

PATTERNS OF CODON USAGE BIAS IN CARBONIC ANHYDRASE REVEAL ADAPTIVE EVOLUTION IN C3 AND C4 PHOTOSYNTHETIC PATHWAYS

Muhammad Hamza Gul, Habib Ullah¹, Sundas Rafiq², Yusra Akhtar³, Sobia Khattak⁴, Alishba Khan⁵, Muhammad Inam ul Haq⁶, Siddiq ur Rehman⁷, Noor ul Haq^{*8}

^{1,2,3,4,5,6,7,*8}Department of Computer Science and Bioinformatics, Khushal Khan Khattak University, Karak 27200, Khyber-Pakhtunkhwa, Pakistan

DOI: <https://doi.org/10.5281/zenodo.21975006>

Keywords

Codon Usage Bias,
Carbonic Anhydrase, CA,
C3, C4, Photosynthesis
Evolution

Article History

Received: 02 April 2026

Accepted: 12 May 2026

Published: 31 May 2026

Copyright @Author

Corresponding Author: *

Noor ul Haq

noorqu@gmail.com,

noorulhaq@kkuk.edu.pk

Abstract

Photosynthesis is an essential biological process that supports life on Earth by transforming carbon dioxide and water into glucose and oxygen while this process involves complex biochemical pathways. Carbonic anhydrase (CA) is a crucial enzyme in this process that convert carbon dioxide and water into bicarbonate and protons thus ensures an adequate CO₂ supply to the Calvin cycle. This study represents the nucleotide composition of CA gene in C3 and C4 plants to understand codon usage patterns at different positions within the genes. Computational analyses such as codon adaptation index (CAI) and relative synonymous codon usage (RSCU) were used to analyze codon usage bias. Correspondence Analysis (COA) was utilized to explore synonymous codon usage variance that reveals evolutionary trends and genetic linkages among the plant species. Neutrality Plot Mapping and Effective Number of Codons (ENC) analysis were conducted to check the impact of natural selection and mutational pressure on codon usage. The study highlights the relations between GC contents and codon bias that offers the insights into the expression and adaptability of CA enzymes across different plant species and their evolutionary relationships. This study offers valuable overview of the complex forces driving codon usage patterns in the carbonic anhydrase gene across various plants. The findings suggest that both random genetic changes and natural selection, particularly favoring specific codon terminations, contribute to the observed codon bias.

Introduction

During Photosynthesis, green plants consolidate water (H₂O) and Carbon Dioxide (CO₂) using light energy to generate Oxygen (O₂) and glucose (C₆H₁₂O₆) is omitted as by product of the process. In photosynthesis, carbon dioxide, water and light energy are the reactants (Gest 2002). Light dependent reaction is the step of photosynthesis in which light energy is changed into NADPH and ATP. Photo system I and II consume the light energy from sunlight to generate ATP and NADPH. This reaction occurs in the thylakoid

membrane. During this cycle water particles are parted into oxygen (O₂), hydrogen (H) and free electron. Another step of photosynthesis called as dark reaction; is a light-free reaction and does not need light energy directly. This cycle happens in the stroma of the chloroplast and converts CO₂ into starches. The CO₂ and H₂O particles are changed over into C₆H₁₂O₆ molecules in this cycle. The light-reliant reaction produces ATP and NADPH, which are utilized in the Calvin cycle or dark reaction (Raven 2005).

C3 photosynthesis is a multi-step process occurring in the mesophyll cells of the leaves and involves Rubisco enzyme that fixes carbon dioxide in plants. It catalyzes the carboxylation of RuBP, leading to two particles of 3-PGA. 2-phosphoglycolate cannot be used in the Calvin cycle and is potentially harmful, it must be recycled through a series of complex reactions that occur in the chloroplasts, peroxisomes and mitochondria. Photorespiration is a metabolic pathway in which RUBISCO interacts with intracellular oxygen (Ehleringer and Monson 1993). In the mesophyll cells of C4 plants, the Phosphoenolpyruvate carboxylase fixes carbon dioxide to generate oxaloacetate, which is composed of four carbon molecules. Malate and aspartate, both four-carbon (C4) acids, serve as primary transport molecules in C4 photosynthesis, effectively replacing oxaloacetate by facilitating the movement of fixed carbon from mesophyll cells to bundle-sheath cells. According to (Ehleringer *et al.* 1997), C4 acids malate or aspartate are then transported to the bundle sheath cells where they are undergone through decarboxylation. In C3 plants, Rubisco's oxygenase activity is reduced by C4 photosynthesis. In climates with low carbon dioxide levels, C4 photosynthesis increases its rate. In leaves, C4 photosynthesis takes place in the mesophyll and bundle sheath cells under the high and low levels of intracellular carbon dioxide levels while C3 plants have high rates of photorespiration (Sage *et al.* 2018).

Carbonic Anhydrase (CA) is a key component in the C4 pathway, catalyzing the conversion of CO₂ in plants bicarbonate. It is found in chloroplast and cytosol with major point that chloroplastic form occurs more prevalent in C3 plants. CA is essential for the fixation of the environmental CO₂ in plants, acting as a primary chemical that facilitates the carboxylation step of C4 photosynthesis (Ludwig 2016; Bhat *et al.* 2017). The term "codon usage bias" describes the uneven frequencies at which synonymous codons appear in gene sequence. Codon is the combination of three nucleotides that codes a polypeptide chain amino acid residue or in the

case of stop codons, the termination of the translation. Codon usage bias has emerged in a variety of plants and animals as a result of genetic drift, natural selection and mutation, codon location, context in genes, GC contents, gene expression degree and length as well as composition of the genome. Only 20 distinct amino acids have been deciphered out of 64 distinct codons (61 codons code amino acids while three stop codon). Many amino acids can be coded by multiple codons. The hereditary code is degenerate because of the repetition. The genetic codes of different organisms are often biased to use one of the few codons that code any specific amino acid more frequently than the others. In other words, a more notable recurrence of one will be observed than expected by chance (Parvathy *et al.* 2022).

Methodologies

Sequence data

Eighteen complete coding sequences (CCDs) of Carbonic anhydrase gene of different plants were retrieved from the National Center for Biotechnology Information <https://ncbi.nlm.nih.gov>, in FASTA format and undesired characters except A, T, C, and G were eliminated (Chakraborty *et al.* 2020).

Nucleotide Composition Analyses

Sequences were analyzed to determine nucleotide distributions and composition of single nucleotides (A%, C%, T%, and G%) as well as the nucleotide compositions of each codons at first, second and third position in C3 and C4 plants using the CAI calculator. The total G + C and A + T contents of the gene sequences and the content of synonymous codons at third position were determined (Chakraborty *et al.* 2020).

Relative synonymous codon usage (RSCU) analysis

RSCU analysis was conducted through CAI calculator to determine the frequency with which synonymous codons that code for the same amino acid appear. All fifty nine codons were

classified in three classes: overrepresented, abundant and underrepresented. codons that have RSCU value more than 1.6 are categorized as overrepresented codon while codons having RSCU values from 1.0 to 1.6 are considered as abundant codons and show a positive bias. Additionally, codons having RSCU values less than 0.6 are considered as underrepresented codons (Duan *et al.* 2021).

Codon Bias Index (CBI)

The Codon Bias Index (CBI) measures the preferential use of particular codon within a gene sequence. It measures the preference for some synonymous codons over others during protein translation. A greater CBI value implies a stronger bias towards specific codons while a lower value indicates a more evenly distributed codons.

Codon bias index (CBI) reflect the tendency how frequently favored codons are located in a gene; Codon Bias Index (CBI) provides a directional evaluation of codon usage bias for a specific subset of optimal codons. The CBI formula is as follows:

$$CBI = (N_{opt} - N_{ran}) / (N_{tot} - N_{ran})$$

where N_{opt} is the number of preferred optimal codons, N_{tot} is the total number of codons and N_{ran} is the predicted number of optimal codons (Li *et al.* 2022).

Codon adaptation index (CAI)

CAI is widely used to estimate gene expression levels. CodonW software has been used to determine the CAI value for carbonic Anhydrase gene of both C3 and C4 plants. Its range is from 0 to 1, with values near to 0 indicating selective pressure to preserve specific codon usage bias and the value of 1 indicates the high expression (Grantham *et al.* 1980).

Correspondence analysis (COA)

Correspondence analysis (COA) is a multivariate statistical approach employed to investigate patterns of synonymous codon usage across different genes. In this study, eighteen different plant genes data were plotted in a

multidimensional space to identify the key factors shaping variation. The plots were based on values of relative synonymous codon usage, except for met, trp and stop codons. The COA was performed using CodonW and Originpro v2023 was used for visual graphing (Parvathy *et al.* 2022).

Neutrality plot mapping analysis

Neutrality plot analysis has been used to evaluate the impact of natural selection and mutation pressure on the codon usage pattern of the carbonic anhydrase gene. If the relationship between GC1,2 (GC content at first and second position of the codon) and GC3 (GC content at third position of the codon) concentration is statistically significant and the regression line slope is close to 1 then mutational pressure is recognized as a primary factor shaping codon usage (Wang *et al.* 2021).

ENC plot analysis

ENC plot analysis is used to examine the relationship between GC3 (the GC content at the third codon position) and the Effective Number of Codons (ENC), thereby assessing codon bias in the gene. ENC values range from 20 to 61 and are calculated based on the total number of codons for each amino acid and their frequencies within the gene sequence. Lower ENC values imply greater bias or selectivity in codon usage while higher values indicate more consistent use of synonymous codons.

The analysis with low ENC value indicates strong preference for codon usage. The expected ENC value for various GC3 content was estimated using the following model:

$$ENC = 2 + S + (29/s^2 + (1-s)^2)$$

Where "S" represents the content of third position of the codon (G+C3s) Originpro v2023 has been used to visualize the ENC plot (Wang *et al.* 2023).

Phylogenetic Analysis

Phylogenetic tree is used to study the genetic diversity and evolutionary relationship of the plants based on Carbonic Anhydrase gene.

ClustalW program was used for sequence alignment and MEGA X software was used for phylogenetic tree building (Yang *et al.* 2014).

Results

Nucleotide Content Analysis

Here in this study, the nucleotide contents of eighteen sequences of Carbonic anhydrase genes from different C3 and C4 were analyzed. This analysis determines the nucleotide limitations influencing the codon usage.

Nucleotide content of entire coding sequences of 18 Carbonic anhydrase genes was determined. We looked at the nucleotide content of the Carbonic anhydrase gene to see if there was any influence from nucleotide limitations on codon usage. According to our findings, The average values of nucleotide content A (338.72), C (320.72), T (341.5), G (302.55), while the value of nucleotide

contents in percentage %A (25.88), %C (24.81), %T (25.92), %G (23.37), the percentage of both G+C (48.19), G+A (49.25), G+T (49.30), A+T (51.80), A+C (50.69), C+T (50.74), with G+A frequency and G+T, G+C is lower than that of A+T, A+C, C+T as shown in Figures 1 and Figure 2.

On third position AT, that analyzed DNA sequence exhibits a preference for C nucleotides. This is evident from the higher mean value and frequency of C₃ compared to A₃, T₃, and G₃. Additionally, the combinations G₃+C₃, A₃+C₃, and C₃+T₃ are more frequent than G₃+A₃, A₃+C₃, and G₃+T₃.

The Carbonic anhydrase gene's nucleotide content across plants showed A and T more frequently occurring than G and C showing in Figure 3 and Figure 4.

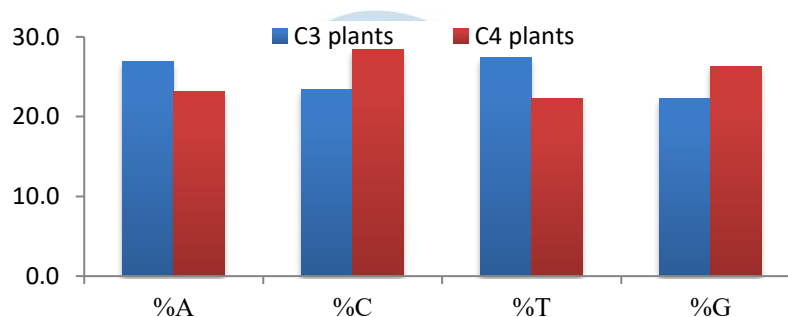


Figure 1: Nucleotide composition analysis: The average value of the A, T, G and C nucleotide composition the entire sequences of C3 and C4 Plants. The nucleotide content of the Carbonic anhydrase gene across different plants revealed that A and T occurred more higher than G & C, indicating a preference for these bases in the gene sequence. In overall C3 and C4 plants Composition, G and C occurs most frequently, suggesting a higher representation of these bases in the codon structures across both plant types.

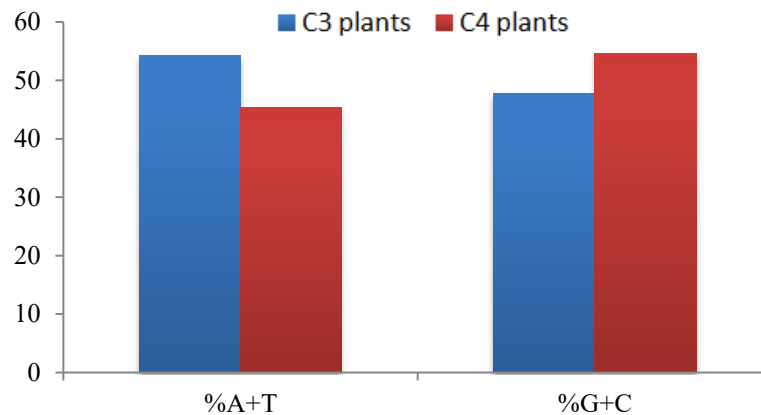


Figure 2: Nucleotide composition analysis: The average value of the A+T C, G+C entire sequences of C3 and C4 Plants. Average composition of Nucleotide (%A+T and %G+C) in C4 and C3 plants. For %A+T, C3 plants show 52%, while C4 plants show about 42%. For %G+C, C3 plants have around 45%, and C4 plants have roughly 52%.

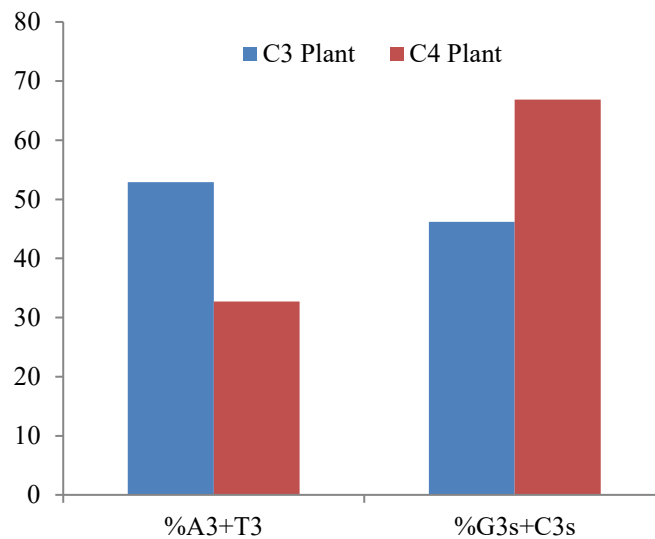


Figure 3: Nucleotide composition analysis: The average value of the AT3 GC3 entire sequences of C3 and C4 Plants. Nucleotide content analysis of 3rd position of codon (%A3+T3 and %G3s+C3s) in C3 and C4 plants. For %A3+T3, C3 plants show 50%, while C4 plants show 30%. For %G3s+C3s, C3 plants show 45%, and C4 plants show 70%, indicating a higher %G3s+C3s in C4 plants.

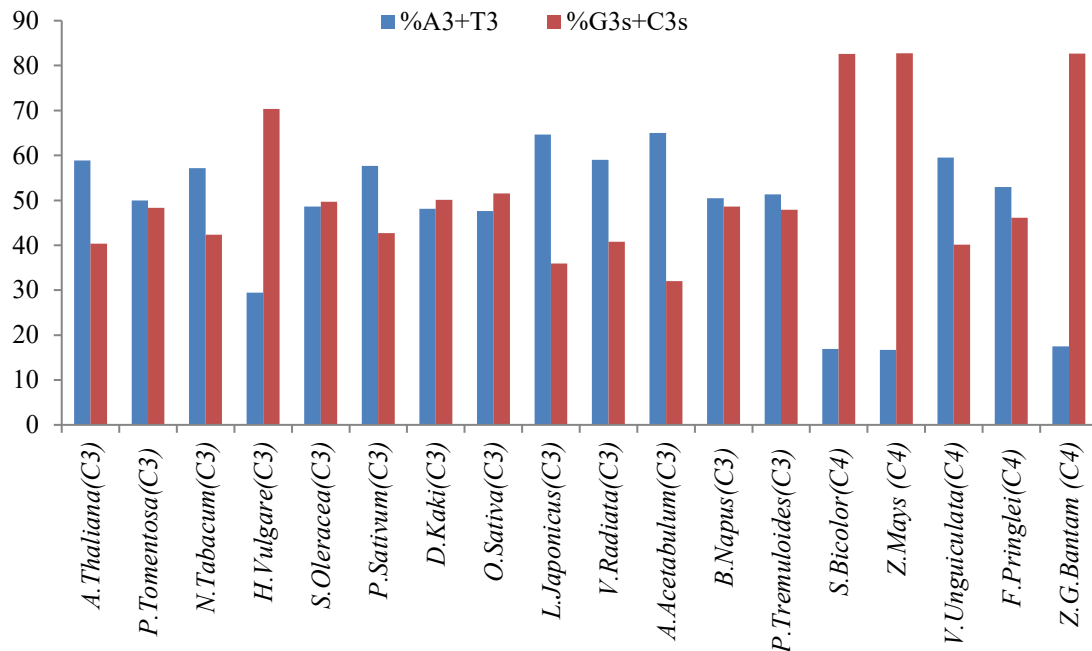


Figure 4: Nucleotide composition analysis: The over All Percentage of AT3 and GC3entire sequences of C3 and C4 Plants. The nucleotide content of (%A3+T3 and %G3s+C3s) across various plant species, categorized as C3 or C4. Most species show %A3+T3 ranging from 40% to 60%, while %G3s+C3s ranges from 20% to 50%. C4 plants, particularly the last three species, exhibit higher %G3s+C3s (up to 70-80%) compared to C3 plants, which generally show lower %G3s+C3s values.

Table 1: The table presents nucleotide composition data for Carbonic anhydrase gene accessions, including percentages for bases (%G, %T, %C, %A) third codon position bases (%G3, %T3, %C3, %A3) Content of GC at position third (GC3s), Codon Bias Index (CBI), ENC, Codon Adaptation Index (CAI value) and Mean Values.

Accession Number	%A	%C	%T	%G	%A3	%C3	%T3	%G3	GC3s	CBI	ENC	CAI value
L18901.1	28.7	21.01	27.81	22.49	31.07	21.6	27.81	19.53	0.404	-0.149	53.3	0.191
MF463312.1	25.82	24.86	25.82	23.51	21.39	26.59	28.61	23.41	0.484	-0.048	56.73	0.229
AF454759.2	26.82	23.47	26.53	23.18	30.17	27.3	27.01	15.52	0.423	0.067	54.1	0.198
L36959.1	22.72	30.19	21.94	25.15	14.15	42.24	15.32	28.29	0.703	0.07	51.16	0.23
M27295.1	27	22.43	27.93	22.64	25.55	16.82	23.05	34.58	0.497	-0.147	53.58	0.119
M63627.1	27.93	21.55	30.51	20.02	29.54	29.06	28.09	13.32	0.427	0.054	52.55	0.203
AB514529.1	30.16	19.88	26.32	23.64	29.82	19.05	18.3	32.83	0.501	-0.122	53.24	0.138
U08404.1	20.16	34.38	21.29	24.17	24.87	20.16	22.77	32.2	0.516	-0.035	60.17	0.125
AJ543754.1	30.02	20.17	27.81	22.01	35.36	19.61	29.28	15.75	0.359	0.015	54	0.201
AF139464.2	27.5	22.96	28.96	20.58	30.18	28.34	28.8	12.67	0.408	-0.005	50.21	0.197
AF170174.1	32.49	13.82	36.44	17.24	32.26	13.71	32.72	21.31	0.32	-0.193	51.17	0.16
KM586049.1	23.88	25.57	27.46	23.08	19.1	28.96	31.34	20.6	0.486	0.054	48.53	0.254
U55837.1	26.98	24.43	26.81	21.78	23.81	27.78	27.51	20.9	0.479	-0.024	56.47	0.224
MF593306.1	20.71	31.48	19.23	28.58	6.68	49.44	10.24	33.63	0.826	0.111	36.95	0.286
HQ697602.1	20.61	31.9	16.22	31.27	9.68	48.12	6.99	35.22	0.827	0.243	39.4	0.256
JQ429799.2	27.99	22.81	28.76	20.44	30.66	27.92	28.83	12.59	0.401	-0.004	49.59	0.198
U19737.1	25.7	23.38	27.53	23.38	22.64	24.38	30.35	22.64	0.461	0.005	56.17	0.243
U08403.1	20.65	32.42	19.32	27.61	7.97	50.55	9.48	32.01	0.827	0.117	38.85	0.258
Mean Values	465.4	446.71	466.6	420.77	424.9	521.6	426.5	427	0.519	0.0005	50.89	0.206111

RSCU Analysis

Synonymous codon frequencies were evaluated using the Relative Synonymous Codon Usage (RSCU) approach. Codons with $RSCU \geq 1.6$ were identified as overrepresented, those with RSCU values of 1 to 1.6 as abundant, and those with $RSCU = 1$ as underrepresented. The study examined 59 codons in Cabonic anhydrase gene of C3, and C4 plants.

First, we examined the findings of 59 codons in C3 plants, where we found only one codons were overrepresented (AGA), and where the RSCU

value is (1.6 or >1.6). 31 abundant codon and 26 underrepresented codon .

The findings of 59 codons in C4 plants are then checked. Five codons overrepresented with RSCU values of 1.6 or >1.6 were concluded while the abundant codons are twenty in number. Under represented codons with value below one were thirty six. This is in contrast to the previous study's findings regarding *Hemiptelea davidii* chloroplast gene and the peach (*Prunus persica*) codon usage bias analysis (Li *et al.* 2022).

Table 2: Comparative RSCU analysis of C3 and C4 plants. The data highlights Relative synonymous codon usage values, where values below than 1.0 indicate underrepresentation. Values between 1.0 and 1.6 denote preferred codons while values exceeding 1.6 indicate highly favored or overrepresented codons.

Comparative RSCU analysis of C3

Amino Acid	Codons	C3 Value	C4 Value
Phe	TTT	1.01	0.58
	TTC	0.98	1.41
Leu	TTA	0.72	0.30
	TTG	1.03	0.65
	CTT	1.46	0.69
	CTC	1.26	2.45
	CTA	0.72	0.48
	CTG	0.78	1.42
Ile	ATT	1.11	0.69
	ATC	0.83	1.83
	ATA	1.05	0.47
Val	GTT	1.23	0.77
	GTC	0.90	1.3
	GTA	0.56	0.4
	GTG	1.29	1.53
Ser	TCT	1.07	0.83
	TCC	1.19	1.84
	TCA	1.30	0.78
	TCG	0.75	0.94
	AGT	0.62	0.34
	AGC	1.04	1.25
Pro	CCT	1.11	1.11
	CCC	1.07	1.2
	CCA	1.30	0.87
	CCG	0.50	0.81
Thr	ACT	1.24	0.44

	ACC	1.20	2.41
	ACA	1.15	0.80
	ACG	0.39	0.33

Comparative RSCU analysis of C4

Amino Acid	Codons	C3 Value	C4 Value
Ala	GCT	1.17	0.78
	GCC	1.05	1.77
	GCA	1.39	0.71
	GCG	0.38	0.71
Tyr	TAT	1.15	0.81
	TAC	0.84	1.18
His	CAT	1.07	0.70
	CAC	0.92	1.29
Gln	CAA	1.10	0.79
	CAG	0.89	1.20
Asn	AAT	1.04	0.40
	AAC	0.95	1.59
Lys	AAA	0.99	0.70
	AAG	1.00	1.29
Asp	GAT	0.99	0.46
	GAC	1.00	1.54
Glu	GAA	1.07	0.62
	GAG	0.92	1.37
Cys	TGT	0.99	0.52
	TGC	1.00	1.47
Arg	CGT	0.81	1.00
	CGC	0.44	1.38
	CGA	0.70	0.24
	CGG	0.59	0.50
	AGA	2.14	1.23
	AGG	1.28	1.63
Gly	GGT	1.00	0.86
	GGC	0.88	1.83
	GGA	1.21	0.63
	GGG	0.90	0.67

Effective number of codon (ENC)

ENC plot analysis to understand the correlation between GC3 and ENC, revealed codon usage bias. Values of ENC for Carbonic Anhydrase genes in C3 plants, ranged between 48.53 to 60.17, with a mean value of 53.48. Similarly, the ENC values for Carbonic Anhydrase genes in C4 plants ranged between 36.95 and 56.17, with a

mean value of 44.19. A low ENC number indicates a highly biasness while a higher ENC value reduces it. The analysis found that all ENC values are higher than 35 which indicate a weak bias in codon usage of Carbonic anhydrase gene.

ENC plot analysis

ENC plot analysis compares ENC values to GC3s to determine a gene's bias codon. Mutational pressure affects the plot, with points on the fitted curve representing random codon use. Natural selection influences the plot, with most points

below the anticipated curve. The findings indicate that natural selection and mutational pressure influence the patterns codon usage bias of the Carbonic anhydrase gene as shown in Figure5.

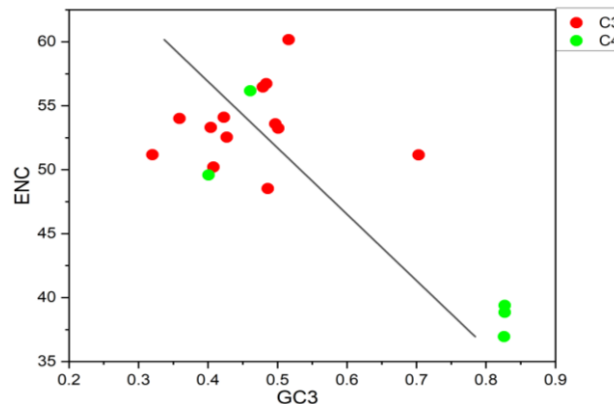


Figure 5: ENC Plot Analysis. This graph shows the relationship between GC3 and ENC for C3 and C4 plant species. The red dots indicate C3 plants, whereas the green dots signify C4 plant. The downward trend line indicates that as ENC increases, GC3 tends to decrease, suggesting a potential interaction between codon usage bias and GC content, with C4 plants generally showing higher GC3 values compared to C3 plants at similar ENC levels. This pattern could be influenced by either mutational pressure or natural selection.

Neutrality plot

The neutrality plot is a valuable tool for examining the relationship between GC content at the first and second codon positions (GC1,2) and GC content at the third codon position (GC3), providing insights into how mutation and natural selection influenced codon usage bias. For Carbonic Anhydrase gene, we constructed a

neutrality plot to explore this relationship. Our analysis revealed a positive A non-significant correlation was found between GC1,2 and GC3, with a correlation coefficient of $r = 0.46$ and $p = 0.05$. These results indicate that codon usage patterns are likely shaped by both mutational pressure and natural selection. Though their effects are not significant as shown in Figure6.

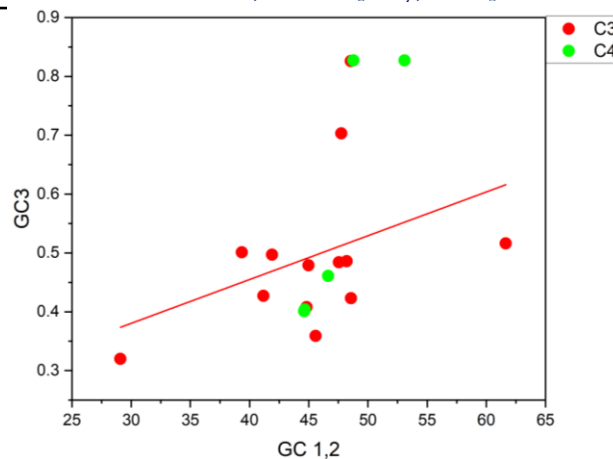


Figure6: Neutrality Plot. This Neutrality Plot analysis graph shows the relationship between GC3 (GC content at the third codon position) and GC12 (GC content at the first and second codon positions) for C3 and C4 plant species. Red dots denote C3 plants, while green dots indicate C4 plants. The upward trend line suggests a positive relationship amongst GC3 and GC12, indicating that the GC content on third codon position may be influenced by mutational bias rather than strict neutrality, with C4 plants showing slightly higher GC3 values compared to C3 plants at similar GC12 levels.

Correspondence analysis (COA)

Correspondence Analysis (COA) was used to examine the codon usage patterns in the Carbonic Anhydrase genes of C3 and C4 plants, focusing on the RSCU values of 59 sense codons. The analysis revealed that most C3 plant species are positioned farther from C4 plant species, indicating a significant dissimilarity in their codon usage patterns. This disparity may be

attributed to differences in genetic and environmental factors. To further investigate these variations, Principal Component Analysis (PCA) was performed. PCA showed that axis 1 accounts for 75.25% of the variance, while axis 2 accounts for 24.75%, highlighting the primary sources of variation between the two axes as shown in Figure7.

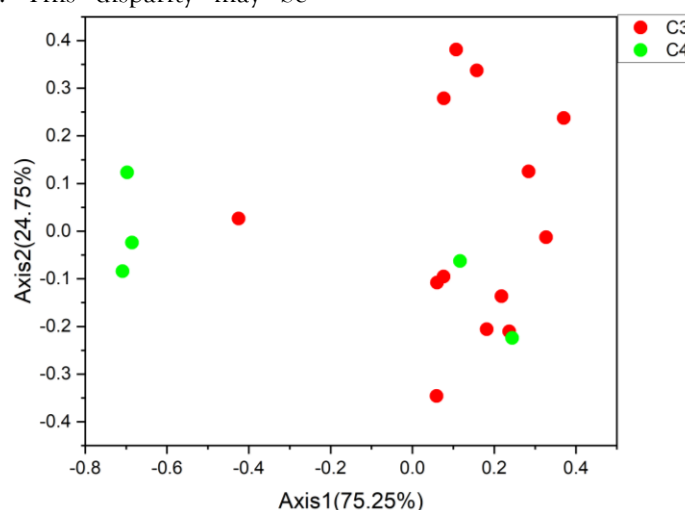


Figure 7: Corresponding analysis plot. This corresponding analysis plot graph shows the relationship between Axis1 (75.25%) and Axis2 (24.75%) for C3 and C4 plant species. Red dots denote C3 plants, while green dots indicate C4 plants. The distribution suggests differences in the genetic composition or

expression patterns between C3 and C4 plants, with C4 plants tending to cluster more towards positive Axis2 values and C3 plants showing a broader spread across both axes, indicating potential variations in their molecular profiles.

Phylogenetic analysis

In phylogenetic analysis of the Carbonic Anhydrase gene, the results indicate that most plant species fall within the same clade, suggesting a recent common ancestor. Within

this clade, there are distinct subgroups comprising sequences from both C3 and C4 plants. The close clustering of several sequences points to a more recent shared evolutionary history among these plants as shown in Figure 8.

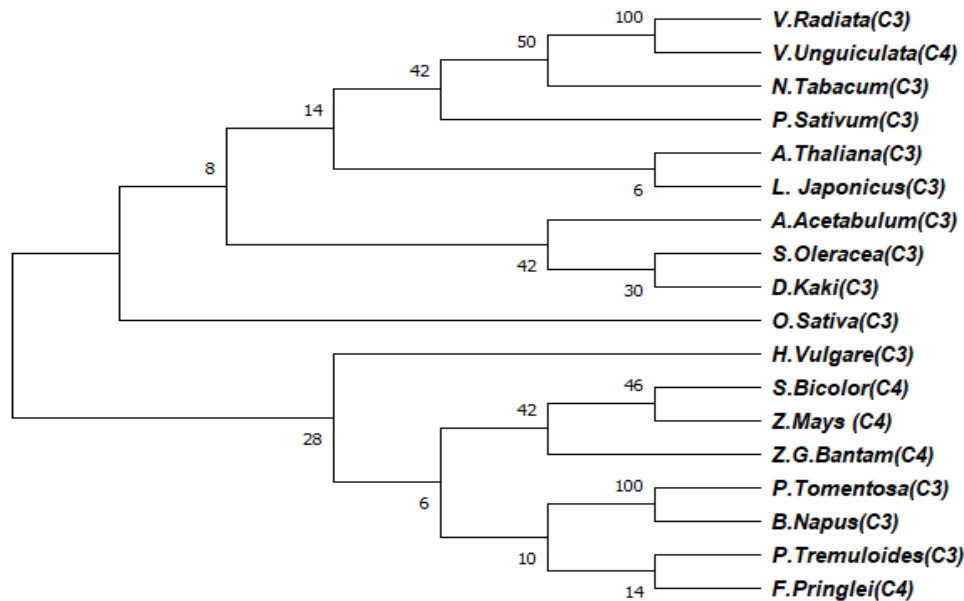


Figure8: PhylogeneticAnalysis The tree illustrates the evolutionary connection between various C3 and C4 plant species. The tree branches show divergence points with bootstrap values indicating the reliability of the nodes, ranging from 6 to 100. C3 plants *N. tabacum*, *A. thaliana* and C4 plants like *V. unguiculata*, *Z. mays* are distributed across the tree, suggesting multiple independent origins of the C4 photosynthetic pathway, with some closely related species showing both C3 and C4 characteristics.

Discussion

In order to explore CUB, a variety of techniques have been used, include the Codon Adaptation Index (CAI), directional codon bias measure, Codon Bias Index (CBI), Intrinsic Codon Deviation Index (ICDI), and Effective Number of Codons (ENC). Codon usage bias (CUB) is important for understanding gene expression levels and the molecular organization of the genome. It also helps to understand the occurrence and impact of lateral gene transfer (Carbone *et al.* 2003), heterologous expression and functional conservation of gene expression

across organisms as well as the molecular evolution of individual genes (Sharp *et al.* 2005).

The present study focuses on the CUB, gene expression and nucleotide content of Carbonic Anhydrase gene across the plant species. The relevance of the carbonic anhydrase gene analysis in this work arises from its critical involvement in the photosynthetic process. Since the patterns of codon usage do not change during translation, it is crucial to comprehend patterns in order to investigate the manner of translational selection of the gene in different plant species. These findings give insight into the patterns of codon use and any possible consequences for the gene

expression. It is interesting that the entire genome's nucleotide makeup can have a substantial influence on CUB (Rahman *et al.* 2022). We investigated the influencing factors, analyzed the patterns of codon usage bias in carbonic anhydrase gene across various plant species and determined the evolutionary relationships among plants based on the gene as the other researchers did (Parvathy *et al.* 2022; Rahman *et al.* 2022).

This study examines the nucleotide composition, codon use bias, and gene expression of the Carbonic anhydrase gene in various plant species. We tried to understand the translational selection of the gene and its effects on gene expression. The findings revealed a general bias towards A and T nucleotides, with a frequency and trend of T (25.92%) > A (25.88%) > C (24.81%) > G (23.68%) in both C3 and C4 plants (Duan *et al.* 2021).

The study shows slightly different nucleotide mean values for synonymous codons at the third codon position similar to the Bulmer theory of Natural selection trend as presented in other studies (Bulmer 1991). We found that A and T nucleotides compose the majority of the carbonic anhydrase gene in various plant species like the codon use patterns of seven Rosales species and chloroplast genomes of ten epimedium species (Zhang *et al.* 2022). Nucleotide content at codon locations revealed significant patterns, suggesting a potential natural selection for certain codons at the third position, as reported by previous studies (Duan *et al.* 2021; Gao *et al.* 2022).

RSCU analysis reveals the preference for certain codons over others in carbonic anhydrase gene's usage patterns in C3 and C4 plants. Results suggest that carbonic anhydrase gene exhibits codon bias in several plant species. Here, thirty one high frequency codons (RSCU > 1 to 1.6) were discovered in RSCU study of C3 plants. Among these 28 high frequency codons, 15 codons (TTT, TTG, CTT, ATT, ATA, GTT, TCT, ACT, ACA, CAT, TAT, CAA, AAT, AAG, GAA) are ending with (A and T) and the rest of the codons (CTC, GTG, TCC, AGC, CCT, CCT, CCC, CCA, ACC, GCA, GCT, GAC,

TGC, GGT, GGA) are ending with G and C. These findings agree with previous studies as Peach (*Prunus persica*) codon usage bias analysis (Li *et al.* 2022). While the other 31 codons ending mostly with (G and C) have RSCU values below 1. These 26 codons include 15 i.e. CTG, GTC, TCG, CCG, ACG, GCG, CAC, CAG, GAG, CGT, CGC, CGA, CGG, GGC, GGG and the rest of the codons i.e. TTC, TTA, CTA, ATC, GTA, AGT, TAC, AAC, AAA, GAT, TGT are terminating with A and T, which is in line with the results of the earlier investigation of *Hemiptelea davidii* chloroplast gene (Liu *et al.* 2020). There was only one over-represented codon i.e. AGA in all C3 plant. While there are five overrepresented codons with RSCU value > 1.6 in RSCU study of C4 plants and those codons are ATC, TCC, GCC, GGC, ACC. Among the 19 high frequency codons (RSCU > 1), fourteen codons i.e. CTG, GTC, GTG, AGC, CCT, CCC, CAC, CAG, GAG, TGC, CGC, AGG, GGC are occurring frequently and end with G and C. While remaining codons (TTC, TAC, AAC, AGA) end with A and T. There are 35 under-represented codons in which twenty two (TTT, TTA, TTG, CTT, CTA, ATT, ATA, GTT, GTA, TCT, TCA, AGT, ACT, ACA, TAT, CAT, CAA, AAT, AAA, GAT, GAA, TGT) are ending with A and T. A very few codons are ending with G and C that include TCG, CCA, CCG, ACC, ACG, GCT, GCA, GCG, CGA, CGG, GGT, GGA, GGG. However, in C4 plant, the high frequency codons are terminated with G and C while the underrepresented codons are terminated with A and T. This is in contrast to the previous study's findings in *Hemiptelea davidii* chloroplast gene and peach (*Prunus persica*) codon usage bias analysis (Li *et al.* 2022).

The codon bias index (CBI) approach was used to assess the overall codon usage bias of the carbon anhydrase gene. Carbonic anhydrase gene does not show any codon usage bias according to the CBI values which vary from -0.149 to 0.117. These findings suggest that the usage patterns for codons in carbonic anhydrase gene are neither significantly overrepresented nor underrepresented for any specific codon. The low

CBI values demonstrate that codon use patterns in carbonic anhydrase gene are mostly unaffected by selection for particular codons (Chakraborty *et al.* 2020).

The ENC plot analysis revealed that mutational pressure and natural selection significantly influenced codon use patterns in carbonic anhydrase gene. The distribution of points around the predicted curve suggests a combination of random codon use and selective pressure, with natural selection being the primary factor affecting codon usage patterns (Li *et al.* 2022; Wang *et al.* 2023).

The neutral assessment analysis was performed to identify the factors influencing codon usage bias in carbonic anhydrase gene. The analysis showed a positive correlation between GC content at various codon locations, indicating that natural selection may affect codon use bias. However, mutational pressure did not significantly impact codon usage. The neutrality plot analysis demonstrated the influence on codon usage patterns of GC, indicating the potential influence of natural selection and mutational pressure.

Correspondence analysis (COA) reveals changes in codon usage among plant species. C3 and C4 plant species have similar patterns, possibly due to distinct carbon fixation mechanisms. However, some C4 plant species are dispersed, suggesting more heterogeneity in codon usage patterns, possibly driven by genetic or environmental factors. This aligns with a prior study revealing a codon usage bias in *Oryza* species. Carbonic anhydrase gene's codon usage patterns are slightly varied across plants, highlighting the complex relationship between genetic and environmental factors that influence the bias in codon usage among plant species (Chakraborty *et al.* 2020).

To investigate for differences, we used a dimension reduction approach i.e. principal component analysis (PCA) to both axes (axis 1 and axis 2). Because axis 1 accounts 75.25% of the variance and axis 2 is 24.75%.

The phylogenetic analysis of carbonic anhydrase gene sequences provided insights into the evolutionary relationships between plant species. The tree reveals genetic similarity and divergences

among species. Distantly related species may show major genetic divergence over time. These results align with previous studies about the complete chloroplast genomes of two *Lespedeza* Species and phylogenetic relationship of the species from *Biebersteiniaceae* and *Nitrariaceae* families (Somaratne *et al.* 2019; Chi *et al.* 2020).

Conclusion

The findings revealed a general bias towards A and T nucleotides, with distinct patterns observed at different codon positions. The RSCU analysis identified preferential codons, with most high-frequency codons ending by G and C in C4, A and T in C3 plants. Interestingly, the codon bias index (CBI) suggested no significant overall bias, indicating a balanced usage pattern. Further analysis using ENC plots and neutrality assessment revealed a complex interplay between natural selection and mutational pressure. The ENC plot distribution suggested a combination of random codon use and selective pressure, while natural selection play more prominent role. Additionally, the neutrality plot analysis indicated a positive correlation between GC content and codon usage bias, supporting the influence of natural selection. Finally, correspondence analysis (COA) showed variations in codon usage among plant species, potentially due to distinct carbon fixation mechanisms and genetic or environmental factors. Phylogenetic analysis confirmed evolutionary relationships among plant species based on genetic similarities and divergences.

References

- Bhat, F. A., B. A. Ganai and B. Uqab (2017). Carbonic anhydrase: mechanism, structure and importance in higher plants. *Asian J Plant Sci Res* 7(3): 17-23.
- Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129(3): 897-907.
- Carbone, A., A. Zinovyev and F. Képes (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19(16): 2005-2015.

- Chakraborty, S., S. Yengkhom and A. Uddin (2020). Analysis of codon usage bias of chloroplast genes in *Oryza* species: Codon usage of chloroplast genes in *Oryza* species. *Planta* 252: 1-20.
- Chi, X., F. Zhang, Q. Dong and S. Chen (2020). Insights into comparative genomics, codon usage bias, and phylogenetic relationship of species from *Biebersteiniaceae* and *Nitrariaceae* based on complete chloroplast genomes. *Plants* 9(11): 1605.
- Duan, H., Q. Zhang, C. Wang, F. Li, F. Tian, Y. Lu, Y. Hu, H. Yang and G. Cui (2021). Analysis of codon usage patterns of the chloroplast genome in *Delphinium grandiflorum* L. reveals a preference for AT-ending codons as a result of major selection constraints. *PeerJ* 9: e10787.
- Ehleringer, J. R., T. E. Cerling and B. R. Helliker (1997). C4 photosynthesis, atmospheric CO₂, and climate. *Oecologia* 112(3): 285-299.
- Ehleringer, J. R. and R. K. Monson (1993). Evolutionary and ecological aspects of photosynthetic pathway variation. *Annual review of ecology and Systematics* 24(1): 411-439.
- Gao, Y., Y. Lu, Y. Song and L. Jing (2022). Analysis of codon usage bias of WRKY transcription factors in *Helianthus annuus*. *BMC Genomic Data* 23(1): 46.
- Gest, H. (2002). History of the word photosynthesis and evolution of its definition. *Photosynth Res* 73(1-3): 7-10.
- Grantham, R., C. Gautier and M. Gouy (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic acids research* 8(9): 1893-1912.
- Li, R., X. Zhang, X. Ma, R. Guo, S. Yan, G. Jin and P. Zhou (2022). Analyses of Codon Usage Patterns and Codon Usage Bias in Peach (*Prunus persica*). *Molecular Plant Breeding* 13.
- Liu, H., Y. Lu, B. Lan and J. Xu (2020). Codon usage by chloroplast gene is bias in *Hemiptelea davidii*. *Journal of Genetics* 99: 1-11.
- Ludwig, M. (2016). Evolution of carbonic anhydrase in C4 plants. *Current opinion in plant biology* 31: 16-22.
- Parvathy, S. T., V. Udayasuriyan and V. Bhadana (2022). Codon usage bias. *Molecular biology reports* 49(1): 539-565.
- Rahman, S. U., M. Abdullah, A. W. Khan, M. I. U. Haq, N. ul Haq, A. Aziz and S. Tao (2022). A detailed comparative analysis of codon usage bias in Alongshan virus. *Virus research* 308: 198646.
- Raven PH, E. R., Eichhorn SE (2005). *Biology of Plants*, (7th Ed). Freeman and Company: 124-127.
- Sage, R. F., R. K. Monson, J. R. Ehleringer, S. Adachi and R. W. Pearcy (2018). Some like it hot: the physiological ecology of C4 plant evolution. *Oecologia* 187: 941-966.
- Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research* 33(4): 1141-1153.
- Somarathne, Y., D.-L. Guan, W.-Q. Wang, L. Zhao and S.-Q. Xu (2019). The complete chloroplast genomes of two *Lespedeza* species: Insights into codon usage bias, RNA editing sites, and phylogenetic relationships in *Desmodieae* (Fabaceae: Papilionoideae). *Plants* 9(1): 51.
- Wang, X., Y. Wang, S. Li, Y. Liu and B. Zhu (2021). Analysis of Codon Usage Bias in the *Platycarya* Chloroplast Genome. *Tree Genetics and Molecular Breeding* 11.
- Wang, Y., D. Jiang, K. Guo, L. Zhao, F. Meng, J. Xiao, Y. Niu and Y. Sun (2023). Comparative analysis of codon usage patterns in chloroplast genomes of ten *Epimedium* species. *BMC Genomic Data* 24(1): 3.

Yang, X., X. Luo and X. Cai (2014). Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasites & vectors* 7(1): 1-11.

Zhang, Y., Z. Shen, X. Meng, L. Zhang, Z. Liu, M. Liu, F. Zhang and J. Zhao (2022). Codon usage patterns across seven *Rosales* species. *BMC Plant Biology* 22(1): 65.

