright has proposed that codon usage variation among the genes can effectively be analyzed by a plot of ENC vs GC3s⁴⁷ where GC3s entirely operate codon usage bias if the value of ENC remains close/fall on the expected curve between GC3s and ENC plot. Figure 6A showed that almost all the points of observed ENC value were either below or above the standard ENC value, suggesting their independence of GC3s except few points (few IN-VSPEI) of cluster IV which was dependent on GC3s, ie, G + C mutational bias (Figure 6B). Observed ENC values of all the individuals of sub-clusters I and II and cluster I were close to standard ENC value toward GC poor regions which were certainly originated from compositional constraints, and rest of the points are below their standard ENC value.

Mutational bias analysis

In the case of mutational bias, generally, GC or AT is used proportionally among the degenerate codon groups in a gene. On the contrary, GC or AT is not proportionally used for codon choice by natural selection.⁴⁸ The relationship among G, C, A, and T content in 4 degenerated codon families are analyzed to know whether these codon bias choices are restricted to the high bias genes or not (Figure 7A and B). Figure 7A showed codon bias for *matK* was mostly determined by a factor-like selection except in a specific few cases where mutation causes codon bias (eg, few individuals of cluster IV). Codon bias choice in sub-cluster II and cluster I of *rbcL* (Figure 7B) is very similar, which was controlled by mutation.

Chemical properties

Physiochemical properties of completely sequenced *matK* and *rbcL* proteins of this family were revealed by using the software ProtParam (Supplemental Tables S3 and S4, respectively). Aliphatic side chains of alanine, valine, isoleucine, and leucine of a protein may be regarded as a positive factor to increase the thermostability for the globular proteins. The aliphatic index for *matK* and *rbcL* proteins of this family

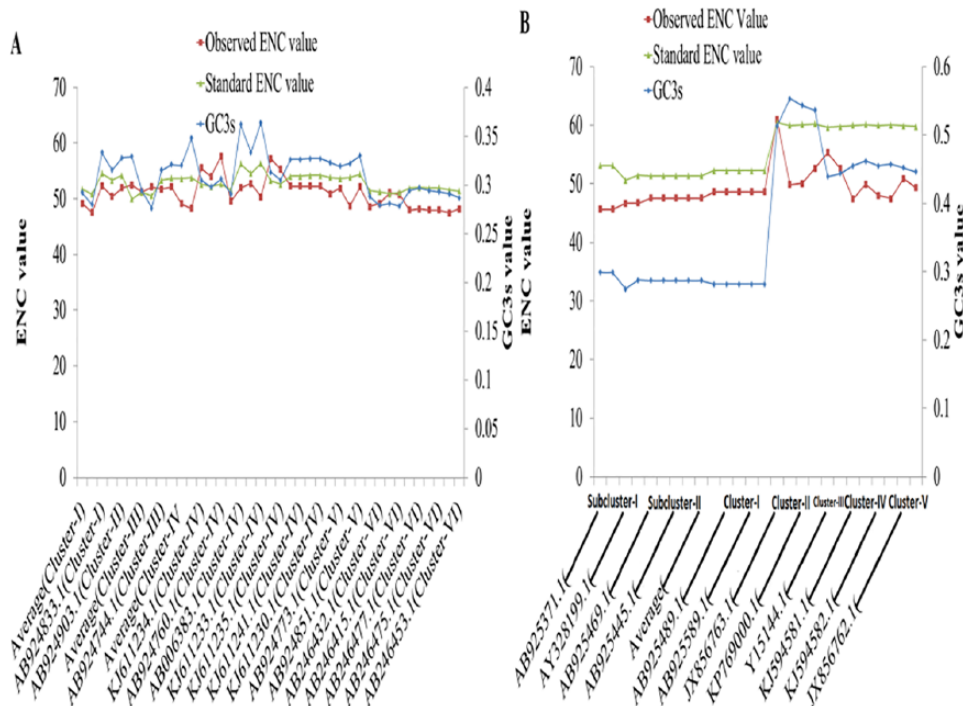


Figure 6. ENC vs GC3s (the portions where observed ENC value occur on the standard curve): (A) *matK* and (B) *rbcL* gene of this family. ENC, effective number of codon.

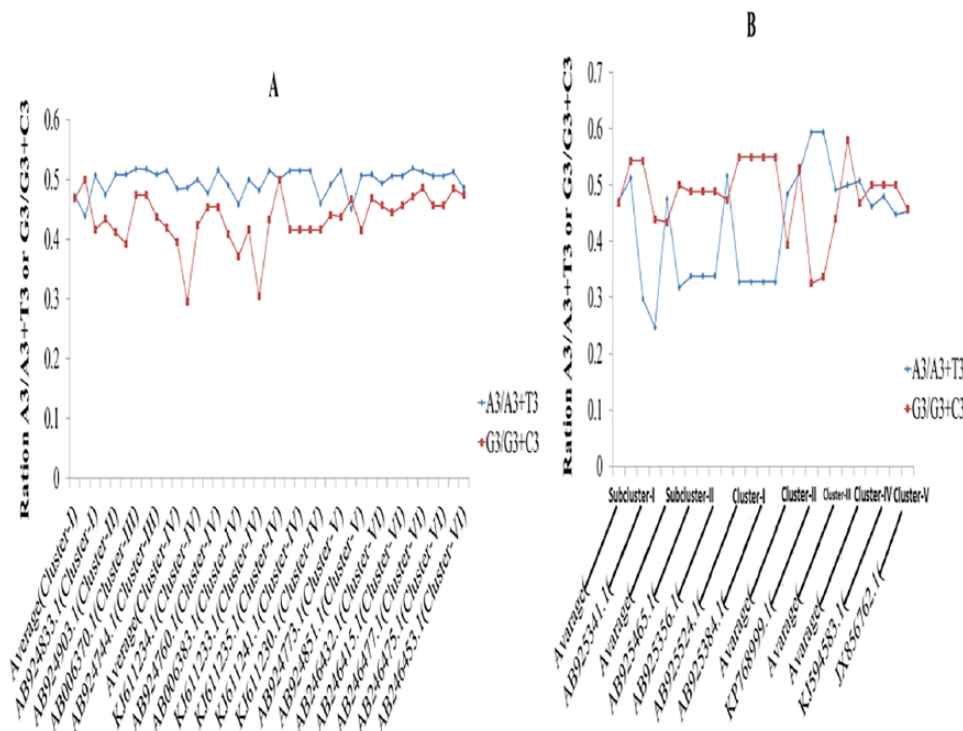


Figure 7. Comparative ratio of A3/A3 + T3 and G3/G3 + C3: (A) *matK* and (B) *rbcL*.

ranged from 92.45% to 106.45% and 79.63% to 85.98%, respectively. Theoretically, the protein having instability index <40 is considered as stable. Thirty-three and 18 individual proteins of *matK* and *rbcL*, respectively, were found to be in vitro stable.

Correspondence analysis

In this study, we investigated on the synonymous codon usage variation among the clusters and the individuals of each cluster of *matK* and *rbcL* genes of this family (considering all

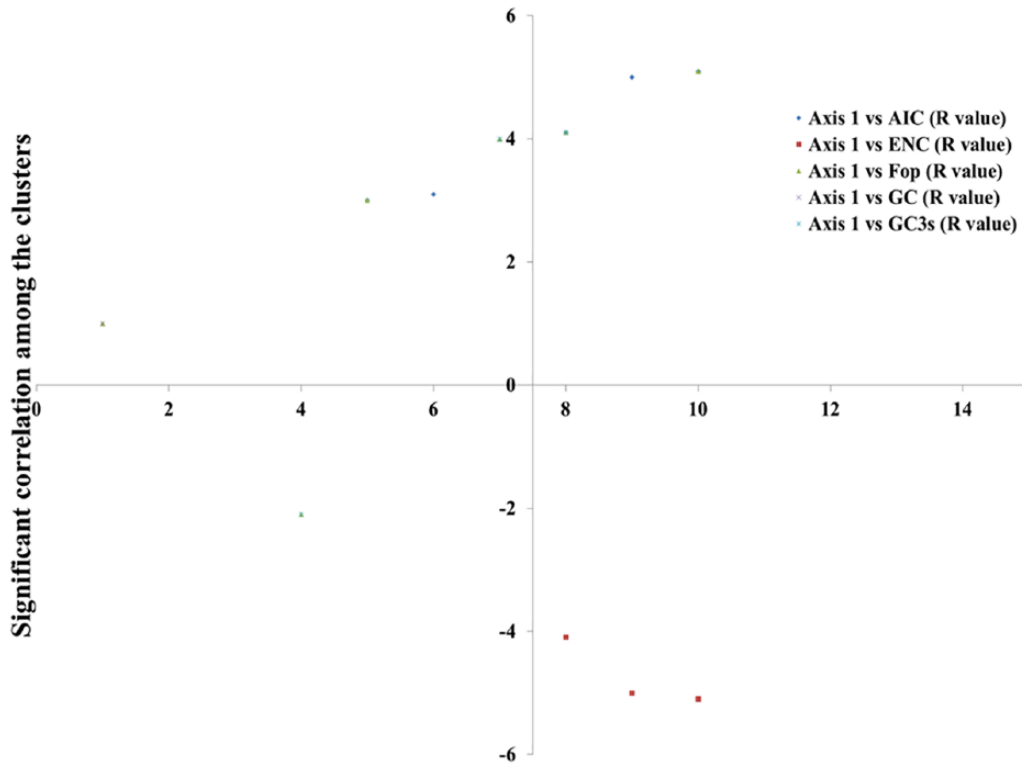


Figure 8. Schematic representation of correlation axis 1 vs CAI or ENC or Fop or GC or GC3s for each cluster of *matK* accordingly on the graph (1-5 points) and the distribution of IN-VSPEI of each cluster of *matK* is also presented on this graph having 1.1 to 5.1 points, where negative (-) and positive (+) sign denotes negative and positive correlation, respectively, with axis 1. CAI, codon adaptation index; ENC, effective number of codon; Fop, frequency of optimal codons; IN-VSPEI, in-vitro stable protein-encoded individual.

the individuals and the stable protein encoding individuals as mentioned in earlier point 3.4). First 2 axes explain major variation which is gradually decreased in the successive axes. Ordinations of genes on the first 4 axes are examined for correlations with indices of codon usage (CAI, ENC, Fop [frequency of optimal codons], GC, and GC3s). In vitro stable protein-encoded individuals of clusters IV and V of *matK* showed more or less same gene expression where genes with higher expression level had a greater degree of codon bias, but the cluster IV had a tendency of GC compositional constraints at synonymous codon position (Figure 8). It was found that there was a significantly positive correlation between axis 1 and Fop in most clusters. This result suggested that selection was one of the major factors influencing the synonymous codon bias of the *matK* gene among the individuals of this family. Highly expressed *rbcl* genes of IN-VSPEI (sub-cluster I) are significantly correlated with GC and GC3s (Table 3).

Sequence analysis

Substitution bias (transition/transversion) ratio at codon position for each cluster revealed evolutionary trend in accordance with their values. The inference was made for both the genes on the basis of overall substitution bias value on their entire codon position (1st + 2nd + 3rd nucleotide). According to selective hypothesis on substitution bias (at 1st + 2nd + 3rd

position), clusters I, II, and V of *matK*, and sub-cluster II and clusters I and IV of *rbcl* (Table 4) exhibited evolutionary less divergence due to its low null expectation value ($<R$), whereas cluster VI of *matK*, and sub-cluster I, cluster II, and cluster V of *rbcl* having high null expectation value ($R = >0.85$) represented less evolutionary divergence, which was depicted in the phylogenetic tree (Figures 1 and 2).

Discussion and Conclusion

Codon usage bias is a complex and important issue regarding evolution in both prokaryotes and eukaryotes. There are some hypotheses that have been proposed to explain the origin of codon usage bias. Neutral theory⁴⁹ and the selection-mutation-drift balance model^{38,50} are one of the best representatives among them (hypothesis of the origin of codon usage bias). According to the neutral theory, random synonymous codon choices are the results of mutation at degenerate coding positions. The selective-mutation-drift model explains that the codon bias is supposed to be determined by the stability among mutational pressures, genetic drift, and selection. However, with the advancement of genome projects in the recent years, these 2 hypotheses are not sufficient for the explanation of codon usage bias. Several parameters like gene length,²³ GC content,^{51,52} recombination rate,^{51,53,54} gene expression level,^{23,53,55} RNA structure,^{24,56,57} protein structure,⁵⁸ intron length,⁵⁹ population size,⁶⁰ evolutionary age of genes,⁶¹

Table 3. Correlation (*R*) of axis 1 with ENC, CAI, Fop, GC content, and GC content at synonymous codon position of *matK* and *rbcL*, respectively.

	AXIS 1 VS CAI (<i>R</i> VALUE)	AXIS 1 VS ENC (<i>R</i> VALUE)	AXIS 1 VS FOP (<i>R</i> VALUE)	AXIS 1 VS GC (<i>R</i> VALUE)	AXIS 1 VS GC3S (<i>R</i> VALUE)
Clusters (<i>matK</i>)					
Cluster I	0.124 (NS)	−0.206 (NS)	0.769*,**	0.716*,**	0.124 (NS)
IN-VSPEI (Cluster I)	0.644 (NS)	−0.639 (NS)	0.533 (NS)	0.567 (NS)	−0.381 (NS)
Cluster II	−0.070 (NS)	0.395 (NS)	0.411 (NS)	−0.142 (NS)	0.496 (NS)
IN-VSPEI (Cluster II)	−0.978 (NS)	−0.974 (NS)	−0.999**	−0.916 (NS)	−0.997**
Cluster III	0.943*,**	0.720 (NS)	0.889*,**	0.706 (NS)	0.732 (NS)
IN-VSPEI (Cluster III)	0.997**	−0.425 (NS)	0.964 (NS)	0.964 (NS)	0.994 (NS)
Cluster IV	0.381 (NS)	−0.373 (NS)	0.734*,**	−0.133 (NS)	0.816*,**
IN-VSPEI (Cluster IV)	0.581**	−0.612**	0.852*,**	−0.276 (NS)	0.849*,**
Cluster V	0.999*,**	−0.998**	−0.986 (NS)	0.702 (NS)	−0.265 (NS)
IN-VSPEI (Cluster V)	0.999*,**	−0.999**	0.998**	0.960 (NS)	0.472 (NS)
Cluster VI	−0.370 (NS)	0.365 (NS)	0.365 (NS)	−0.400 (NS)	−0.244 (NS)
IN-VSPEI (Cluster VI)	−0.375 (NS)	0.385 (NS)	0.460 (NS)	−0.421 (NS)	−0.246 (NS)
Clusters (<i>rbcL</i>)					
Sub-cluster I	−0.0302 (NS)	0.598 (NS)	−0.204 (NS)	−0.272 (NS)	0.197 (NS)
IN-VSPEI (Sub-cluster I)	1*,**	−1*,**	1*,**	−1*,**	1*,**
Sub-cluster II	Not identified	Not identified	Not identified	Not identified	Not identified
IN-VSPEI (Sub-cluster II)	Not identified	Not identified	Not identified	Not identified	Not identified
Cluster I	Not identified	Not identified	Not identified	Not identified	Not identified
IN-VSPEI (Cluster I)	Not identified	Not identified	Not identified	Not identified	Not identified
Cluster II	−0.982 (NS)	0.992 (NS)	0.994 (NS)	0.405 (NS)	−0.939 (NS)
IN-VSPEI (Cluster II)	−0.982 (NS)	0.992 (NS)	0.994 (NS)	0.405 (NS)	−0.939 (NS)
Cluster III	NA	NA	NA	NA	NA
IN-VSPEI (Cluster III)	NA	NA	NA	NA	NA
Cluster IV	−0.235 (NS)	0.050 (NS)	−0.015 (NS)	−0.296 (NS)	−0.339 (NS)
IN-VSPEI (Cluster IV)	−0.144 (NS)	−0.237 (NS)	0.445 (NS)	−0.445 (NS)	−0.445 (NS)
Cluster V	0.306 (NS)	0.306 (NS)	0.811**	0.332 (NS)	0.743 (NS)
IN-VSPEI (Cluster V)	NA	NA	NA	NA	NA

Abbreviations: CAI, codon adaptation index; ENC, effective number of codon; Fop, frequency of optimal codon; IN-VSPEI, in-vitro stable protein-encoded individual; NS, non-significant correlation.

NA, not applicable, because sample size is not more than 2. Not identified means codonW could not produce data of 4 axis. * and ** means significant at 0.01 and 0.05 probability levels, respectively.

environmental stress,⁶² hydrophobicity and aromaticity of encoded proteins,^{63,64} and so on may influence the codon usage bias. In this study, gene expression level and gene compositional constraint have been given primary focus. In vitro stable protein of both *matK* and *rbcL* genes of the family Dipterocarpaceae selected through natural selection had higher barcode success than mutational selection. Thus, from this study, it could be suggested that mutation and selection are operated randomly on the genes—both *rbcL* and *matK*.

GC rich organisms such as bacteria, archaea, fungi, *Triticum aestivum*, *Hordeum vulgare*, and *Oryza sativa*^{46,65} tend to use G or C in their third position. Meanwhile, AT rich organisms like *Onchocerca volvulus*, *Mycoplasma capricolum*, *Plasmodium falciparum*, etc^{66,67} show A or T preference in the third position. The core barcode genes of this family (Dipterocarpaceae) showed AT richness and the overall codon usage was biased toward A- and U-ending codon except individuals of cluster II of *rbcL* gene. By combining analyses of both selective

Table 4. Substitution bias at codon position of *matK*.

	CODON POSITION				INFERENCE ON SUBSTITUTION BIAS BASED ON SELECTIVE HYPOTHESIS
	1ST	2ND	3RD	1ST + 2ND + 3RD (CODON)	
Clusters (<i>matK</i>)					
Cluster I	0.97	0.72	0.44	0.45	Evolutionary less divergent
Cluster II	0.34	0.49	0.0	0.027	Less divergent
Cluster III	0.64	0.88	0.51	0.79	Evolutionary divergent
Cluster IV	0.29	0.89	0.42	0.78	Evolutionary divergent
Cluster V	0.03	0.05	0.00	0.26	Less divergent
Cluster VI	1.72	0.27	1.82	1.02	Highly divergent
Clusters (<i>rbcL</i>)					
Sub-cluster I	2.01	1.00	02.02	1.77	Highly divergent
Sub-cluster II	0.50	0.50	0.50	0.50	Evolutionary divergent
Cluster I	1.93	0.00	0.00	0.37	Less divergent
Cluster II	1.93	0.00	0.00	0.95	Highly divergent
Cluster III	NA	NA	NA	NA	NA
Cluster IV	0.50	0.00	0.50	0.00	Less divergent
Cluster V	0.67	1.93	0.75	0.86	Highly divergent

NA, not applicable, because more than 2 samples are required to calculate substitution bias in MEGA 7.0.

hypothesis on substitution bias (Table 4) and phylogenetic tree (Figures 1 and 2), it was reflected that generally null expectation, ie, *R* values greater than 0.85 and 0.5 or <0.5, showed less evolutionary divergence because very high and very low transition may negatively affect the fitness of organism.

Phylogenetic analysis helped to identify the variations, patterns, transition/transversion bias, and codon bias in nucleotide sequence. Genome-based phylogeny is found to be effective in this concern, and it has been practiced in bacterial system (due to smaller genome size). In angiosperm, whole genome phylogeny is being challenged, because of very hard processing of so large massive information unlike bacterial genome. However, it is quite obvious to looking at DNA barcode. A software, MEGA, provided information about inter- and intra-specific relationship of the family Dipterocarpaceae. Phylogenetic tree analysis showed that cluster IV of *matK* and cluster I of *rbcL* had longer branch length making the evolutionary sense between the species of this family. *matK* is a good candidate for DNA barcoding among the members of this family because of their high discrimination success than *rbcL* and showed a predictable range of mutational events, ie, nucleotide substitution (Supplemental Table S5), whereas *rbcL* involved additions and deletions. It was also found that the genetic diversity was affected by habitat alteration creating a situation like fragmented population, limited seed dispersal, etc.

Acknowledgements

Raju Biswas is thankful to CSIR for Junior Research Fellowship (File No: 09/025(0216)/2015-EMR-I). Authors are thankful to UGC-Center of Advanced Study and DST-FIST at Department of Botany, The University of Burdwan for pursuing research activities.

Author Contributions

Both RB designed the work. RB conducted the work. Both RB and ASP analysed the data. RB and ASP wrote the paper. RB checked the paper. All authors finalized and submitted the paper.

Supplemental Material

Supplemental material for this article is available online.

ORCID iD

Rajib Bandopadhyay  <https://orcid.org/0000-0002-8318-5631>

REFERENCES

- Harnelly E, Thomy Z, Fathiya N. Phylogenetic analysis of Dipterocarpaceae in Ketambe Research Station, Gunung Leuser National Park (Sumatra, Indonesia) based on *rbcL* and *matK* genes. *Biodiversitas*. 2018;19:1074–1080. doi:10.13057/biodiv/d190340.
- Trang NTP, Duc NM, Sinh NV, Triest L. Application of DNA barcoding markers to the identification of *Hopea* species. *Genet Mol Res*. 2015;14:9181–9190. doi:10.4238/2015.August.7.28.

3. Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 2007;23:167–172. doi:10.1016/j.tig.2007.02.001.
4. Cho Y, Qiu YL, Kuhlman P, Palmer JD. Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci U S A.* 1998;95:14244–14249.
5. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol Phylogenet Evol.* 2003;29:380–395.
6. Cho Y, Mower JP, Qiu YL, Palmer JD. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *Proc Natl Acad Sci U S A.* 2004;101:17741–17746.
7. Muller K, Borsch AT. Phylogenetics of Utricularia (Lentibulariaceae) and molecular evolution of the trnK intron in a lineage with high substitutional rates. *Plant Syst Evol.* 2005;250:39–67.
8. Doebley J, Durbin M, Golenberg EM, Clegg MT, Ma DP. Evolutionary analysis of the large subunit of carboxylase (rbcL) nucleotide sequence among the grasses (Gramineae). *Evolution.* 1990;44:1097–1108.
9. CBOL Plant Working Group. A DNA barcode for land plants. *Proc Natl Acad Sci U S A.* 2009;106:12794–12797.
10. Hilu KW, Liang H. The matK gene: sequence variation and application in plant systematic. *Am J Bot.* 1997;84:830–839.
11. Kress WJ, Erickson DL, Jones FA, et al. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proc Natl Acad Sci U S A.* 2009;106:18621–18626.
12. Chase MW, Cowan RS, Hollingsworth PM, et al. A proposal for a standardised protocol to barcode all land plants. *Taxon.* 2007;56:295–299.
13. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcoding. *PLoS ONE.* 2011;6:e19254. doi:10.1371/journal.pone.0019254.
14. Kress WJ, Erickson DL. A two locus global DNA barcode for land plants: the coding rbcL gene complements the noncoding trnH-psb region. *PLoS ONE.* 2007;2:e508. doi:10.1371/journal.pone.0000508.
15. Burgess KS, Fazekas AJ, Kesanakurti PR, Graham SW, Husband BC, Newmaster GS. Discriminating plant species in a local temperate flora using the RbcL+matK DNA barcode. *Meth Ecol Evol.* 2011;2:333–340. doi:10.1111/j.2041210X.2011.00092.x.
16. Ennos RA, French GC, Hollingsworth PM. Conserving taxonomic complexity. *Trends Ecol Evol.* 2005;20:164–168.
17. Hollingsworth ML, Clark AA, Forrest LL, et al. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol Ecol Resour.* 2009;9:439–457. doi:10.1111/j.1755-0998.2008.02439.x.
18. Fazekas AJ, Kesanakurti PR, Burgess KS, et al. Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Mol Ecol Resour.* 2009;9:130–139. doi:10.1111/j.1755-0998.2009.02652.x.
19. Petit RJ, Excoffier L. Gene flow and species delimitation. *Trends Ecol Evol.* 2009;24:386–393.
20. Ghosh T. Studies on codon usage in *Entamoeba histolytica*. *Int J Parasitol.* 2000;30:715–722.
21. Grantham R, Gautier C, Gouy M. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 2000;8:49–62.
22. Ponger L, Duret L, Mouchiroud D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* 2001;11:1854–1860.
23. Duret L, Mouchiroud D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc Natl Acad Sci U S A.* 1999;96:4482–4487.
24. Carlini DB, Chen Y, Stephan W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes *Adh* and *Adhr*. *Genetics.* 2001;159:623–633.
25. Gu W, Zhou T, Ma J, Sun X. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems.* 2004;73:89–97.
26. Kahali B, Basak S, Ghosh TC. Reinvestigating the codon and amino acid usage of *S. cerevisiae* genome: a new insight from protein secondary structure analysis. *Biochem Biophys Res Commun.* 2007;354:693–699.
27. Versteeg R, Van Schaik BDC, Van Batenburg MF, et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 2003;13:1998–2004. doi:10.1101/gr.1649303.
28. Sugiura M. The chloroplast genome. *Plant Mol Biol.* 1992;19:149–168.
29. Morton BR. Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proc Natl Acad Sci U S A.* 1999;96:5123–5128.
30. Morton BR. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast. *DNA J Mol Evol.* 2003;56:616–629.
31. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *Mol Evol.* 1998;46:449–459.
32. Rosenberg MS, Subramanian S, Kumar S. Patterns of transitional mutation biases within and among mammalian genomes. *Mol Biol Evol.* 2003;20:988–993.
33. Wakeley J. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends Ecol Evol.* 1996;11:158–162.
34. Keller I, Bensasson D, Nichols RA. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genet.* 2007;3:e22.
35. Stoltzfus A, Yampolsky LY. Climbing mount probable: mutation as a cause of nonrandomness in evolution. *J Hered.* 2009;100:637–647.
36. Stoltzfus A, Norris RW. On the causes of evolutionary transition: transversion bias. *Mol Biol Evol.* 2015;33:595–602. doi:10.1093/molbev/msv274.
37. Vogel F, Kopun M. Higher frequencies of transitions among point mutations. *J Mol Evol.* 1977;9:159–180.
38. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol.* 1986;24:28–38. doi:10.1007/BF02099948.
39. Wright F. The effective number of codons used in a gene. *Gene.* 1990;87:23–29. doi:10.1016/03781119(90)904919.
40. Comeron JM, Aguade M. An evaluation of measures of synonymous codon usage bias. *J Mol Evol.* 1998;47:268–274.
41. Roychoudhury S, Mukherjee D. A detailed comparative analysis on the overall codon usage pattern in herpesviruses. *Virus Res.* 2010;148:31–43.
42. Sharp PM, Li WH. The codon adaptation index—a measure of directional synonymous codon usage, and its potential application. *Nucleic Acids Res.* 1987;15:1281–1295.
43. Kumar S, Tamura K. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* 2004;5:150–163.
44. Felsenstein J. Confidence limits phylogenies: an approach using the Bootstrap. *Evolution.* 1985;39:783.
45. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A.* 1988;85:2653–2657. doi:10.1073/pnas.85.8.2653.
46. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. *Genes Genet Syst.* 2003;78:343–352. doi:10.1266/ggs.78.343.
47. Duret L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 2000;16:287–289.
48. Sueoka N, Kawanishi Y. DNA G + C content of the third codon position and codon usage biases of human genes. *Gene.* 2000;261:53–62. doi:10.1016/S03781119(00)004807.
49. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res.* 1997;25:244–245.
50. Bulmer M. Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *J Evol Biol.* 1988;1:15–26.
51. Comeron JM, Kreitman M, Aguade M. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics.* 1999;151:239–249.
52. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. *Proc Natl Acad Sci U S A.* 2001;98:5688–5692.
53. Hey J, Kliman RM. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics.* 2002;160:595–608.
54. Kliman RM, Hey J. Hill Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret. *Genet Res.* 2003;81:89–90.
55. Stenico M, Lloyd AT, Sharp PM. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 1994;22:2437–2446.
56. Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. *Genetics.* 1994;138:227–234.
57. Chen Y, Carlini DB, Baines JF, et al. RNA secondary structure and compensatory evolution. *Genes Genet Syst.* 1999;74:271–286. doi:10.1266/ggs.74.271.
58. Oresic M, Dehn M, Korenblum D, Shalloway D. Tracing specific synonymous codon-secondary structure correlations through evolution. *J Mol Evol.* 2003;56:473–484.
59. Vinogradov AE. Intron length and codon usage. *J Mol Evol.* 2001;52:2–5.
60. Berg OG. Selection intensity for codon bias and the effective population size of *Escherichia coli*. *Genetics.* 1996;142:1379–1382.
61. Prat Y, Fromer M, Linial N, Linial M. Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evol Biol.* 2009;9:285.
62. Goodarzi H, Torabi N, Najafabadi HS, Archetti M. Amino acid and codon usage profiles: adaptive changes in the frequency of amino acids and codons. *Gene.* 2008;407:30–41.
63. Romero H, Zavala A, Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res.* 2000;28:2084–2090.
64. Ripe C, Delmotte F, van Ham RC, Moya A. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res.* 2004;14:44–53.
65. Hershberg R, Petrov DA. General rules for optimal codon choice. *PLoS Genet.* 2009;5:e1000556.
66. Saul A, Battistutta D. Codon usage in *Plasmodium falciparum*. *Mol Biochem Parasitol.* 1988;27:35–42.
67. Muto A, Yamao F, Osawa S. The genome of *Mycoplasma capricolum*. *Prog Nucleic Acid Res Mol Biol.* 1987;34:29–58.