Computers in Biology and Medicine 114 (2019) 103449

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Exploring the codon patterns between *CCD* and *NCED* genes among different plant species

R. Priya [a], P. Sneha [a], J. Febin Prabhu Dass [a], George Priya Doss C [a], M. Manickavasagam [b], Ramamoorthy Siva [a,*]

[a] School of BioSciences and Technology, Vellore Institute of Technology, Vellore, 632014, Tamil Nadu, India
[b] Department of Biotechnology, Bharathidasan University, Trichy, 620024, Tamil Nadu, India

ARTICLE INFO

Keywords:
Carotenoid cleavage dioxygenase
Nine cis-epoxy carotenoid dioxygenases
Phylogenetic tree
Codon usage bias
Correspondence analysis
Compositional mutation bias

ABSTRACT

Plant carotenoid cleavage oxygenase (*CCO*) is an enzyme which catalyzes carotenoids to apocarotenoid products that are involved in several vital physiological functions. The *CCO* exists in two forms, namely, CCD (Carotenoid Cleavage Dioxygenase) and NCED (Nine-Cis Epoxycarotenoid Dioxygenase). This paper relates to a comparative study on *CCD* and *NCED* genes through phylogeny and codon usage analysis. The result of the phylogenetic analysis indicates a closer relationship between *CCD* and *NCED* subclass genes, while the RSCU values indicate a high preference for CUC codon in both *CCD* and *NCED* gene families. The mean ENc value of *NCED* genes was found to be 48.76, suggesting a higher codon bias compared to *CCD* genes. However, the ENc-GC_{3S} plot suggests that both the gene families are under mutational pressure with variations according to their species-specific role. Similarly, the multivariate analysis also suggests that nucleotide mutation bias influences codon usage. Correlation analysis of Axis I and codon adaptation index values indicate a significant correlation between critical indices. Even though the prominence of the variations in codon usage between the two gene families, they are exerted towards the time-specific functional requirement for that plant species. This is evident from the cleaving roles of these enzymes against various carotenoids at different growth stages. The result of this investigation indicates that *CCD* and *NCED* genes are under mutational pressure. This codon bias study paves the way for increasing the production of apocarotenoids, which have a great significance in the industry.

1. Introduction

Carotenoids are nutritionally valuable tetraterpenoid pigments, residing in the plastids of plants contributing to the yellow and orange colors to various flowers and fruits [1,2]. To date more than1181 carotenoids have been reported in 700 organisms [3]. The enzyme carotenoid cleavage oxygenases (COO) break the carotenoid in their conjugated double bonds to produce various apocarotenoids [4]. The phytohormone abscisic acid (ABA) which are involved in plant defense mechanism [5–7], the retinal and retinoic acid, the volatile aromatic compounds (i.e., α- and β-damascenone, β-cyclocitral, geranial, genaryl acetone, and β-ionone) are among the well-known apocarotenoids. Further examples of commercially essential pigments such as, bixin and crocin derived from *Bixa orellana* and *Crocus sativus* are being widely used in food and cosmetic industries [8,9].

The first carotenoid cleaving enzyme *Vp14* was isolated from viviparous maize mutant [10]. The four carotenoid cleavage dioxygenases (*CCDs*) and five 9-*cis*-epoxycarotenoid dioxygenases (*NCEDs*) were identified in the *Arabidopsis thaliana* genome based on the *Vp14* sequence homology. In these two genes, *CCDs* cleave a variety of *trans* carotenoid substrates and *NCEDs* are involved in ABA biosynthesis [11]. Substrate specificities and activities in plants have created a faint relationship between *CCDs,* and *NCEDs* [12]. However, the *NCED* and *CCD* genes share a common RPE65 (Retinoid Isomerohydrolase RPE65) domain. For members of the *CCO* family other than RPE65, iron is thought as activating oxygen for the cleavage of carotenoids or

Abbreviations: CCO, carotenoid cleavage oxygenase; CCD, Carotenoid Cleavage Dioxygenase; NCED, Nine-Cis Epoxycarotenoid Dioxygenase; ENc, Effective Number of codon; ABA, abscisic acid; DNA, Deoxyribonucleic acid; tRNA, transfer RNA; ML, Maximum likelihood; NJ, Neighbour Joining; BI, Bayesian inference; RSCU, relative synonymous codon usage; COA, correspondence analysis; CAI, codon adaptation index; CDS, Complete coding sequence; AIC, Akaike Information Criterion; ANOVA, Analysis of Variants; SD, Standard Deviation; RPE65, Retinoid Isomerohydrolase.

* Corresponding author.
E-mail address: siva.ramamoorthy@gmail.com (R. Siva).

https://doi.org/10.1016/j.compbiomed.2019.103449
Received 6 April 2019; Received in revised form 13 September 2019; Accepted 13 September 2019
Available online 14 September 2019
0010-4825/

lignostilbenes [13,14].

The genetic message from DNA is converted into their analogous protein using three-nucleotide codes known as codons. Many of these codons are translated into the same amino acids due to the degeneracy of the genetic code. The codon stipulating the similar amino acid is termed as synonymous codons, wherein variations in the frequencies of these synonymous codons are termed as codon bias [15]. Despite the genetic code being universal, codon usage between different species is extremely inconstant [16]). These differences are identified within species as well as genes [17], which are mainly connected to the functionality of the gene [18,19]. Several studies have reported on the codon biasness, tRNA abundance [20,21] and gene expression level [22,23] in various organisms such as *Saccharomyces cerevisiae*, *Escherichia coli*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Drosophila melanogaster* [24,25]. Selection and mutation are the two most important components that are responsible for the varaitions due to codon usage between genomes in numerous organisms. Based on the above mentioned two factors, selection in codon bias subsidizes the efficiency and accuracy of amino acid sequence [26], leading to the intent association of codon bias with the tertiary structure of the protein [27]. Mutation in codon bias can occur later due to the non-randomness in the mutational patterns, whereby some codons would be more mutable and, therefore, would have lower equilibrium frequencies [28,29]. The mutation leads to variation in the occurrence of codon usages by genomic G + C composition [30]. Studies relating to synonymous codons help in better understanding of the process of synonymous codon bias usage, host expression system, gene prediction from genomic sequence and protein functional classification [31,32]. A study of various organisms such as plants [33,34] and micro-organism has been documented [35,36]. Such studies have already been carried in various plant genes such as *PHE* [37], ribulose 1, 5 bisphosphate (RuBP) and chlorophyll a/b binding protein [38] and mitochondrial genes [39]. They have been used in the exposure of information relating to the molecular evolution of individual genes [34]. Hence, understanding the scope and reasons for codon biases is essential for gaining the knowledge of evolution.

Promisingly, the availability of a vast number of sequenced *CCO* genes of the plant provides an excellent opportunity to reveal the synonymous codon usage bias in them. Despite the presence of various reports on *CCO* genes, there is no information on their patterns of codon usage bias. Hence, for the first time, we have computationally analyzed 79 peptide coding mRNA sequences from 17 plant species for examining the factors and patterns that motivate the *CCD* and *NCED* genes. The objective of the study is to get an understanding of the codon usage bias in these two genes and their molecular significance. The authors have utilized the state-of-the-art computational tools for the comprehension of the codon bias in the *CCD* and *NCED* genes, which would help in understanding the mutational stress of both the genes. As these genes play a significant role in the production of the apocarotenoids that have industrial significance, research such as this may help to understand and increase the production of commercially important pigments.

2. Materials and methods

2.1. Sequence dataset

The authors have retrieved the *CCD* and *NCED* mRNA coding sequence (CDS) from 17 plant species belonging to 13 different families using EMBL-EBI (ENA- European Nucleotide Archive) database (http://www.ebi.ac.uk/services). Seventy-nine coding sequences were compiled for *CCD* and *NCED* families. Diverse plant families that include Asteraceae, Apiaceae, Bixaceae, Brassicaceae, Iridaceae, Fabaceae, Poaceae, Lauraceae, Oleaceae, Rubiaceae, Rosaceae, Rutaceae, and Vitaceae have been encompassed in this study. There are numerous plant species that contain either *CCD* or *NCED* genes. Hence we have considered the plant species that possess *CCD* and NCED genes in the same group of species. The dataset was constructed with great attention in eliminating incomplete and terminated codons in sequence considering only the complete (CDS) coding sequence. The comprehensive particulars of *CCD* and *NCED* gene sequences along with their respective accession numbers are depicted in Table 1.

2.2. Phylogenetic analysis

The CCD and NCED coding sequences collected from various plant species were subjected to a phylogenetic study for getting an inference relating to the relationships among them. Initially, the CDS were aligned using ClustalW with default settings [40]. The alignment was further refined using a tool trimAI [41] which exclude spurious sequences or regions which are poorly aligned from a multiple sequence alignment to enable the improvement in the quality of the phylogenetic analysis. Here, three standard tree-building methods have been adapted for phylogenetic reconstruction. An initial phylogenetic tree was built using the Neighbour-joining (NJ) method available in MEGA v5.05 software [42]. The most applicable nucleotide substitution model for Maximum Likelihood (ML) tree was selected using the Akaike Information Criterion (AIC) applied in Modeltest v3.7 [43]. Based on this, a rooted ML tree was constructed using the PhyML v3.0 [44]program by setting the bootstrap values to 1000. Next, the Bayesian inference (BI) was performed using MrBayes 3.1.2 [45]. Log-likelihood scores were generated by setting 3×10^6 generations and further sampled for every hundred generations. Once the log-likelihood scores were stabilized, a calculation of a consensus tree was computed by the omission of the first 25% trees as burn-in. In all the three phylogenetic reconstruction methods, the fungi *CCD* enzyme was utilised as an outgroup to root the tree of *CCO* gene family. The resulting *CCO* phylogeny tree was found to be robust in the varaible methods for tree reconstruction. Therefore ML method with Bayesian inference burn-in value tree is presented in this study.

2.3. Exploration of RSCU in CCD and NCED genes

The Relative Synonymous Codon Usage (RSCU) [46] was computed for each *CCD* and *NCED* genes using the CodonW1.4.2 program [47]. The RSCU is a robust method that recognizes the highly favored codon in the gene set. It is possible to observe whether these codons are GC or AT-ending for an specific amino acid. The RSCU values can be defined by the observed frequency to the expected frequency of the codon when all the synonymous codons for those amino acids are utilised in equal measure. The RSCU value, which is higher than 1.0 indicates positive codon usage bias (ample codons) of that specific codon. A value which is less than 1.0 indicates negative codon usage bias (less-ample codons) of a specific amino acid [46]. The synonymous codons with RSCU value greater than 1.6 indicates over-representation, while the RSCU value less than 0.6 is regarded as under-represented [48].

2.4. Major gene indices in CCD and NCED genes

Heterogeneity was identified by computing major codon indices that included GC, $GC_{3S,}$ and ENc values acquired from the *CCD* and *NCED* gene sequences. The indices GC, GC_{3S} and ENc have a specific impact for the elucidation of the heterogeneity in a group of the gene sequence or a genome. In these three indices, the GC measures the quantity of G+C in relation with A+T base pairs in the genome and also deliberates as an essential pointer for genome stability. The second index GC_{3S} is the frequency of G + C at the synonymous third codon position, which is the best indicator for the compositional properties of *CCD* and *NCED* genes. The third indices ENc, estimate the codon usage in the biasness in the gene set. The ENc values range from 20 to 61; the later value indicates the absence of codon bias where all codons have been utilized equally in contrast to the initial value, that is, only one codon is preferred for each amino acid in the sequence [49]. The heterogeneity in the codon usage from each group of *CCD* and *NCED* gene sequences of different plant species is shown by a frequency distribution graph with mean and

Table 1
CCD and *NCED* coding sequence data set.

Family Name	Plant Name	Gene Name		Gene Name	
		CCD	Acc. No	*NCED*	Acc. No
Apiaceae	*Daucus carota subsp. Sativus*	*CCD1*	ABB52081	*NCED3*	ABB52080
				NCED2	ABB52080
				NCED1	ABB52078
Asteraceae	*Chrysanthemum morifolium*	*CCD4a*	BAF36654	*NCED3a*	BAF36655
		CCD4b	BAF36656	*NCED3b*	BAF36657
Brassicaceae	*Arabidopsis thaliana*	*CCD1*	CAA06712	*NCED2*	AEE84030
		CCD4	AEE84154	*NCED3*	AEE75526
		CCD7	BAF01693	*NCED5*	AEE31178
		CCD8	AEE86121	*NCED6*	AEE76875
				NCED9	AEE36100
	Brassica napus	*CCD1*	AEN94300	*NCED3*	AEN94304
Bixaceae	*Bixa orellana*	*CCD1*	CAD71148	*NCED*	CAD71149
Fabaceae	*Phaseolus vulgaris*	*CCD1*	AAK38744	*NCED1*	AAF26356
				NCED2	AAK38744
	Pisum sativum	*CCD1*	BAC10549	*NCED2*	BAC10550
				NCED3	BAC10551
				NCED4	BAC10552
	Glycine max	*CCD4*	CAW43074	*NCED1*	AEK69514
		CCD7	HM366150	*NCED2*	AEK69515
		CCD8	ADK26571		
Iridaceae	*Crocus sativus*	*CCD1*	CAC79592	*NCED4*	ACD44928
		CCD4a	ACD62476		
		CCD4b	ACD62477		
Lauraceae	*Persea americana*	*CCD1*	AAK00622	*NCED1*	AAK00623
				NCED3	AAK00632
Orchidaceae	*Oncidium gower ramsey*	*CCD1*	ACP27629	*NCED*	ACP27628
Poaceae	*Oryza sativa*	*CCD1*	ABA99623	*NCED5*	AAW21321
		CCD7	CAD41601	*NCED4*	AAW21320
		CCD8a	BAB63485	*NCED3*	AAW21319
		CCD8b	BAC05598	*NCED2*	AAW21318
		CCD8c	EAY74583	*NCED1*	AAW21317
		CCD8d	EAZ42528		
	Sorghum bicolor	*CCD1*	EER91008	*NCED9*	EER95877
		CCD4	EER92020	*NCED3*	EER93751
		CCD7	EES11230		
		CCD8a	EES08398		
		CCD8b	EES03597		
		CCD8c	EER89247		
		CCD8c-1	EER89248		
		CCD8d	EER98721		
		CCD8d-1	EES15183		
Rosaceae	*Malus hupehensis*	*CCD1*	ACF75911	*NCED*	ACH85193
Rubiaceae	*Coffea canephora*	*CCD1*	ABA43900	*NCED3*	ABA43901
Rutaceae	*Citrus clementine*	*CCD4a*	DQ309330	*NCED3*	ABC26013
		CCD4b	ABC26012		
Vitaceae	*Vitis vinifera*	*CCD1*	AAX48772	*NCED1*	AFP28804
		CCD4a	AGT3321	*NCED2*	AAR11194
		CCD4b	KF008003	*NCED3*	JQ319644
		CCD7	CBI16625		
		CCD8	CCB46699		

The *CCD* (Carotenoid cleavage dioxygenase) and *NCED* (Nine- Cis epoxycarotenoid cleavage dioxygenase) mRNA coding sequence (CDS) from 17 plant species belongs to 13 different families were retrieved using data base with their accession number.

standard deviation (SD) of the above three significant indices (GC, GC_3s, and ENc). Moreover, the ENc value is related to translational efficiency owing to the occurrence of optimal codons[50]. *CCD* and *NCED* genes are featured with codon usage for validation. A total of 50 nuclear gene sequences were randomly collected from a model species *A. thaliana* genome (TAIR) database as a control (Supplementary Table 1). Here the heterogeneity of the significant indices of *Arabidopsis* genes has been compared with the *CCO* gene dataset. The differential variations in codon usage between the *CCO* and *Arabidopsis* genes were then tested using 1-way ANOVA.

The focus was on the codon composition in *CCD* and *NCED* genes involving the correlation between the GC_3s with $GC_{1,2}$. The pressure existing within the codon composition in a gene set was determined based on correlation analysis. The content of the GC in each of the three codon positions (GC1, GC2, and GC3) was calculated using the tool codon O [51]. The affiliation between ENc and GC_3s values to inspect the mutation bias was plotted. When the ENc values fall on or just beneath the predictable curve, the codon choice could be influenced by G+C mutation bias. Otherwise, unknown factors were responsible for the biasness in *CCD* and *NCED* gene sequences.

2.5. Correspondence analysis of CCD and NCED genes

Correspondence analysis (COA) is a usual method for the estimation of the synonymous codon usage bias in the gene set. A series of orthogonal axes are created by COA for identifying trends that explain the data variations, with each subsequent axis explaining a decreasing amount of the variations [52]. In this study, the COA on RSCU values have been implemented to enable examination of the differences in the codon usage between the *CCD* and *NCED* gene families. Here, every gene was plotted as a 59-dimensional vector space meant for reducing the impact of amino acid composition on codon usage. Measurement of the

relative inertia within the genes can help in to observe the significant trends in gene variations. Finally, ordering of *CCD* and *NCED* genes was carried out according to the position along the axis of significant inertia.

2.6. CAI in CCD and NCED family genes

The codon adaptation index (CAI) is a measureof relative adaptiveness towards the codon usage of genes which are highly expressed [53,54]. CAI values can be implicated to enable prediction of the expression level and evaluation of the variations in the codon usage pattern of a gene. The CAI value for a gene ranges between 0 and 1.0, respectively. A high value of CAI indicates a strong codon usage bias and probably a higher level of gene expression [53].

3. Results

3.1. Phylogeny reconstruction

The evolutionary relationship of the *CCD* and *NCED* genes were inferred using standard tree reconstruction methods. GTR + I + G was found to be the best fit model of evaluation for 79 mRNA-coding genes as reported by the Modeltest. ML and BI analyses showed equal weight for each position that resolved a single robust tree presented in Fig. 1. The other tree reconstructed method Neighbour-joining was given as (Supplementary Fig. 1). Most of the *CCD* and *NCED* genes from different plant species share high sequence homology with them. Hence they are clustered very closely representing a functional similarity between genes. For instance, in the *CCD* family, the entire *CCD1* cluster together, where earlier reports have also shown *CCD1* expression causing the distinguished emission of various volatile characteristics of fruits and flowers [55]. Likewise, *CCD7/8* was clustered together under a single clade formed as a sister group. The involvement of both the subclasses genes present in *Arabidopsis*, peas, and rice in generating the branching, and inhibition of strigolactone hormones formed in roots [56]. Similarly, *CCD4* groups and their functional cleavage of the carotenoid were involved in petal color formation. In the *NCED* family, all genes form a single cluster according to their sequence homology, and functionally cleave the double bond of 9-cis- violaxanthin or 9-cis-neoxanthin. They also catalyze the first step in ABA biosynthesis in most of the plant species [10].

3.2. Highly preferred codon of CCO genes

CCD and *NCED* genes RSCU value were calculated separately. The values showed the C and G-ending codons as greater in *CCD* compared to U and T-ending codons. The most often use of 31 codons was also seen in *CCD* gene family (RSCU > 1.0). Among them, we found CUC (leucine) as the most frequently used codons with the highest RSCU value of 1.54. Among the 31 codons, 13 and 7 were ending with C (Cytosine), G (Guanine) respectively. The inference of this analysis was that the predominance of codons ending with C was more than the nucleotide ending codons in *CCD* genes. In addition, the codons, ACA, and CCG, which codes for threonine and proline exhibited the least RSCU value. Further, the preferred codon RSCU value ranged from 1.02 to 1.54, as shown in Fig. 2A.

In the *NCED* gene family, the C and G-ending codons were seen as higher than U and T-ending codons. Twenty-six codons were highly preferred in the *NCED* gene family. Among the 26 codons, CUC (leucine) codon was observed with the highest RSCU values of 1.92, and on the other hand, AAA (lysine) and, AGA (arginine) showed the least RSCU values. The codons with an RSCU value ranged from 1.01 to 1.92 (Fig. 2B). Among the 26 codons, 14 and 7 ends with C (Cytosine) and G (Guanine) respectively. The observed results suggest that C ending codons ending with C were overall dominant in comparison to any other nucleotide ending codons in the *NCED* genes. The RSCU values of *CCD* and *NCED* genes are shown in Table 2. The RSCU values statistical

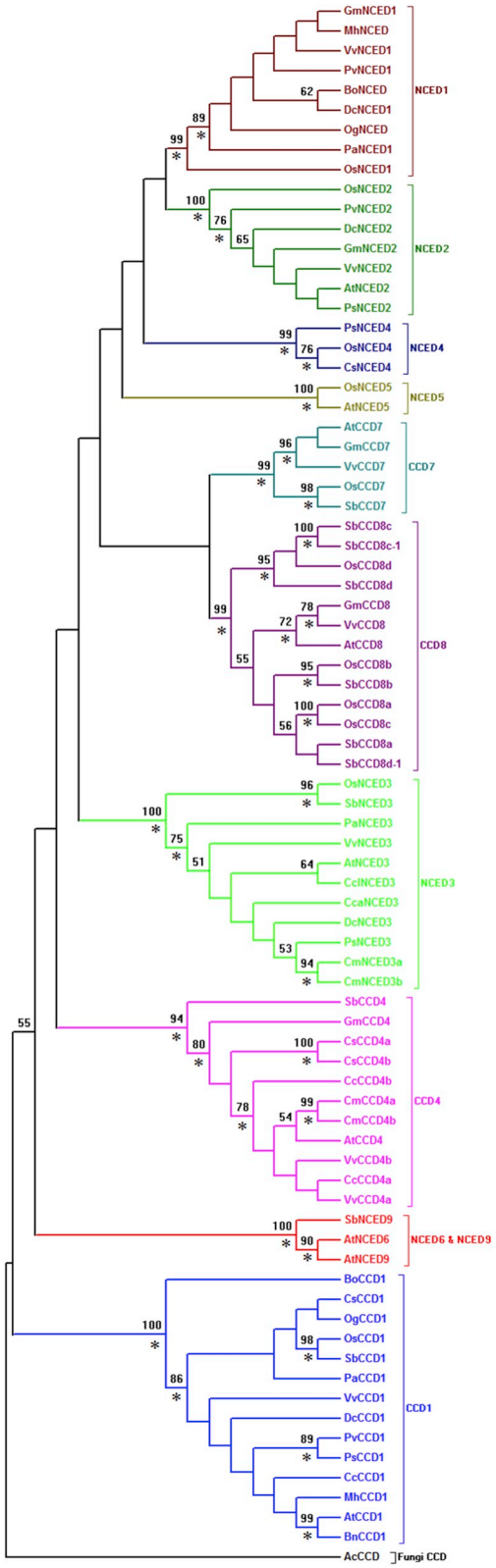

Fig. 1. Phylogenetic tree indicates the evolutionary relationship of the *CCO* gene using the Maximum Likelihood and Bayesian inference method. Both the methods show equal weight for each position that resolved a single robust tree. The bootstrap values greater than 50% are shown above the branch. Asterisks symbols indicate Bayesian posterior probabilities >70%.

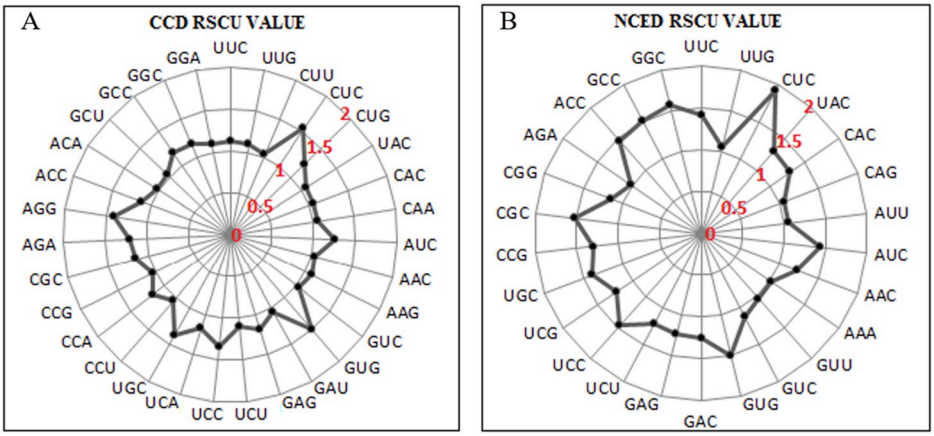

Fig. 2. Moderate to the high relative synonymous codon Usage value of *CCO* gene family, 2A. RSCU value of *CCD* (Carotenoid Cleavage Dioxygenase) gene family and 2B. RSCU value of *NCED* (Nine-Cis Epoxycarotenoid Dioxygenase) gene family.

significance was assessed using t-tests within the *CCD* and *NCED* genes ($P < 0.05$).

3.3. Variations in codon usage among the CCD and NCED gene family

The heterogeneity in codon usage among *CCD* and *NCED* genes from different plant species was determined using frequency distribution of gene indices, namely GC, GC_3s, and ENc values. Here the mean GC value was 0.51 ± 0.096, and the values range from 0.40 to 0.70 for the *CCD* family, as shown in Fig. 3A. In addition to variation in the GC content, GC_3s values range from 0.29 to 0.95, with an average of 0.54 ± 0.195. The ENc is another larger index with an average value of 50.09 ± 7.396 for the entire *CCD* family. The mean GC value in the *NCED* family was found to be 0.54 ± 0.091, and the values range from 0.40 to 0.70. The GC_3s values range from 0.30 to 1.00, with a mean of 0.60 ± 0.194. The mean ENc value was 48.76 ± 8.98. The distribution mean and standard deviation (SD) values in the *NCED* family genes are shown in Fig. 3B. *Arabidopsis* genes that were selected as a control for validation purpose showed a mean GC, GC_3s, and ENc value of 0.45 ± 0.03, 0.41 ± 0.06 and 53.26 ± 3.27, respectively. We found the differences in significant indices for *CCD* and *NCED* genes against the *Arabidopsis* genes (one-way ANOVA, two d.f., $P < 0.001$). These results strongly suggest the evidence of heterogeneity in *CCD* and *NCED* genes against the *Arabidopsis* genes (control).

3.4. Discriminating translational efficiency of CCD and NCED gene

Defining variations in the degree of codon usage biasness, the ENc values can also be associated with the translational efficiency among the *CCD* and *NCED* genes. A gene is said to contain optimal codons if the ENc value favors a lower range of 20 can be correlated with higher level of gene expression. Conversely, the ENc does not indicate the codons are more common than others, but it makes a relative estimation of the total disappearance from arbitrary synonymous codon choice [57]. The average of ENc values with SD were depicted in (Fig. 4) for finding the translational efficiency between *CCD* and *NCED* families. The mean ENc value of *CCD* gene (50.09 ± 7.396) was higher than that of the *NCED* genes representing a lower codon bias. However, a higher bias was seen in the individual ENc values for *CCD* genes from *Sorghum bicolor CCD4* (34.82), *Oryza sativa CCD8b* (32.03) and *S. bicolor CCD8b* (32.01). The *NCED* gene family with a mean value of 48.76 ± 8.979 indicates a good codon bias. The individual ENc values for *NCED* genes from *O. Sativa NCED5* (28.96), *O. Sativa NCED4* (31.13), *O. Sativa NCED3* (30.66), *O. Sativa NCED1* (32.10), *S. bicolor NCED*9 (28.81) and *S. bicolor NCED3* (29.74) were found to be highly biased with the ENc value < 35. The result shows the possibility of a slight bias between *CCD* and *NCED* genes. These can be correlated for better translational efficiency (Supplementary Table 2).

3.5. Mutation pressure in CCD and NCED gene

The mutation pressure in *CCD* and *NCED* gene was determined by examining the most crucial factor that affects the codon composition. The two types of the existing force which affect the codon composition could be mutation pressure and natural selection, as identified by the relationship between GC_3s and $GC_{1,2}$ content. The substantial positive correlation of GC_3s and $GC_{1,2}$ (Fig. 5) in *CCD* ($r = 0.847$, $p < 0.01$) and *NCED* ($r = 0.735$, $p < 0.01$) genes were identified using the Spearman's rank correlation analysis. The result suggests the occurrence of the base composition in codon positions as a result of mutation pressure.

We observed the correlation between GC_3s and ENc values for the validation of mutation pressure in *CCD* and *NCED* genes. The genes affected by G + C mutational pressure or any other factor could be identified based on the expected GC_3s shown in Fig. 6. The ENc – GC_3s plot indicates all the genes lying under the predictable GC_3s curve in both the CCD and NCED gene families. The ENc – GC_3s plot also specifies the occurrence of the codon usage bias by other factors in the *CCD* and *NCED* genes. Thus the correlation analysis confirms the presence of mutation pressure in *CCD* and *NCED* genes rather than the natural selection.

3.6. Mutation bias and gene function as the dominant factor

We investigated the relative synonymous codon usage values of 79 coding sequences using correspondence analysis (COA) for the characterization of selective forces operating on *CCD and NCED* genes. The greatest difference was identified in the first axis of 79 genes on the basis of correspondence analysis in all the four top axes. The first axis contains 58.62% differences, and the next three axes contain 19.76%, 5.89%, and 8.56% within the gene. The observed inference is that Axis-I was the principal axis, and it signified as key factor which helped in the clarification of the gene set. The primary source of the difference in the complete *CCD* and *NCED* gene can be located with the assistance of correlation analysis utilising the principal axis. The Spearman's rank correlation was applied to both *CCD* and *NCED* families individually considering the values of Axis-I with regard to major indices. The correlation values of *CCD* versus major indices for ENc, GC and GC_3s values are $r = 0.359$, $r = -0.889$ and $r = -0.856$ $p < 0.01$ respectively. Similarly the ENc, GC and GC_3s values for NCED are $r = 0.487$, $r = 0.941$ and $r = 0.835$ $P < 0.01$ respectively. A negative correlation was seen in axis one against GC, GC_3s in *CCD* and *NCED* showed a strong positive correlation with GC and GC_3s. Fig. 7 indicates the closer relationship of

Table 2
Cumulative codon usage of *CCD* and *NCED* genes family.

AA	Codon	N	RSCU	Codon	N	RSCU
Phe	UUU	571	0.89	UUU	319	0.59
	UUC	714	1.11*	UUC	770	1.41*
Leu	UUA	163	0.50	UUA	210	0.71
	UUG	359	1.10*	UUG	309	1.05*
	CUU	343	1.05*	CUU	281	0.96
	CUC	500	1.54*	CUC	564	1.92*
	CUA	191	0.59	CUA	129	0.44
	CUG	397	1.22*	CUG	271	0.92
Tyr	UAU	394	0.93	UAU	215	0.69
	UAC	450	1.07*	UAC	406	1.31*
His	CAU	346	0.95	CAU	232	0.70
	CAC	381	1.05*	CAC	429	1.30*
Gln	CAA	257	1.06*	CAA	306	0.94
	CAG	226	0.94	CAG	346	1.06*
Ile	AUU	375	0.95	AUU	351	1.06*
	AUC	497	1.26*	AUC	472	1.43*
	AUA	313	0.79	AUA	170	0.51
Met	AUG	610	1.00	AUG	491	1.00
Asn	AAU	446	0.96	AAU	350	0.78
	AAC	479	1.04*	AAC	552	1.22*
Lys	AAA	608	0.92	AAA	569	1.01*
	AAG	714	1.08*	AAG	563	0.99
Val	GUU	436	0.97	GUU	423	1.04*
	GUC	466	1.04*	GUC	457	1.13*
	GUA	224	0.50	GUA	137	0.34
	GUG	667	1.49*	GUG	607	1.50*
Asp	GAU	737	1.06*	GAU	442	0.75
	GAC	652	0.94	GAC	733	1.25*
Glu	GAA	587	0.81	GAA	510	0.76
	GAG	854	1.19*	GAG	832	1.24*
Ser	UCU	282	1.11*	UCU	310	1.21*
	UCC	341	1.35*	UCC	379	1.47*
	UCA	296	1.17*	UCA	239	0.93
	UCG	232	0.92	UCG	313	1.22*
	AGU	124	0.49	AGU	104	0.40
	AGC	245	0.97	AGC	197	0.77
Cys	UGU	101	0.62	UGU	87	0.61
	UGC	225	1.38*	UGC	198	1.39*
Trp	UGG	247	1.00	UGG	195	1.00
Pro	CCU	422	1.04*	CCU	324	0.82
	CCC	316	0.78	CCC	411	1.04
	CCA	470	1.16*	CCA	335	0.85
	CCG	411	1.02*	CCG	515	1.30*
Arg	CGU	164	0.75	CGU	146	0.80
	CGC	255	1.17*	CGC	278	1.52*
	CGA	107	0.49	CGA	97	0.53
	CGG	213	0.94	CGG	209	1.15*
	AGA	265	1.21*	AGA	185	1.01*
	AGG	307	1.41*	AGG	180	0.99
Thr	ACU	297	0.97	ACU	233	0.84
	ACC	347	1.13*	ACC	409	1.47*
	ACA	312	1.02*	ACA	243	0.87
	ACG	271	0.88	ACG	228	0.82
Ala	GCU	420	1.04*	GCU	394	0.96
	GCC	478	1.19*	GCC	627	1.52*
	GCA	382	0.95	GCA	224	0.54
	GCG	331	0.82	GCG	404	0.98
Gly	GGU	399	0.84	GGU	335	0.84
	GGC	563	1.18*	GGC	630	1.58*
	GGA	531	1.11*	GGA	332	0.83
	GGG	412	0.87	GGG	302	0.76

AA: Three letter amino acid code; N: Total number of codon; RSCU: Relative synonymous codon Usage; Bold and Asterisks: Highly preferred codons.

nearly all the *CCD* and *NCED* genes to a family from differential species related by an underlying codon usage. The closer relationship between the codon usage of functionally homologous gene values is shown by plotting the first two axes. From the observed results we conclude that the major common factor shaping the codon usage in the *CCD* and *NCED* genes is nucleotide composition mutation bias.

3.7. Relative adaptiveness of codon usage in CCD and NCED

The codon adaptation index (CAI) is the simplest and the most efficient measure utilized in the evaluation of the codons bias that finds favor in the genes expression level. CAI measures the expression standards of a gene on a reference gene set. The set of reference sequences used in the calculation of CAI values in this study were the genes that encode ribosomal proteins [34]. The cross-species comparison was made with *Saccharomyces cerevisiae*. In the *CCD* gene family their significance correlation with the gene expression level was assessed by the CAI against the ENC values ($r = -0.721$), GC_3s ($r = 0.651$) and GC ($r = 0.642$) at $P < 0.01$ in Fig. 8A. In *NCED* family, all the three indices ENc, GC_3s and GC have significant correlation between the CAI values and the respective indices ($r = -0.604$, $r = 0.636$ and $r = 0.643$, $P < 0.01$) in Fig. 8B. All the above correlations were statistically significant, indicating codon usage in *CCD* and *NCED* family affected by the gene expression.

4. Discussion

Codon usage bias has been observed from most primitive unicellular microorganisms to complex eukaryotes. In general, the codon bias may be considered as a balance between natural selection and mutational bias phenomenon. Considerable research work done in the past explains the various phenomenon that controls in different organisms. However, plant codon usage pattern shows a closer resemblance to higher eukaryotes than unicellular organisms. This may be due to a higher preference to GC content much similar to *homo sapiens* shown by plants [38]. A consensus phylogenetic tree was obtained by the use of all the three methods that showed a closer relationship to the subclass genes for *CCD* and *NCED*. These relationships documented previously were also reconciled with our tree [58–61]. This closeness of the *CCD* genes was observed at a structure and function level [62]. Moreover, there was a functional closeness reflection within these genes wherein functionally similar genes from different species formed together as one group. Interestingly, the constructed phylogeny also implied that *CCD7/8* having observed similar evolutionary trends [63] with both the genes showing a common function with branching inhibition.

In this study, we have analyzed the phylogenetic relationship of *CCD* and *NCED* genes through standard models of tree building such as Neighbour-Joining (NJ), Maximum likelihood (ML) and Bayesian inference (BI) methods. In extension to the above methods, the codon and nucleotide usage in every *CCD* and *NCED* genes with other significant related factors were computed. The adequate number of codons (ENc) and relative synonymous codon usage (RSCU) along with significant indices such as GC, GC1, GC2, and GC3 were measured to enable reading the codon usage bias in the gene set. Moreover, the primary source of variation was identified from the correlation analysis of these indices with the correspondence analysis (COA). Finally, analysis of the expression patterns of these gene sets was made using the codon adaptation index (CAI). The methods mentioned above find extensive use with consideration as the most excellent statistical method for investigating codon biasness for the selected dataset. The key factors that influence the codon usage 79 coding sequences of plant *CCD* and *NCED* genes were determined and analyzed using the software's such as Codon W version 4.3, GCUA version 1.2, and Graphpad prism 5. A similar codon usage bias analysis has been performed in various species earlier [64,65] while no investigation till date has been made in the *CCD* and *NCED* genes. Hence, the findings of this research will enable understanding of the molecular evolution of these genes.

The codon usage with its corresponding indices indicates the existence facts. For instance, CUC (leucine) codon in high frequency is used in both *CCD* and *NCED* genes. In contrast, the control gene set (*Arabidopsis*) shows a strong preference for AGA (arginine) codon with RSCU value of 2.04 (Supplementary Table 3). Highly preferred CUC can be correlated with the common RP65 domain with cleavage activity within

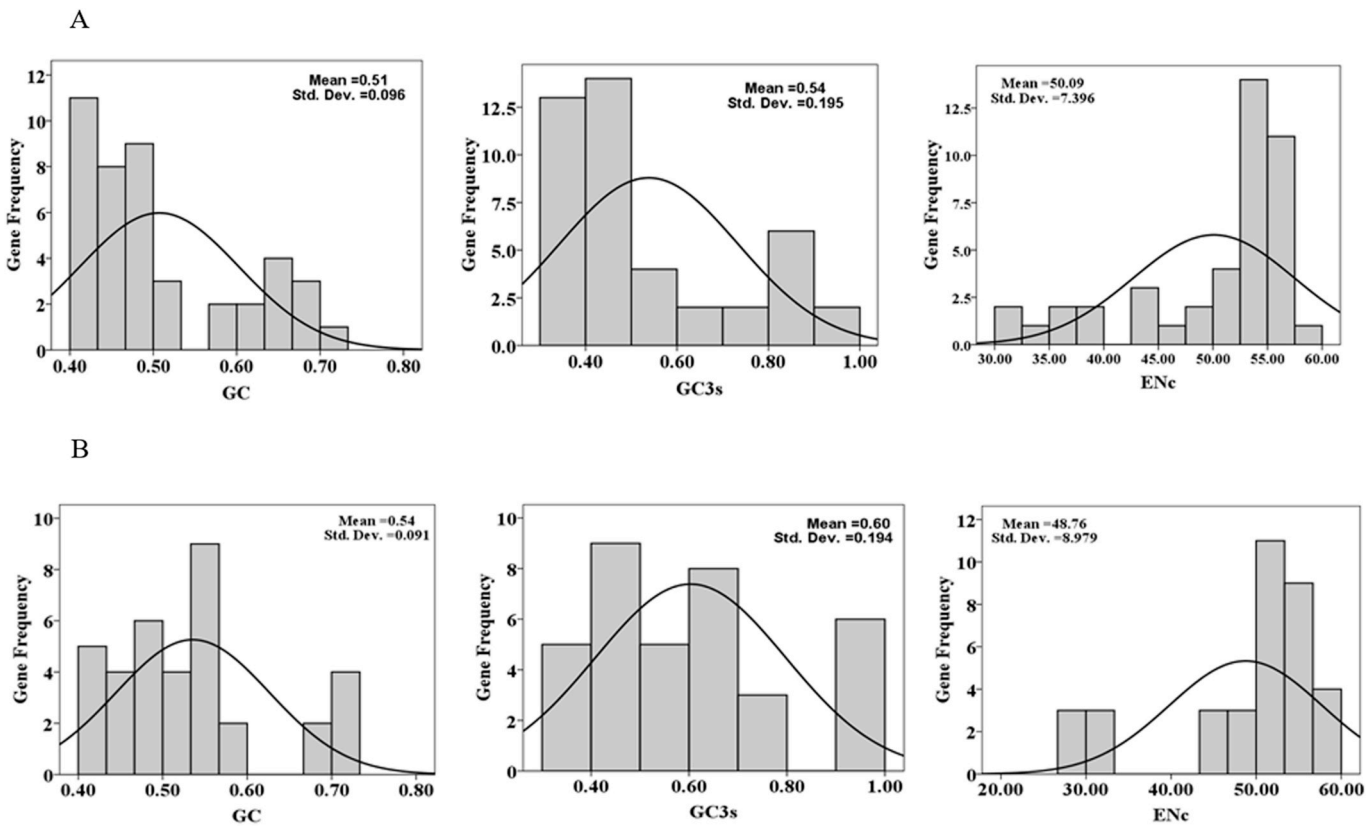

Fig. 3. Frequency distribution of GC, GC3, and ENc for both *CCO* gene families, 3A. Graph illustrating the mean and standard deviation of GC, GC3, and ENc for *CCD* gene family and 3B. Illustrates the mean and standard deviation of GC, GC3, and ENc for *NCED* gene family.

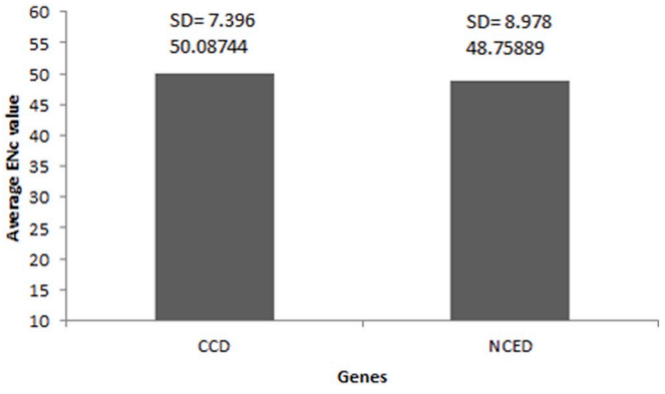

Fig. 4. Average ENc values and its analogous SD value for each *CCD* and *NCED* gene group among different plant species.

the *CCO* family. However, the variation in the leading indices could be attributed to the structural difference with function-specific roles. Next, the average GC content of the entire genes and at the third codon positions was slightly higher in *NCED* gene family than in *CCD* (Table 3), suggesting a possible variation in the codon composition in *NCED* and *CCD* genes. Further, the total GC content has shown slight variations in *CCD* and *NCED*. These findings suggest the likelihood of *NCED* genes displaying stability better than the *CCD* genes. Consequently, the gene index GC_3s is higher on an average of 0.07 (>0.05 cutoff) for *NCED* family than *CCD* family. We have observed similar findings where most of the eukaryotic species tends to show preference to the GC content [66](Romero et al., 2003). Many earlier studies have reported mutation stress as the major reason behind the preference to GC in monocot plants during codon usage bias [67-68] . This research work shows *NCED* genes with GC3s richness and under selective pressure. This would provide an additional transcriptional advantage [69] for the *NCED* genes.

On the other hand, the mean value of ENc is low in *NCED* genes, reflecting the selective translational efficiency more than *CCD* genes. However, in both *CCD* and *NCED* families, few genes were highly biased with ENc <35. As inferred from the result, the highly biased genes for *CCD* and *NCED* were from the same plant species which were also monocots. Similarly, a significant positive correlation was observed in codon composition for both *CCD* and *NCED* families. A study by Zhong et al (2012); reported a similar kind of ENc value showing mutational stress in the RNA viruses rather than a natural selection during codon bias [70]. This suggests the occurrence of observed patterns of base composition in codon positions as a result of mutation pressure, which is not positioned specific, as in the case of natural selection. All the genes were just below the expected GC3s curve in ENc-plots for *CCD* and *NCED,* indicating that, apart from the compositional constraints, other factors might have influenced codon usage variation. The COA analysis stated that the discrimination on the first axis in the *CCD* family resulted mainly from differences in the overall GC composition. As opposed to the negative correlation of axis one against GC, GC_3s in *CCD*, *NCED* showed a strong positive correlation with GC and GC_3s. Although numerous methods help in understanding the codon bias, the most adaptive method would be the CAI method referred to in many studies [71,72]. In this study, the relationship between the gene expression levels assessed by the CAI against three major indices was statistically significant, indicating the impact of gene expression on codon usage in *CCD* and *NCED* families. The numerical values of the codons usage have been subjected to various correlation and statistical studies. Only a few may have been biased towards certain outcomes, as the mutational pressure is the key for the set of genes. However, natural selection could also lead to the distinction between the genes from different species. Also, the entire dataset for the *CCD* and *NCED* genes for the within the

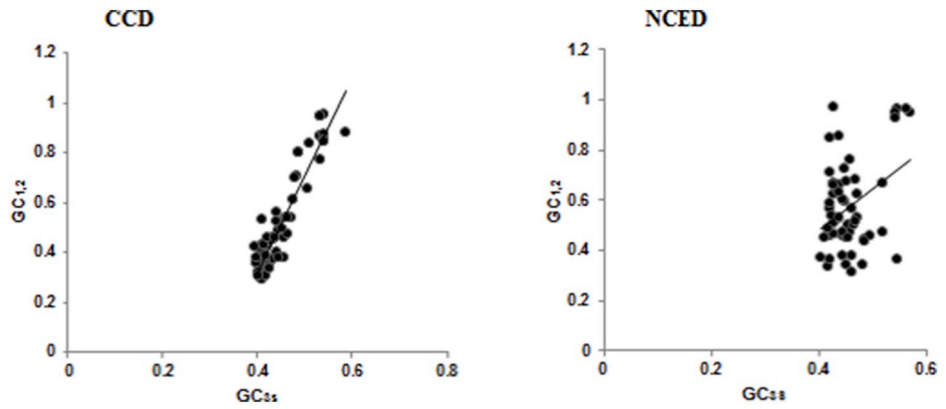

Fig. 5. Correlation between G + C content at the first and second codon position ($GC_{1,2}$) with synonymous third positions (GC_3S) in both *CCD* and *NCED* gene family.

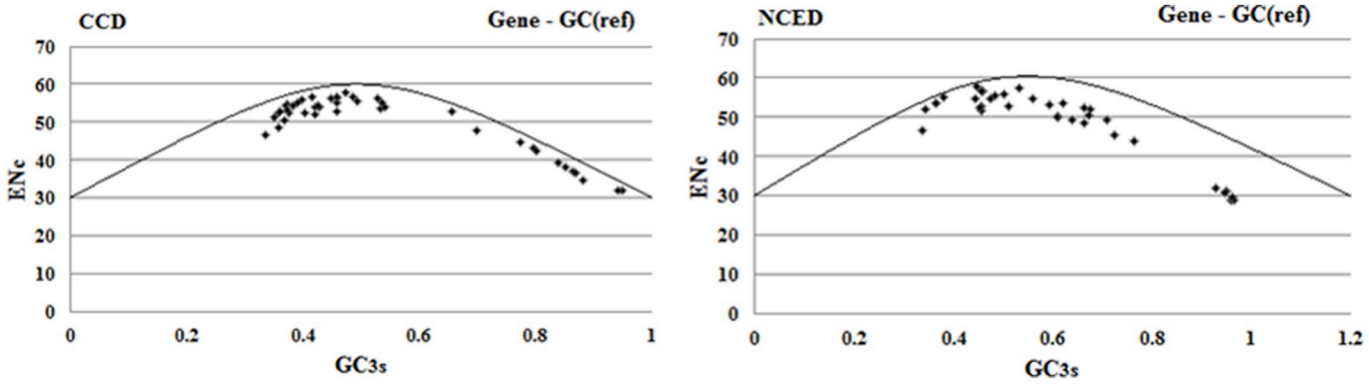

Fig. 6. The ENc value for each of *CCD* and *NCED* genes from 17 different plants examined and plotted against the corresponding G + C content at the synonymously variable third position (GC_3S).

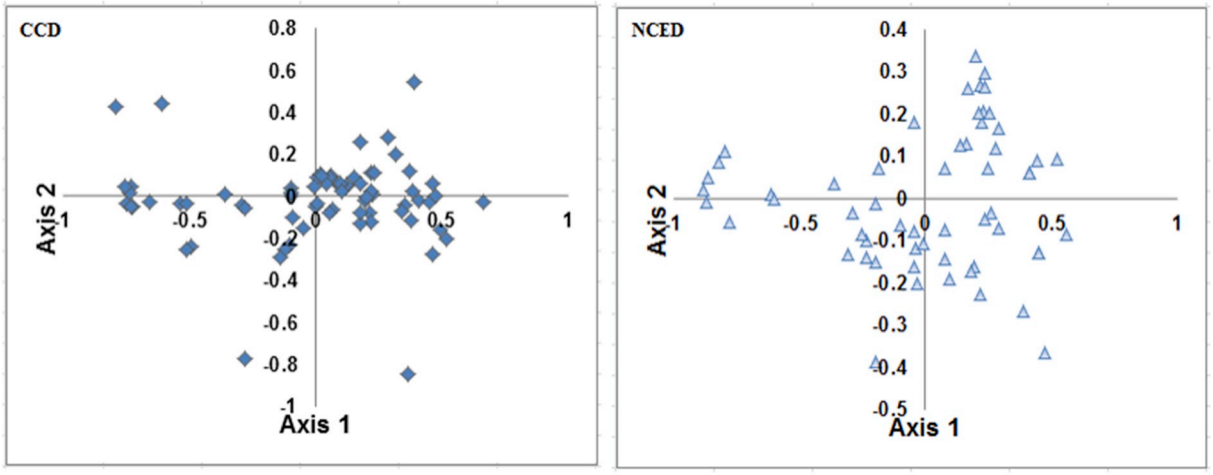

Fig. 7. A plot of the values of the two most principal axes generated by the COA of each gene.

carotenoid family of genes were missing in the dataset. Further, the partially coding sequences were also made negated. Similarly, under the codon usage study, the model organism used is *Arabidopsis thaliana* was utilized as the reference codon table in computing the dataset. We have made extensive use of multiple ways for the study of codon usage bias. However, the major limitation would be non-availability of the complete gene sequences of *NCED* and *CCD*. Despite the availability of a large dataset and massive deposition of the gene sequences, lack of high-quality datasets limit the values and integrity of the current research. This limitation can be overcome by sequencing all the missing genes within this family. The increase in reliable datasets and improved databases along with advanced state-of-the-art bioinformatic tools would enable understanding the molecular evolution in the near future.

5. Conclusion

To conclude, this in-depth investigation of codon usage bias in *CCO* genes has shed light on a broader sense and brought out the most common and discriminative features at the molecular level on the function in different plant species. The inference from this study is that

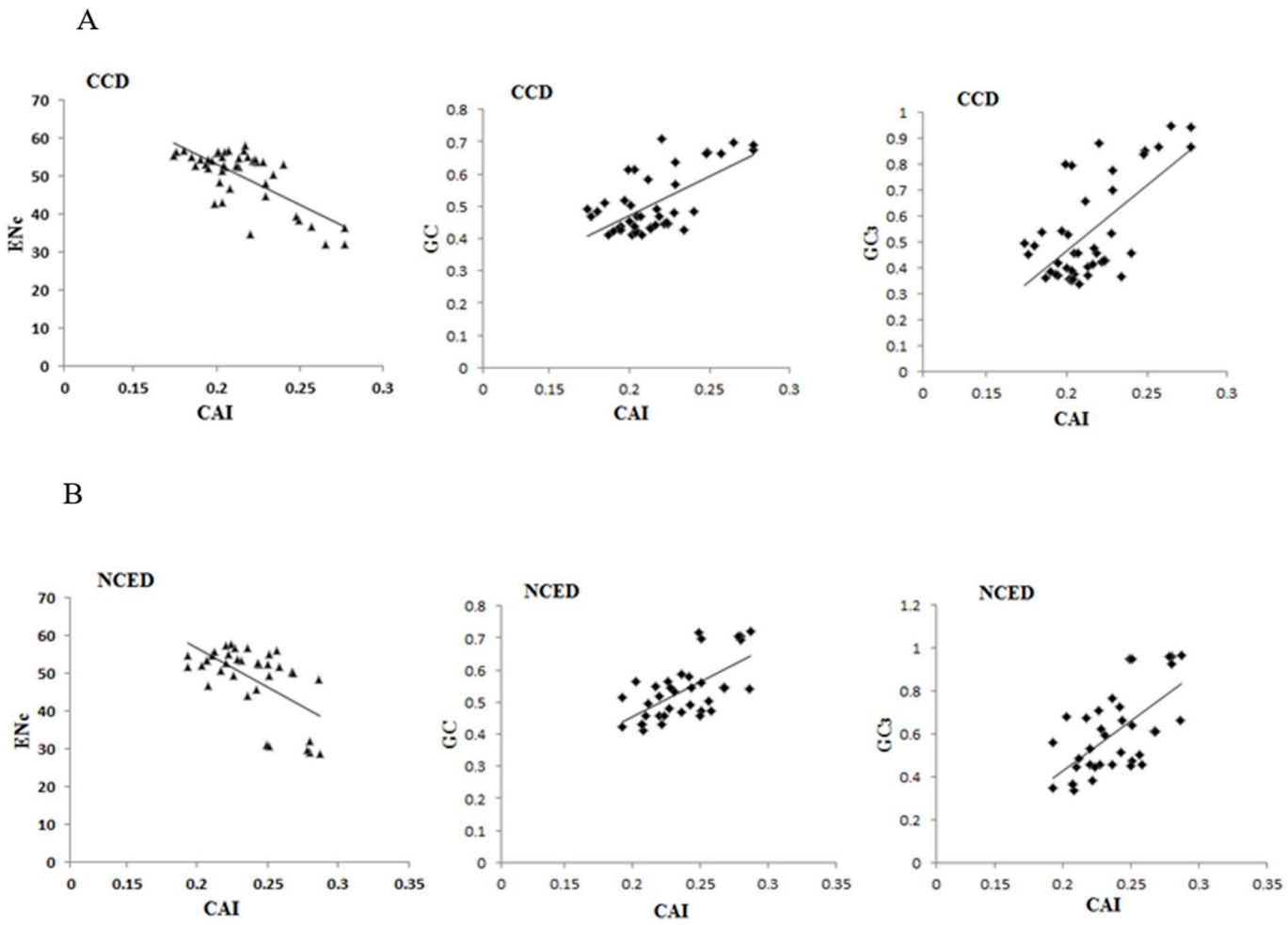

Fig. 8. Codon adaptation index (CAI) calculate the expression of a gene from 17 different plant species, 8A. Shows the correlation of CAI with major indices for *CCD* genes and 8B. Illustrates the correlation of CAI with major indices for *NCED* gene.

Table 3
Average base composition of GC1, GC2 and GC3 in *CCO* and *Arabidopsis thaliana* (control) gene.

S.No	GC1s	GC2s	GC3s
CCD	0.517268	0.376527	0.504569
NCED	0.528278	0.390038	0.573948
Arabidopsis nuclear gene (Control)	0.61353	0.507435	0.631926

the mutation bias due to nucleotide composition is the significant common factor carving the codon usage in the *CCD* and *NCED* genes. Knowledge of the codon usage may also help to increase the production of the desired range of important apocarotenoid. However, for species-specific pigmentation products, additional experimental studies such as gene expression studies are required. The current research will also benefit further transgenic study and build a genetic profile for future research in the production of economically valuable pigments.

Author's contributions

RP, SP, JFPD, GPDC, and RS involved in the design, acquisition of data, and interpretation of the data. JFPD, GPDC, and RS supervised the entire study and drafting the manuscript. The manuscript was reviewed and approved by all the authors RP, SP, JFPD, GPDC, and RS. All the authors have equally contributed.

Conflicts of interest

We have no conflicts of interest to disclose.

Acknowledgment

We express our sincere gratitude to the Science and Engineering Research Board – Department of Science and Technology, New Delhi, India for the support extended through the project [SR/FT/LS-75/2011]. The authors are grateful to VIT University management for their constant support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2019.103449.

References

[1] M. Ibdah, Y. Azulay, V. Portnoy, B. Wasserman, E. Bar, A. Meir, et al., Functional characterization of CmCCD1 a carotenoid cleavage dioxygenase from melon, Phytochemistry (Oxf.) 67 (2006) 1579–1589.
[2] R. Priya, F. Prabhu Doss, R. Siva, Gene expression prediction and hierarchical clustering analysis of plant CCD genes, Plant Mol. Biol. Report. 34 (2016) 618–627.
[3] J. Yabuzaki, Carotenoids Database. http://Carotenoiddb.jp/2018.
[4] E.M. Auldridge, A. Block, T.J. Vogel, C. Dabney-Smith, I. Mila, M. Bouzayen, et al., Characterization of three members of the Arabidopsis carotenoid cleavage dioxygenase family demonstrates the divergent roles of this multifunctional enzyme family, Plant J. 45 (2006) 982–993.

[5] F. Bouvier, J.C. Isner, O. Dogbo, B. Camara, Oxidative tailoring of carotenoids: a prospect towards novel functions in plants, Trends Plant Sci. 10 (2005) 187–194.
[6] R. Siva, Status of natural dyes and dye-yielding plants in India, Curr. Sci. 92 (2007) 916–925.
[7] M. Sankari, H. Hridya, A. Amirtha, S. Babu, R. Rivera Madrid, C. George Priya Doss, et al., Identifying a carotenoid cleavage dioxygenase 4a gene and its efficient Agrobacterium mediated genetic transformation in Bixa orellana. L. Appl. Biochem. Biotechnol. 179 (2016) 697–714.
[8] R. Siva, G.J. Mathew, A. Venkat, C. Dhawan, An alternative tracking dye for gel electrophoresis, Curr. Sci. 94 (2008) 765–767.
[9] R. Siva, F. Prabhu Doss, K. Kundu, V.S.V. Satyanarayana, V. Kumar, Molecular characterization of bixin-An important industrial product, Ind. Crops Prod. 32 (2010) 48–53.
[10] S.H. Schwartz, B.C. Tan, D.A. Gage, J.A. Zeevart, D.R. McCarty, Specific oxidative cleavage of carotenoids by VP14 of maize, Science 276 (1997) 1872–1874.
[11] Z. Sun, J. Hans, M.H. Walter, R. Matusova, J. Beekwilder, F.W.A. Verstappen, et al., Cloning and characterization of a maize carotenoid cleavage dioxygenase (ZmCCD1) and its involvement in the biosynthesis of apocarotenoids with various roles in mutualistic and parasitic interactions, Planta 228 (2008) 789–801.
[12] A. Ohmiya, Carotenoid cleavage dioxygenases and their apocarotenoid products in plants, Plant Biotechnol. 26 (2009) 351–358.
[13] D.P. Kloer, G.E. Schulz, Structural and biological aspects of carotenoid cleavage, Cell. Mol. Life Sci. 63 (2006) 2291–2303.
[14] T. Borowski, M.R. Blomberg, P.E. Siegbahn, Reaction mechanism of apocarotenoid oxygenase (ACO): a DFT study, Chemistry 14 (2008) 2264–2276.
[15] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, Codon catalog usage and the genome hypothesis, Nucleic Acids Res. 8 (1980) 49–62.
[16] P.M. Sharp, E. Cowe, D.G. Higgins, Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, and Homo sapiens: a review of the considerable within-species diversity, Nucleic Acids Res. 16 (1988) 8207–8211.
[17] J. Ma, M.N. Nguyen, J.C. Rajapakse, Gene classification using codon usage and support vector machines IEEE/ACM Trans, Comput. Biol. Bioinf. 6 (2009) 134–143.
[18] A. Fuglsang, Strong associations between gene function and codon usage, Apmis 111 (2003) 843–847.
[19] Q. Liu, S. Dou, Z. Ji, Q. Xue, Synonymous codon usage and gene function are strongly related in Oryza sativa, Biosystems 80 (2005) 123–131.
[20] L. Duret, tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes, Trends Genet. 16 (2002) 287–289.
[21] P. Shah, A.M. Gilchrist, Effect of correlated tRNA abundances on Translation errors and evolution of codon usage bias, PLoS Genet. 6 (9) (2010), e1001128.
[22] U. Roymondal, S. Das, S. Sahoo, Predicting gene expression level from relative codon usage bias: an application to Escherichia coli genome, DNA Res. 16 (2009) 13–30.
[23] A.W. Carrie, S. Yu, H. Johannesson, Evolution of synonymous codon usage in Neurospora tetrasperma and Neurospora discreta, Genome. Biol. Evol. 3 (2011) 332–343.
[24] T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms, Mol. Biol. Evol. 2 (1985) 13–34.
[25] R.M. Goetz, A. Fuglsang, Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli, Biochem. Biophys. Res. Commun. 327 (2005) 4–7.
[26] M. Bulmer, The selection–mutation–drift theory of synonymous codon usage, Genetics 129 (1991) 897–907.
[27] W.J. Gu, T. Zhou, J.M. Ma, X. Sun, Z.H. Lu, The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens, Biosystems 73 (2004) 89–97.
[28] H. Akashi, Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy, Genetics 136 (1994) 927–935.
[29] N. Stoletzki, A. Eyre-Walker, Synonymous codon usage in Escherichia coli: selection for translational accuracy, Mol. Biol. Evol. 24 (2007) 374–381.
[30] S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro, H.H. McAdams, Codon usage between genomes is constrained by genome-wide mutational processes, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 3480–3485.
[31] Y. Zheng, W.M. Zhao, H. Wang, Y.B. Zhou, Y. Luan, M. Qi, et al., Codon usage bias in Chlamydia trachomatis and the effect of codon modification in the MOMP gene on immune responses to vaccination, Biochem. Cell Biol. 85 (2007) 218–226.
[32] K. Lin, Y. Kuang, J.S. Joseph, P.R. Kolatkar, Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics, Nucleic Acids Res. 30 (2002) 2599–2607.
[33] H. Chiapello, F. Lisacek, M. Caboche, A. Henaut, Codon usage and gene function are related in sequences of Arabidopsis thaliana, Gene 209 (1998) 1–38.
[34] C. Xu, X. Cai, Q. Chen, H. Zhou, Y. Cai, A. Ben, Factors affecting synonymous codon usage bias in chloroplast genome of oncidium gower ramsey, Evol. Bioinf. 7 (2011) 271–278.
[35] M. Bailly-Bechet, A. Danchin, M. Iqbal, M. Marsili, M. Vergassola, Codon usage domains over bacterial chromosomes, PLoS Comput. Biol. 2 (2006) e37.
[36] H.M.W. Emily, K.S. David, R. Raul, P. Malik, L.M.P. Leo, Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus, BMC Evol. Biol. 10 (2010) 253.
[37] S. Sahooa, S.S. Das, R. Rakshit, Codon usage pattern and predicted gene expression in Arabidopsis thaliana, GeneX 2 (100012) (2019).
[38] E.E. Murray, J. Lotzer, M. Eberle, Codon usage in plant genes, Nucleic Acids Res. 17 (2) (1989) 477–498.
[39] W. Xu, T. Xing, M. Zhao, X. Yin, G. Xia, M. Wang, Synonymous codon usage bias in plant mitochondrial genes is associated with intron number and mirrors species evolution, PLoS One 0 (6) (2015), e01315081.
[40] J.D. Thompson, D.G. Higgins, T.J. Gibson, W. Clustal, Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[41] S. Capella-Gutiérrez, J.M. Silla-Martínez, T. Gabaldon, trimAI: a tool for automated alignment trimming in large-scale phylogenetic analyses, Bioinformatics 25 (2009) 1972–1973.
[42] K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0, Mol. Biol. Evol. 24 (2007) 1596–1599.
[43] D. Posada, ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online, Nucleic Acids Res. 34 (2006) W700–W703.
[44] S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, PHYML Online – a web server for fast maximum likelihood-based phylogenetic inference, Nucleic Acids Res. 33 (2005) W557–W559.
[45] J.P. Huelsenbeck, F. Ronguist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics 17 (2001) 754–755.
[46] P.M. Sharp, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1986) 28–38.
[47] J.F. Peden, Analysis of Codon Usage, PhD' Thesis, University of Nottingham, 1999.
[48] E.H. Wong, D.K. Smith, R. Rabadan, M. Peiris, L.L. Poon, Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus, BMC Evol. Biol. 10 (2010) 253.
[49] F. Wright, The 'effective number of codons' used in a gene, Gene 87 (1990) 23–29.
[50] R. Hershberg, D.A. Petrov, General rules for optimal codon choice, PLoS Genet. 5 (2009), e1000556.
[51] M.C. Angellotti, S.B. Bhuiyan, G. Chen, X.F. Wan, Codon usage bias analysis within and across genomes, Nucleic Acids Res. 35 (2007) W132–W136.
[52] M.J. Greenacre, Theory and Applications of Correspondence Analysis, Academi Press London, UK, 1984, p. 364.
[53] P.M. Sharp, W.H. Li, The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Res. 5 (1987) 1281–1295.
[54] D.C. Shields, P.M. Sharp, Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases, Nucleic Acids Res. 15 (19) (1987) 8023–8040.
[55] A.J. Simkin, S.H. Schwartz, M. Auldridge, M.G. Taylor, H.J. Klee, The tomato carotenoid cleavage dioxygenase 1 genes contribute to the formation of the flavor volatiles beta-ionone, pseudoionone, and geranylacetone, Plant J. 40 (2004) 882–892.
[56] R. Vallabhaneni, L.M.T. Bradbury, E.T. Wurtzel, The carotenoid dioxygenase gene family in maize, sorghum, and rice, Arch. Biochem. Biophys. 504 (2010) 104–111.
[57] O.U. Araxi, D.H. Laurence, Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection, Genetics 159 (2001) 1191–1199.
[58] J.G. Lashbrooke, P.R. Young, S.J. Dockrall, K. Vasanth, M.A. Vivier, Functional characterization of three members of the Vitis vinifera L. carotenoid cleavage dioxygenase gene family, BMC Plant Biol. 13 (2013) 156.
[59] H. Cui, Y. Wang, S. Qin, Genome wide analysis of carotenoid cleavage dioxygenase in unicellular and filamentous cyanobacteria, Comp. Funct. Genom. (2012), 164690, 13.
[60] R. Priya, R. Siva, Phylogenetic and functional diverge in plant nine-cis epoxy carotenoid cleavage dioxygenase gene family, J. Plant Res. 128 (2015) 519–534.
[61] R. Priya, H. Hridya, C. Soundarya, G. Somasundari, C. George Priya Doss, P. Sneha, et al., Astaxanthin biosynthetic pathway: molecular phylogenies and evolutionary behaviour of Crt genes in eubacteria, Plant Gene 8 (2016) 32–41.
[62] R. Priya, R. Siva, Phylogenetic analysis and evolutionary studies of plant carotenoid cleavage dioxygenase gene, Gene 548 (2) (2014) 223–233.
[63] R.K. Wang, J.J. Lu, G.N. Xing, J.Y. Gai, T.J. Zhao, Molecular evolution of two consecutive carotenoid cleavage dioxygenase genes in strigolactone biosynthesis in plants, Genet. Mol. Res. 10 (2011) 3664–3673.
[64] A. Kawabe, N.T. Miyashita, Patterns of codon usage bias in three dicot and four monocot plant species, Genes Genet. Syst. 278 (5) (2003) 343–352.
[65] P. Mazumdar, R.B. Othman, K. Mebus, N. Ramakrishnan, J.A. Harikrishna, Codon usage and codon pair patterns in non-grass monocot genomes, Ann. Bot. 120 (6) (2017) 893–909.
[66] H. Romero, A. Zavala, V. Musto, G. Bernardi, The influence of translational selection on codon usage in fishes from the family Cyprinidae, Gene 317 (2003) 141–147.
[67] L. Wang, M.J. Roossinck, Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants, Plant Mol. Biol. 61 (2006) 699–710.
[68] H.-C. Wang, D.A. Hickey, Rapid divergence of codon usage patterns within the rice genome, BMC Evol. Biol. 7 (2007) 1–10.
[69] E. Elhaik, T.V. Tatarinova, GC3 biology in eukaryotes and prokaryotes, in: DNA Methylation —From Genomics to Technology, 2012, pp. 55–68.

[70] Q. Zhong, W. Xu, Y. Wu, H. Xu, Patterns of synonymous codon usage on human metapneumovirus and its influencing factor, J. Biomed. Biotechnol. 2012 (2012) 7.
[71] A. Carbone, A. Zinovyev, F. Képès, Codon adaptation index as a measure of dominating codon bias, Bioinformatics 19 (16) (2003) 2005–2015.
[72] J.M. Comeron, M. Aguadé, An evaluation of measures of synonymous codon usage bias, J. Mol. Evol. 47 (3) (1998) 268–274.