Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

Maejo International Journal of Science and Technology

e-ISSN 2697-4746
Available online at www.mijst.mju.ac.th

Full Paper

Genomic features and phylogenetic analysis of complete chloroplast genome of *Caragana bicolor*

Yongming Fan [1, *], Mengshan Wang [1], Lingjun Li [2, *], Le Chen [3], Can Zhao [1], Jian Hu [1] and Heng Zeng [4]

[1] School of Human Settlements, North China University of Water Resources and Electric Power, Zhengzhou, China
[2] Bureau of Comprehensive Development Ministry of Water Resources, Beijing, China
[3] College of Landscape Architecture, Beijing Forestry University, Beijing, China
[4] Hydrology and Water Resources Monitoring Center for the Upper Reaches of Ganjiang
* Corresponding authors, e-mail: fanyongming@ncwu.edu.cn; vlingjun@126.com

Received: 26 February 2026 / Accepted: 9 July 2026 / Published: 27 July 2026

Abstract: *Caragana* is an important genus in the Fabaceae family with significant ecological, ornamental and economic values. It includes over 100 species, many of which exhibit highly similar morphological characteristics, making species identification using traditional taxonomic methods quite challenging. Therefore, employing scientific, accurate and efficient methods for the identification of *Caragana* plants is of great significance. In this study the complete chloroplast genome of *Caragana bicolor* was sequenced using the Illumina HiSeq 4000 platform, and a systematic analysis was conducted by combining the data with the published chloroplast genomes of 56 closely related species. The results indicated that the complete chloroplast genome of *C. bicolo*r is 131,396 bp in length, with an overall guanine and cytosine content of 35.14%. The chloroplast genome of *C. bicolor* does not exhibit the typical quadripartite structure, suggesting possible degeneration of its inverted repeat regions. A total of 110 genes were identified in the *C. bicolor* chloroplast genome. Additionally, 33 simple sequence repeat loci were detected in the *C. bicolor* chloroplast genome, with lengths ranging from 10 to 113 bp. Phylogenetic analysis revealed that *C. bicolor* is most closely related to *C. tibetica*, and this sister clade, together with species such as *C. jubata* and *C. sinica*, forms the core branch of the genus *Caragana*, confirming the taxonomic placement of *C. bicolor* within the genus. This study elucidates the key characteristics of the *C. bicolor* chloroplast genome, providing a reference for phylogenetic analyses of *Caragana* using chloroplast genomes and laying the foundation for further investigation into the origin, conservation and utilisation of *C. bicolor*.

Keywords: *Caragana bicolor*, chloroplast genome, comparative genomics, simple sequence repeat, phylogenetic analysis

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

INTRODUCTION

Caragana is a genus of shrubs in the Fabaceae, comprising approximately over 100 species worldwide. These species are primarily distributed across arid and semi-arid regions of Asia and Europe. In China 66 species have been identified, with their presence recorded in north-eastern, northern, north-western and south-western provinces [1]. *Caragana* plants exhibit remarkable cold and drought tolerance, along with a nitrogen fixation capability. They possess the typical Fabaceae symbiotic nitrogen-fixation characteristics, forming root nodules that convert atmospheric nitrogen into plant-available nitrogen nutrients. This process effectively enhances soil fertility and promotes nutrient cycling in degraded ecosystems [2].

Traditional taxonomy primarily relies on differences in floral structures and morphological characteristics of vegetative organs for species delimitation [3, 4]. However, most species within the genus *Caragana* exhibit a high degree of morphological convergence, making it very difficult to achieve stable and reliable species identification based solely on external morphology [5]. Although the chloroplast genomes of several *Caragana* species have been sequenced, the phylogenetic relationships at the whole-genome level between newly sequenced species and previously reported ones have not been adequately integrated and analysed, leaving the evolutionary framework within the genus unclear [6-11]. With the rapid development of molecular biology techniques, particularly the widespread application of high-throughput sequencing and comparative genomics methods, new pathways have emerged to advance research in plant phylogenetics [12]. The chloroplast genome has a relatively simple, stable and highly conserved structure, making it widely used in studies of plant population evolution, population genetics and phylogenetic relationships, serving as an ideal tool for investigating evolutionary relationships among plant species [13, 14].

Caragana bicolor is a species within the genus *Caragana*. It exhibits close morphological affinity to many other species within the same genus and is difficult to distinguish [15]. In this study *C. bicolor* was selected as the research subject, and its complete chloroplast genome was sequenced and assembled using the Illumina high-throughput sequencing platform. Based on the obtained genome sequence, we systematically analysed the fundamental structural characteristics of the chloroplast genome of this species including genome size, guanine and cytosine (GC) content, gene composition and gene order. Furthermore, through comparative genomic analysis of 56 chloroplast genomes of its closely related species and representative Fabaceae species, we identified highly variable regions. These regions can serve as potential DNA barcodes for the rapid and accurate identification of *C. bicolor* and its closely related species. This study enriches the chloroplast genome database of *Caragana* and provides a reference for phylogenetic research on *Caragana* species, as well as for the conservation and utilisation of their germplasm resources.

MATERIALS AND METHODS

Plant Material and DNA Extraction and Sequencing

Leaves of *C. bicolor* were collected in Deqin County, Yunnan Province, China. The species identification was verified by Associate Researcher Juntong Chen from the Kunming Institute of Botany, Chinese Academy of Sciences. The leaf material was preserved at -80°C in the Aquatic Plant Application and Germplasm Enhancement Laboratory of Yellow River Basin, North China University of Water Resources and Electric Power, for subsequent DNA extraction. A voucher specimen was deposited in the same laboratory (Specimen No.: NCWU-20200617). Total genomic DNA was extracted from the fresh leaves using magnetic bead-based method with the CretMag™

Multi Sample DNA Kit (Suzhou Cretaceous Biotechnology Co., China), following the manufacturer's protocol [16,17]. The quality of the extracted DNA was assessed. The concentration was measured as 130.4 ng/μL using a Qubit fluorometer, and the A260/280 ratio was determined to be 1.81 using a NanoDrop spectrophotometer. Furthermore, agarose gel electrophoresis confirmed the integrity of the DNA, showing a clear band without significant degradation, thus meeting the requirements for downstream experiments. A DNA library was constructed using the VAHTS Universal Plus DNA Library Prep Kit (Vazyme Biotech Co., China). Sequencing was performed on an Illumina HiSeq 4000 platform (Shanghai OE Biotech Co., China) with a read length of 150 bp. Quality control of the raw sequencing data was conducted using the FastQC software package (version 0.12.0, available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accessed on December 1, 2024) [18,19].

Complete Genomes Assembly and Annotation

Prior to the assembly of the *C. bicolor* chloroplast genome, raw sequencing reads were subjected to quality control using fastp v0.23.4, a widely used bioinformatics software tool, specifically an ultra-fast all-in-one preprocessing tool for FASTQ sequencing data (available at: https://github.com/OpenGene/fastp, accessed on 18 October 2024) to generate clean reads [18]. Subsequently, the chloroplast genome of *C. bicolor* was assembled using GetOrganelle v1.7.7.1, a dedicated bioinformatics toolkit for precise assembly of organelle genomes from whole genome sequencing data (available at: https://github.com/Kinggerm/GetOrganelle, accessed on March 18, 2025) with its default settings [20], using the chloroplast genome of *C. tibetica* (NCBI accession number: OQ942026.1) as the reference. The assembled chloroplast genome sequence was annotated using CPGAVAS2, an integrated online server for chloroplast genome sequence annotation and analysis (available at: http://www.herbalgenomics.org/cpgavas/) [21]. The circular genome map was visualised using OGDRAW v1.2, an online visualisation tool for generating high-quality circular or linear maps of organelle genomes (available online: http://ogdraw.mpimp-golm.mpg.de/, accessed on 16 April 2025) [22].

Identification of Simple Sequence Repeats (SSRs)

SSRs identification were performed using MISA, a bioinformatics tool for identifying and localising SSRs (available at: https://webblast.ipk-gatersleben.de/misa/, accessed on 3 August 2025) [23]. The minimum number of repeat units for SSR motif lengths of 1, 2, 3, 4, 5 and 6 were set to 10, 6, 5, 5, 5 and 5 respectively [17].

Comparative Genomic Analysis

This study focused on *C. bicolor* and its ten closely related species, i.e. *C. acanthophylla* (accession no.: PP897765.1), *C. erinacea* (PP897766.1), *C. frutex* (PP316084.1), *C. halodendron* (MW760845.1), *C. jubata* (MT211963.1), *C. korshinskii* (NC_035229.1), *C. kozlowii* (NC_035228.1), *Calophaca sinica* (MN696543.1), *Calophaca soongorica* (PP897764.1) and *Astragalus tumbatsica* (NC_083912.1). Using the chloroplast genome of *C. acanthophylla* (accession no.: PP897765.1) as a reference, a comparative analysis of the chloroplast genomes was conducted using the mVISTA tool in Shuffle-LAGAN mode (available at: https://genome.lbl.gov/vista/mvista/submit.shtml, accessed on August 3, 2025) [24]. Nucleotide diversity (Pi) was calculated using CPStools v2.5, a Python package designed for versatile analysis

of chloroplast genomes [25]. The chloroplast genome sequences of these species, excluding that of *C. acanthophylla* which served as the reference, were downloaded from the NCBI database for the mVISTA analysis.

Codon Usage Analysis

Codon usage bias analysis was evaluated using CodonW v1.4.2, a codon usage bias analysis software package (available online: https://sourceforge.net/projects/codonw/files/codonw/Win32-Executables-1.4.2/, accessed: 3 August 2025) by calculating the relative synonymous codon usage (RSCU) values [26]. Based on conventional interpretation, an RSCU value equal to 1.0 indicates no usage preference for a given codon, whereas a value greater than 1.00 reflects a preferential or higher-frequency usage of that codon.

Phylogenetic Analysis

Phylogenetic analysis was performed using the complete chloroplast genome sequences of 56 species downloaded from the NCBI database, along with the newly sequenced chloroplast genome of *C. bicolor* obtained in this study. *Glycyrrhiza uralensis* was selected as the outgroup. PhyloSuite v1.2.3, a one-stop, graphical desktop platform for molecular phylogenetic analysis, was employed to extract coding sequences and conduct integrated bioinformatic analysis [27]. Genome sequence alignment was carried out using MAFFT v7.505, a widely-used multiple sequence alignment (MSA) software, which automatically selects the optimal algorithm for different sequence types based on the auto strategy [28]. The aligned sequences were refined with trimAl v1.2, a tool for automatic trimming of MSA, to remove poorly aligned regions and gaps [29]. Batch optimisation of protein-coding genes was conducted using MACSE v2.0.3, an MSA tool specifically for protein-coding DNA sequences (CDS) [30]. Finally, a phylogenetic tree was constructed based on the maximum likelihood method with IQ-TREE, a phylogenetic inference software package (available at: http://www.iqtree.org/#download, accessed on August 2, 2025), and the robustness of the tree topology was assessed with 5,000 bootstrap replicates [31].

RESULTS

Chloroplast Genome Features of *C. bicolor*

Sequencing generated 54,306,446 raw reads (8.15 Gb raw bases). After quality filtering, 54,235,560 clean reads were obtained, totalling 8.08 Gb of clean bases, with 96.38% of bases having a quality score above Q30 (indicating a base-calling error probability of less than 0.1%). The chloroplast genome of *C. bicolor* is 131,396 bp in length (Figure 1). The overall GC content is 35.14%. The chloroplast genome of *C. bicolor* does not conform to the typical quadripartite structure, suggesting possible degeneration of its inverted repeat (IR) regions, which are no longer homologous and symmetrical sequences.

A total of 110 genes were identified in the chloroplast genome of *C. bicolor*, namely 75 protein-coding genes, 31 tRNA genes and 4 rRNA genes. Based on their functions, these genes can be categorised into four groups: photosynthesis, self-replication, other genes and unknown (Table 1). Among them, photosynthesis-related genes are the most abundant, with a total of 44 genes including subunits of ATP synthase, photosystem II, NADH-dehydrogenase, cytochrome b/f complex, photosystem I and the subunit of rubisco. The self-replication category includes the large subunit of ribosome, DNA-dependent RNA polymerase and the small subunit of ribosome. Other

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

genes comprise subunits of Acetyl-CoA carboxylase, c-type cytochrome synthesis genes, envelope membrane proteins, proteases and maturases.

Seventeen genes containing introns were identified in the chloroplast genome of *C. bicolor*. Among them, *ycf3* and *clpP* contain two introns each, representing the most complex structures. The remaining 15 genes (*rpl16*, *petD*, *petB*, *trnV-UAC*, *trnL-UAA*, *trnG-UCC*, *atpF*, *rpoC1*, *rps16*, *trnK-UUU*, *rpl2*, *ndhB*, *trnI-GAU*, *trnA-UGC*, *ndhA*) contain one intron each. Among these 15 single-intron genes, four (*trnI-GAU*, *trnA-UGC*, *ndhB*, *rpl2*) are located in the IR regions and are present in one copy on each of the positive and negative strands. The original chloroplast genome sequence data of *C. bicolor* have been deposited in the China National GeneBank Sequence Archive [32] under accession number CNX1292652.

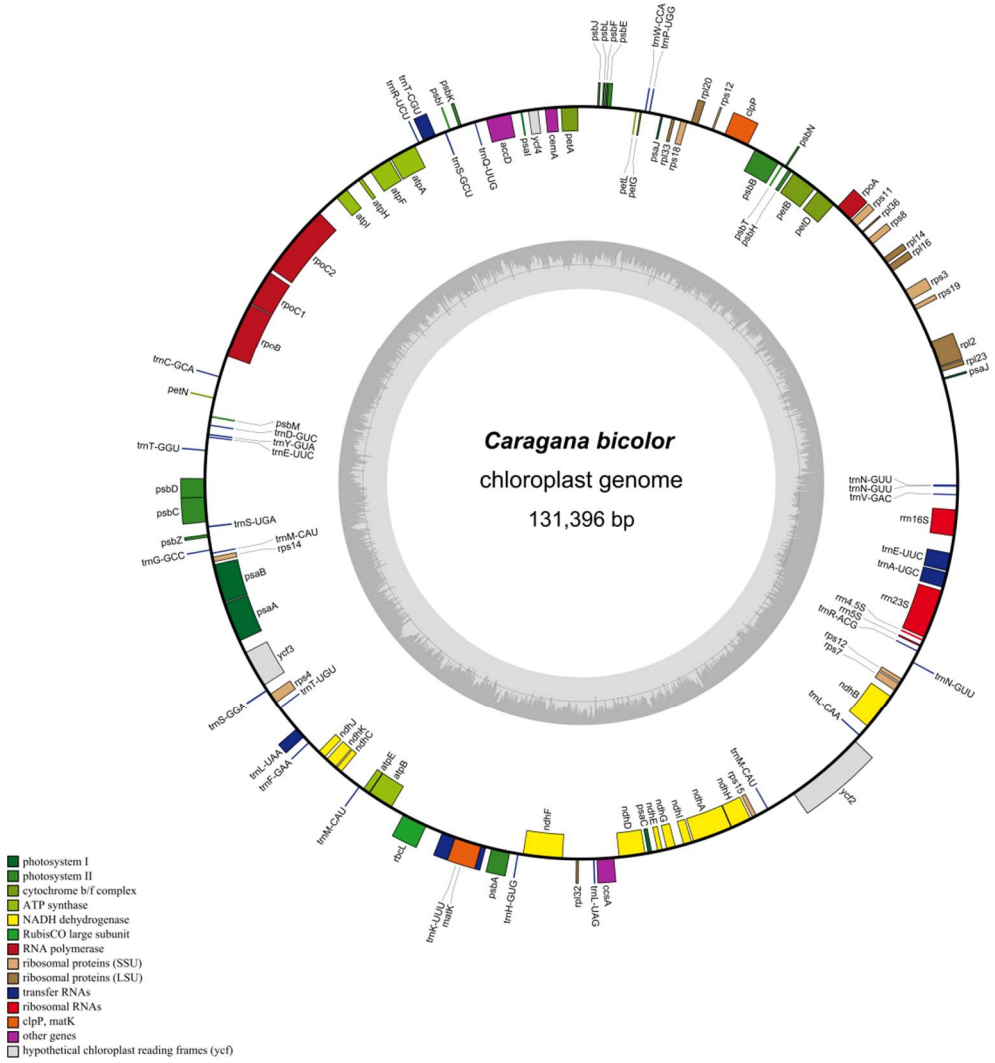

Figure 1. Circular complete chloroplast genome map of *C. bicolor*. Genes outside the circle are transcribed clockwise, while those inside are transcribed counterclockwise.

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

Table 1. Gene composition of *C. bicolor* chloroplast genome

Category of genes	Group of genes	Name of genes
Photosynthesis	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Subunits of NADH-dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	Subunits of photosystem I	*psaA, psaB, psaC, psaI, psaJ**
	Subunit of rubisco	*rbcL*
Self replication	Large subunit of ribosome	*rpl14, rpl16, rpl2, rpl20, rpl23, rpl32, rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps12, rps14, rps15, rps18, rps19, rps3, rps4, rps7, rps8*
Other genes	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA*
	Envelop membrane protein	*cemA*
	Protease	*clpP*
	Maturase	*matK*
Unkown	Conserved open reading frames	*ycf2, ycf4*

* Two gene copies in IRs

Analysis of SSRs

In the chloroplast genome of *C. bicolor* a total of 33 SSR loci were identified, ranging in length from 10 to 113 bp (Figure 2). Among these, 25 (75.8%) were mononucleotide repeats, all consisting of A or T bases, indicating a strong A/T bias in the chloroplast genome SSRs. Seven (21.2%) were dinucleotide repeats, all of the AT/TA type (Figures 2A, 2B). Figure 2C summarises the SSR counts by type.

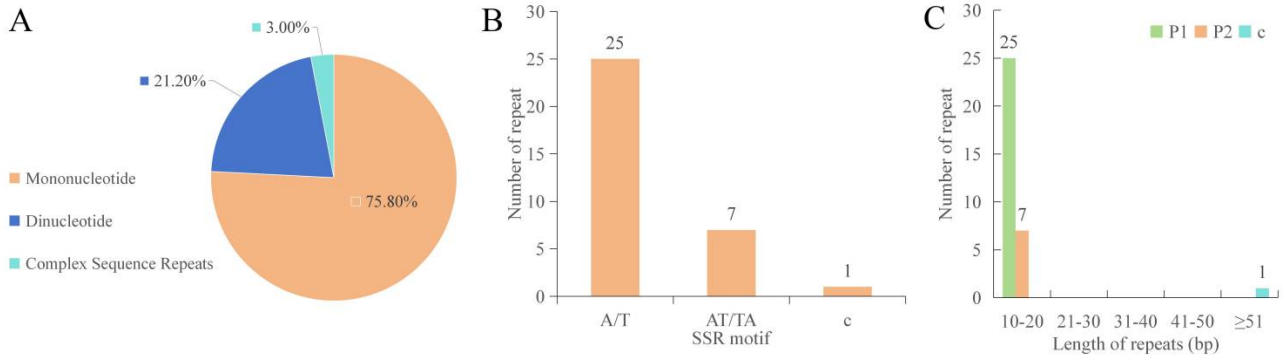

Figure 2. Analysis of repeat sequences in *C. bicolor* chloroplast genome: A. Composition and proportion of SSR types; B. Distribution of different SSR repeat motifs; C. Statistical count of SSR numbers by type

Additionally, one complex compound SSR was identified, located at positions 26,871-26,983 bp with a total length of 113 bp. It is composed of multiple T-repeat motifs interspersed with specific spacer sequences (Figure 2A). This structure may possess a high mutation potential and represents a potential hypervariable region. Analysis of the length distribution of the 33 identified

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

SSRs in the *C. bicolor* chloroplast genome showed that 32 SSRs were concentrated within the 10-20 bp range, indicating that its SSRs possess typical short-fragment characteristics. Within this range, there were 25 mononucleotide (A/T) repeats and 7 dinucleotide (AT/TA) repeats.

Comparative Analysis of Genome Structure in *C. bicolor* and Its Related Species

To clarify the sequence conservation characteristics of the chloroplast genomes of the genus *Caragana* and its related genera *Calophaca* and *Astragalus*, this study performed a global alignment of the chloroplast genomes of 11 species, including *C. bicolor*, using the mVISTA tool (Figure 3). With *Caragana acanthophylla* as the reference sequence, the results indicate that the overall structure and gene arrangement of the chloroplast genomes of these species are relatively consistent. However, the coding regions, particularly the *psaA* and *rbcL* gene regions, exhibit significantly higher sequence conservation compared to the non-coding regions. The non-coding regions show greater sequence variability, making them potential target regions for subsequent molecular marker development and studies on species differentiation.

Analysis of Codon Usage Bias in Chloroplast Genome of *C. bicolor*

To elucidate the codon usage bias characteristics of the *C. bicolor* chloroplast genome, this study statistically analysed the codon usage frequency and RSCU values of its CDS. The results show that the CDS of the *C. bicolor* chloroplast genome encode the 20 standard amino acids and 3 stop codons, utilising 64 codons in total, of which 61 are sense codons. The codon usage frequency ranges from 7 to 750 (Table 2).

Analysis of amino acid composition reveals variation in the total usage frequency of codons for different amino acids. Leucine (Leu) is the most abundant amino acid, corresponding to 6 synonymous codons with a total usage frequency of 1873. Among the leucine codons, TTA is used most frequently (628 times) and has an RSCU value of 2.01, making it the most preferred codon across the entire genome. Cysteine (Cys) is the least abundant amino acid, with only 2 synonymous codons and a total usage frequency of merely 204. Its codon TGC has the lowest single usage frequency (51) and an RSCU of 0.50, classifying it as a non-preferred codon. Furthermore, the total codon usage frequencies for glutamic acid (Glu, 941), phenylalanine (Phe, 990) and lysine (Lys, 847) are also relatively high. Their predominant codons GAA, TTT and AAA each has usage frequencys exceeding 600, indicating relatively high expression activity.

Using an RSCU threshold of 1.00 to determine codon preference, the analysis identifies 30 preferred codons (RSCU > 1.00) and 32 non-preferred codons (RSCU < 1.00) in the *C. bicolor* chloroplast genome. Among the preferred codons, 16 are highly preferred (RSCU ≥ 1.50). The stop codon TAA, with an RSCU of 1.75, is the dominant termination codon. Among the non-preferred codons, Leu-CTC has the lowest RSCU value (0.34), followed by Gly-GGC (0.35) and Ser-AGC (0.35). Ser-AGC is used only 75 times, making it the least frequently used codon in the entire genome and a typical rare codon.

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

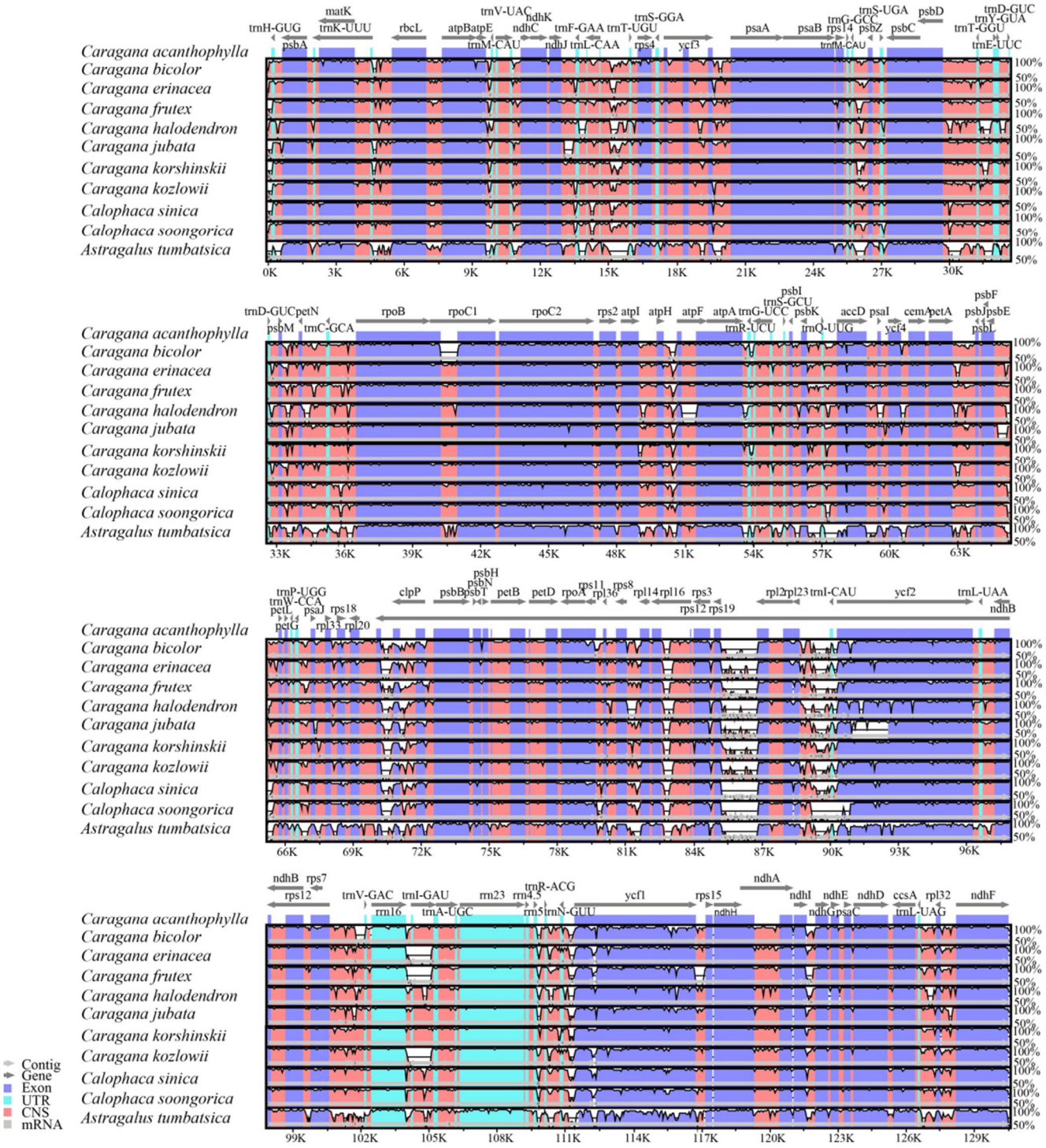

Figure 3. Comparative analysis of chloroplast genomes of 11 species across *Caragana*, *Calophaca* and *Astragalus* using mVISTA, with *Caragana acanthophylla* as reference. The x-axis represents chloroplast genome coordinates and the y-axis indicates per cent sequence identity (50-100%). Grey arrows show the transcriptional direction of genes (labelled above) and troughs descending to approximately 50% identity denote highly variable or absent regions. Protein-coding, non-coding, and tRNA regions are coloured blue/purple, red and cyan respectively.

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

Table 2. Codon usage and codon-anticodon recognition patterns of *C. bicolor*

Amino acid	Symbol	Codon	RSCU	Number	Amino acid	Symbol	Codon	RSCU	Number
A	Ala	GCA	1.07	277	P	Pro	CCT	1.58	297
A	Ala	GCC	0.63	162	P	Pro	CCG	0.44	82
A	Ala	GCG	0.42	108	P	Pro	CCA	1.17	220
A	Ala	GCT	1.89	488	P	Pro	CCC	0.81	153
C	Cys	TGT	1.50	153	Q	Gln	CAA	1.58	500
C	Cys	TGC	0.50	51	Q	Gln	CAG	0.42	133
D	Asp	GAT	1.64	590	R	Arg	AGG	0.60	102
D	Asp	GAC	0.36	129	R	Arg	CGA	1.39	236
E	Glu	GAA	1.47	693	R	Arg	AGA	1.65	280
E	Glu	GAG	0.53	248	R	Arg	CGC	0.40	67
F	Phe	TTT	1.38	681	R	Arg	CGG	0.49	83
F	Phe	TTC	0.62	309	R	Arg	CGT	1.47	249
G	Gly	GGT	1.45	469	S	Ser	AGC	0.35	75
G	Gly	GGG	0.58	186	S	Ser	AGT	1.33	285
G	Gly	GGA	1.62	525	S	Ser	TCT	1.69	362
G	Gly	GGC	0.35	113	S	Ser	TCG	0.58	124
H	His	CAT	1.56	335	S	Ser	TCA	1.14	245
H	His	CAC	0.44	95	S	Ser	TCC	0.92	197
I	Ile	ATC	0.63	329	T	Thr	ACC	0.75	169
I	Ile	ATA	0.92	476	T	Thr	ACA	1.20	269
I	Ile	ATT	1.45	750	T	Thr	ACG	0.38	85
K	Lys	AAG	0.50	212	T	Thr	ACT	1.68	377
K	Lys	AAA	1.50	635	V	Val	GTA	1.48	378
L	Leu	TTG	1.25	389	V	Val	GTC	0.46	117
L	Leu	TTA	2.01	628	V	Val	GTG	0.56	143
L	Leu	CTA	0.78	245	V	Val	GTT	1.50	381
L	Leu	CTT	1.26	393	W	Trp	TGG	1.00	311
L	Leu	CTC	0.34	106	Y	Tyr	TAC	0.36	117
L	Leu	CTG	0.36	112	Y	Tyr	TAT	1.64	529
M	Met	ATG	1.00	396	*	Ter	TAG	0.44	7
N	Asn	AAC	0.45	176	*	Ter	TGA	0.81	13
N	Asn	AAT	1.55	611	*	Ter	TAA	1.75	28

Nucleotide Diversity Analysis

This study conducted a nucleotide diversity (Pi) analysis on 97 coding genes from the chloroplast genomes of *C. bicolor* and its nine closely related species (Figure 4). The results show that the Pi values for all genes ranged from 0.00,000 to 0.06,362, with an average of 0.01,427. Six tRNA genes (*trnE-UUC*, *trnN-GUU*, *trnK-UUU*, *trnP-UGG*, *trnM-CAU*, *trnW-CCA*) are completely conserved (Pi = 0.00,000).

Following general screening criteria for highly variable genes in plant chloroplast genomes and using Pi ≥ 2 times the average value as the core statistical basis, the screening threshold was adjusted for data stability to Pi ≥ 0.03. This process ultimately identifies nine highly variable genes (Figure 4). The Pi values of these genes are 2.12 to 4.46 times the average, significantly exceeding the overall level of variation. Functionally, these genes are primarily involved in pathways closely related to environmental adaptation such as ribosomal protein synthesis, amino acid transport and protein quality control. Among them, *trnL-CAA* (Pi = 0.06,362) and *clpP* (Pi = 0.06,319) exhibit the highest degree of variation and can be considered core candidate molecular markers for species identification and phylogenetic analysis of *C. bicolor*.

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

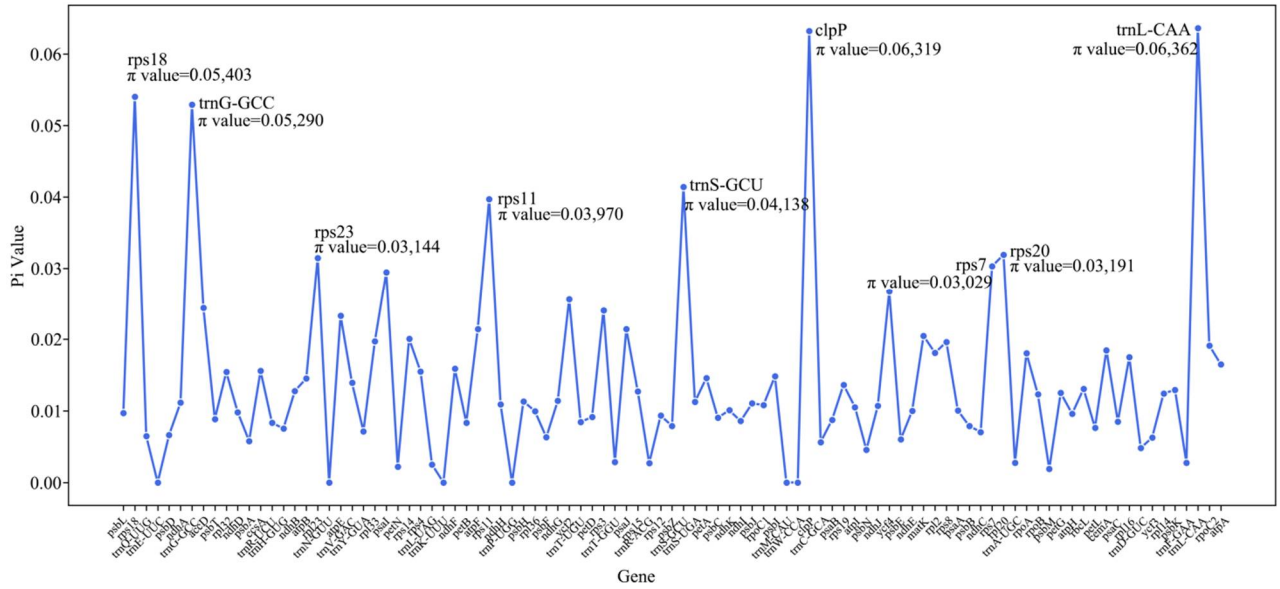

Figure 4. Nucleotide diversity (π) values of chloroplast coding regions in *C. bicolor* and its nine related species

Phylogenetic Analysis of Species Clusters

Using the maximum likelihood method, a phylogenetic analysis was conducted based on the complete chloroplast genome sequences of 56 Fabaceae species obtained from the NCBI database, along with the chloroplast genome sequence of *C. bicolor*, with *Glycyrrhiza uralensis* designated as the outgroup (Figure 5). The bootstrap support values for the inferred phylogenetic relationships in the maximum likelihood tree range from 44 to 100, indicating strong support for nearly all clades (Figure 5).

In the phylogenetic tree, all *Caragana* species cluster into a highly supported monophyletic clade, confirming the monophyly of the genus *Caragana*. Within this clade, *C. bicolor* and *C. tibetica* form a sister group, indicating their closest phylogenetic relationship. This sister lineage, together with species such as *C. jubata* and *C. sinica*, constitutes the core branch of the genus *Caragana*, thereby clarifying the taxonomic placement of *C. bicolor* within the genus. Species of *Calophaca* (*C. sinica* and *C. soongorica*) form a highly supported sister group with the *Caragana* clade, suggesting a close phylogenetic relationship between these two genera within the Fabaceae family. The *Hedysarum* group independently clusters as a monophyletic lineage, forming a sister group to the *Caragana* and *Calophaca* clades, indicating that these three groups share a relatively recent common ancestor but have since diverged into independent evolutionary lineages. *Astragalus* forms the most species-rich group, comprising multiple highly supported subclades, reflecting its characteristic pattern of radiative diversification. *Oxytropis* shows a close phylogenetic relationship with *Astragalus*. The *Campylotropis* group forms a well-supported monophyletic branch. However, its branching support relative to the aforementioned groups is lower, suggesting it may represent a relatively ancestral or early-diverging lineage. *Gueldenstaedtia* and *Galega* appear as independent small lineages, displaying more distant phylogenetic relationships with the aforementioned groups.

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

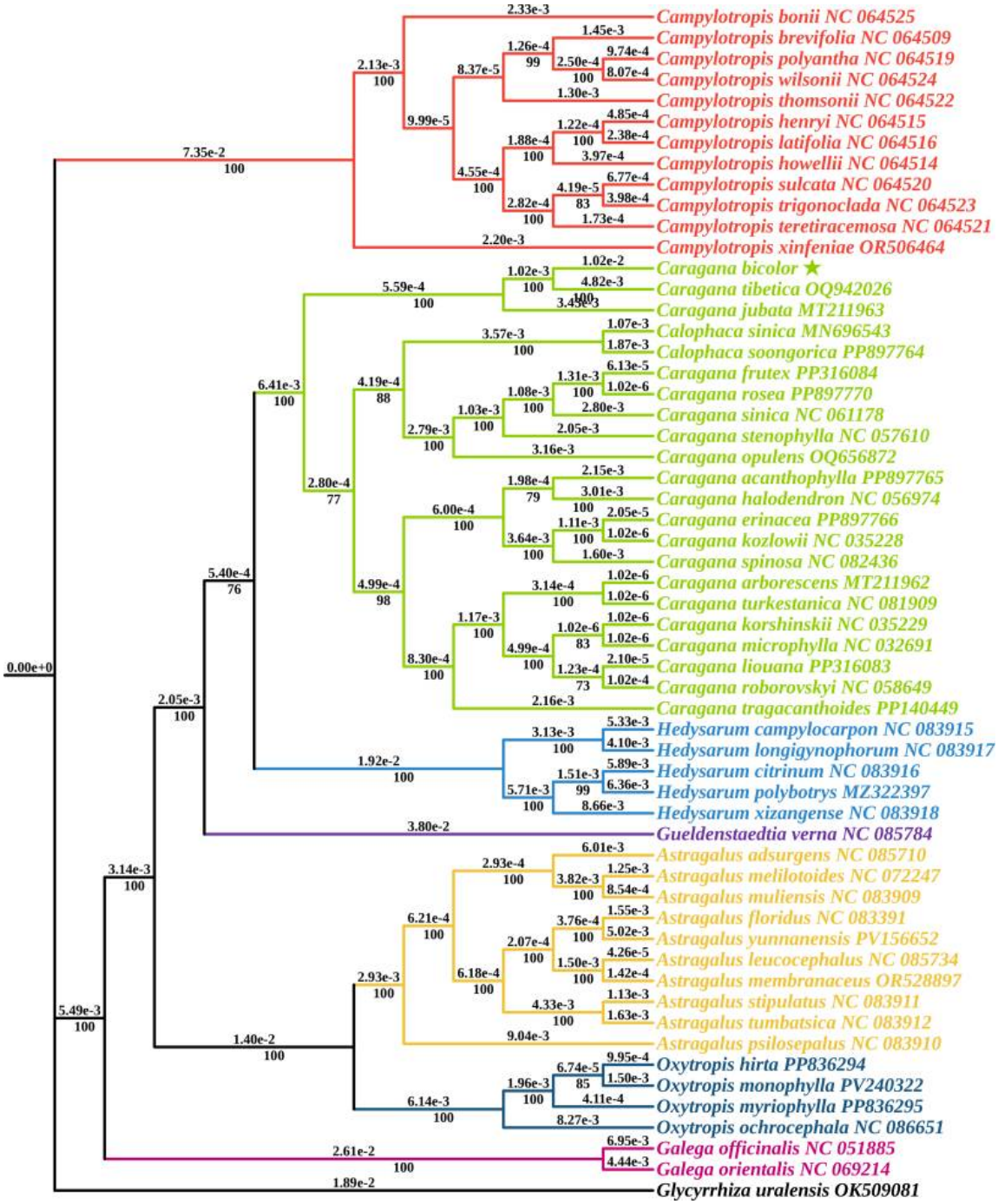

Figure 5. Maximum likelihood phylogenetic tree of 57 complete chloroplast genomes. Bootstrap values and branch lengths are shown near each node. Different colours represent differences in the clustering results.

DISCUSSION

This study sequenced, assembled and annotated the chloroplast genome of *C. bicolor*. The genome exhibits core features (genome size, GC content and gene number) that are highly consistent with those previously reported for *C. tibetica* (128,433 bp, 34.65% GC, 110 genes) and *C. turkestanica* (129,453 bp, 34.30% GC, 111 genes) [10] and eight other *Caragana* species (128,458-135,401 bp, GC content 34.3%-34.8%, 110 genes) [7]. This consistency reflects the conservation of chloroplast genome size, GC content and gene number within the genus *Caragana*. In *C. bicolor*

the genome lacks the typical quadripartite structure. This feature is consistent with reported species in the genus, such as *C. arborescens*, *C. jubata*, *C. roborovskyi*, *C. opulens*, *C. stenophylla*, *C. tibetica*, *C. turkestanica* and *C. erinacea* [6, 9-11], suggesting that IR region degeneration/loss may be a common characteristic of the genus *Caragana*, rather than a unique phenomenon in *C. bicolor*. The IR region serves as a stability-regulating element in chloroplast genomes. Its degeneration is relatively common within the IR-lacking clade of Fabaceae [6, 9].

In terms of gene structure, among the 17 intron-containing genes in the chloroplast genome of *C. bicolor*, *ycf3* and *clpP* contain two introns each while the remaining 15 genes possess a single intron. This pattern is identical to that of *C. tibetica*, whereas *clpP* in *C. turkestanica* contains only one intron, reflecting subtle differentiation within the genus [9]. Furthermore, four single-intron genes *trnI-GAU*, *trnA-UGC*, *ndhB* and *rpl2* are located in the IR regions and retain copies on both strands. Even with IR region degeneration, these genes maintain the stability of core functional elements. This evolutionarily conserved feature, which prioritises functional maintenance, is highly consistent across the genus *Caragana*, ensuring the stability of essential processes such as ribosome assembly and photosynthesis [9, 11].

Among the 33 SSR loci identified in the chloroplast genome of *C. bicolor*, mononucleotide repeats account for 75.8% and consist exclusively of A/T repeats, while dinucleotide repeats account for 21.2% and are all of the AT/TA type. These findings are consistent with the characteristics reported by Liu et al. [10] for nine *Caragana* species, of which mononucleotide repeats represent 62.76%-63.64% and show an A/T bias. They are also consistent with Cui et al. [7], who observed that mononucleotide repeats are the dominant type and A/T bases are overwhelmingly predominant, which confirms the nucleotide composition bias in SSR loci within the genus *Caragana*. Additionally, this study identified in *C. bicolor* one compound SSR of 113 bp, which may possess higher mutation potential and could serve as a specific molecular marker for this species.

The mVISTA global alignment results show that the overall structure and gene arrangement of the chloroplast genomes among 11 species, including *C. bicolor* and species of *Caragana*, *Calophaca* and *Astragalus*, are highly consistent. However, the coding regions *psaA* and *rbcL* exhibit higher conservation than the non-coding regions. These findings align with the conclusions of Liu et al. [9] and Chen et al. [1], who also reported greater conservation in coding regions and concentrated variation in non-coding regions based on mVISTA comparisons. Nucleotide diversity (Pi) analysis identified nine highly variable genes (Pi $\geq$ 0.03). Although these genes do not completely overlap with the five highly variable regions *rps2-atpI*, *accD-psaI-ycf4*, *cemA-petA*, *psbN-psbH* and *rpoA-rps11*) reported by Liu et al. [10] or with the variable regions *trnK-UUU* and *rbcL* described by Yuan et al. [11], they are all concentrated in non-coding regions or a few functional genes. This further validates the distribution pattern of highly variable regions within the genus *Caragana*.

The protein-coding sequences of the *C. bicolor* chloroplast genome encode the 20 standard amino acids and 3 stop codons, utilising 64 codons in total, of which 61 are sense codons. Leucine (Leu) shows the highest usage frequency (1,873 times), while cysteine (Cys) exhibits the lowest (204 times). A total of 30 preferred codons (RSCU > 1.00) were identified, most of which end with A or U. These findings are highly consistent with the codon usage patterns reported for *Caragana* species by Liu et al. [10] and Cui et al. [7]. *C. bicolor* also displays species-specific characteristics. The preference for the dominant codon TTA (RSCU = 2.01) is stronger than that reported for all

other *Caragana* species. Furthermore, the dominance of the stop codon TAA (RSCU = 1.75) is at a relatively high level within the genus.

The phylogenetic tree based on 57 chloroplast genomes shows that *C. bicolor* and *C. tibetica* form a sister group (bootstrap value 100), which further clusters with *C. jubata* to constitute a core branch, collectively belonging to the core clade of the genus *Caragana*. This differs from the conclusion of Yuan et al. [11], who proposed a closer relationship between *C. bicolor* and *C. opulens*. The discrepancy is likely attributable to differences in the selection of closely related species across studies. While Yuan et al. [11] included only eight *Caragana* species, this study expanded the analysis to 56 closely related species, resulting in higher sample coverage and greater reliability of the findings. Simultaneously, this study found that *Calophaca sinica* and *Calophaca soongorica* are nested within the evolutionary clade of *Caragana*, with *Caragana* and *Calophaca* forming a sister group. This result aligns with the phylogenetic findings of Zhang et al. [32], suggesting that *Caragana* may not be monophyletic and that *Calophaca* is nested within it. Furthermore, Liu et al. [10] identified *C. tibetica* as most closely related to *C. jubata*, whereas in this study, *C. tibetica* clusters with *C. bicolor*. This difference may be related to geographical distribution. Both *C. tibetica* and *C. bicolor* are distributed in high-altitude regions of south-western China, with significant geographical overlap, while *C. jubata* has a more northern distribution range. Geographical isolation may have promoted species differentiation. This observation is consistent with the conclusion of Yuan et al. [11] that the phylogenetic relationships within *Caragana* are associated with altitude and geographical distribution, further supporting the notion that geographical isolation may be an important factor influencing species differentiation in *Caragana*.

CONCLUSIONS

Through the integration of existing research findings of *Caragana*, this study has clarified the core characteristics and evolutionary position of the *C. bicolor* chloroplast genome. The results show that the complete chloroplast genome of *C. bicolor* is 131,396 bp in length with a total GC content of 35.14%. Its structure is not typically quadripartite, supporting the observed degeneration of the IR region in the genus *Caragana* and revealing new characteristics of structural variation in the *Caragana* chloroplast genome. A total of 110 genes were annotated in the *C. bicolor* chloroplast genome, comprising 75 protein-coding genes, 31 tRNA genes and 4 rRNA genes. Phylogenetic analysis indicates that *C. bicolor* and *C. tibetica* are most closely related. Together with species such as *C. jubata* and *C. sinica*, they form the core clade of the genus *Caragana*. This study provides a molecular basis for taxonomic and evolutionary studies of the genus *Caragana*.

DATA AVAILABILITY STATEMENT

The original data of chloroplast genome *C. bicolor* have been deposited into China National GeneBank Sequence Archive with accession number CNX1292652.

ACKNOWLEDGEMENTS

This research was supported by the Natural Science Foundation of Henan (Grant No. 252300423627) and the High-level Talent Scientific Research Startup Project of North China University of Water Resources and Electric Power (Grant No. 202110007).

Maejo Int. J. Sci. Technol. **2026**, *20*(02), 211-225

REFERENCES

1. X. Cui, E. Li, J. He, Y. Wang, C. Shang, B. Zhong, J. Viruel, W. Dong and Z. Zhang, "Ancient hybridization drives arid adaptation and species diversification in Caragana (Fabaceae)", *New Phytol.*, **2025**, *247*, 2454-2472.
2. M. L. Qiao, Z. J. Ji, X. X. Yuan and Q. H. Shi, "Study on saline-alkali tolerance characteristics of symbiotic rhizobium by Leguminous Caragana", *J. Inner Mongolia Minzu Univ. (Nat. Sci. Ed.)*, **2024**, *39*, 46-51 (in Chinese).
3. Y. Fan, X. Jin, M. Wang, H. Liu, W. Tian, Y. Xue, K. Wang, H. Li and Y. Wu, "Flower morphology, flower color, flowering and floral fragrance in *Paeonia* L.", *Front. Plant Sci.*, **2024**, *15*, Art.no.1467596.
4. C. Suwanphakdee, P. Sutthinon and T. R. Hodkinson, "Characterization of anatomical characters of *Peperomia* (Piperaceae) from Asia for taxonomy", *Bot. J. Linn. Soc*., **2024**, *205*, 268-291.
5. S. A. Rather, S. Wang, M. D. Dwivedi and Z. Y. Chang, "Molecular phylogeny and systematic evaluation of the *Caragana opulens* species complex (Fabaceae, Papilionoideae) based on the molecular and morphological data", *Phytotaxa*, **2021**, *478*, 179-200.
6. W. Chen, X. Zhang, Q. Li, L. Jiang, M. Jiang, X. Bai and L. Wang, "Comparative analyses of the complete chloroplast genomes of four *Caragana* species, a specific genus distributed in arid and semi-arid area", *Planta*, **2025**, *262*, Art.no.16.
7. X. Cui, K. Liu, E. Li, Z. Zhang and W. Dong, "Chloroplast genomes evolution and phylogenetic relationships of *Caragana* species", *Int. J. Mol. Sci.*, **2024**, *25*, Art.no. 6786.
8. X. Li, L. Liu, Q. Ren, T. Zhang, N. Hu, J. Sun and W. Zhou, "Analysis of synonymous codon usage bias in the chloroplast genome of five Caragana", *BMC Plant Biol.*, **2025**, *25*, Art.no.322.
9. L. Liu, H. Li, J. Li, X. Li, N. Hu, H. Wang and W. Zhou, "Chloroplast genome analyses of *Caragana arborescens* and *Caragana opulens*", *BMC Genom. Data*, **2024**, *25*, Art.no.16.
10. L. Liu, H. Li, J. Li, X. Li, N. Hu, J. Sun and W. Zhou, "Chloroplast genomes of *Caragana tibetica* and *Caragana turkestanica*: Structures and comparative analysis", *BMC Plant Biol.*, **2024**, *24*, Art.no.254.
11. M. Yuan, X. Yin, B. Gao, R. Gu and G. Jiang, "The chloroplasts genomic analyses of four specific Caragana species", *PLOS ONE*, **2022**, *17*, Art.no. e0272990.
12. W. Wang, C. Tan, L. Li, X. Li, L. Zhang, X. Li, J. Wang, Z. He, T. Yang, K. Ma, Q. Hu, W. Yang, Z. Li, M. Zhang, W. Du, F. Yang, Z. Xu, X. Ma, J. Tong, J. Cai, C. Hua, F. Chen, L. You, L. Li, W. Zeng, B. Wang, X. Xu and X. Wei, "The China National GeneBank Sequence Archive (CNSA) 2024 update", *Hortic. Res*., **2025**, *12*, Art.no. 36.
13. L. A. Cauz-Santos, Z. P. da Costa, M. A. Sader, C. van den Berg and M. L. C. Vieira, "Chloroplast genomic insights into adaptive evolution and rapid radiation in the genus *Passiflora* (Passifloraceae)", *BMC Plant Biol.*, **2025**, *25*, Art.no.192.
14. Y. Jiang, H. Li, M. Wu, X. Zhang, S. Baasanmukh, H. Li, H. Sun and S. Chen, "Comparative chloroplast genomes of *Incarvillea* species (Bignoniaceae) unveiled genomic diversity and shed light on phylogenetic relationships", *BMC Plant Biol*., **2025**, *25*, Art.no.399.
15. C. Y. Chang and F. C. Shi, "Taxonomical notes on *Caragana bicolor* Kom. and *C. franchetiana* Kom.", *Acta Bot. Boreal.-Occident. Sin.*, **2010**, *30*, 2331-2333 (in Chinese).
16. Y. Zhou, X. L. Cheng, Y. Gao, Y. L. Sun, H. Wang, J. P. Bai and J. Liu, "Development and application of high-throughput semi-automatic genome DNA extraction instrument", *J. Plant Genet. Resour.*, **2022**, *23*, 1224-1232 (in Chinese).

17. Y. Fan, L. Chen, Y. Wu, X. Wang, W. Tian, M. Ge, C. Li and Y. Xie, "The complete chloroplast genome of *Synotis solidaginea* and phylogenetic analysis", *Ecol. Evol.*, **2025**, *15*, Art.no. e72806.
18. S. Chen, Y. Zhou, Y. Chen and J. Gu, "Fastp: An ultra-fast all-in-one FASTQ preprocessor", *Bioinform.*, **2018**, *34*, i884-i890.
19. S. W. Wingett and S. Andrews, "FastQ Screen: A tool for multi-genome mapping and quality control", *F1000Res.*, **2018**, *7*, Art.no.1338.
20. J. J. Jin, W. B. Yu, J. B. Yang, Y. Song, C. W. dePamphilis, T. S. Yi and D. Z. Li, "GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes", *Genome Biol.*, **2020**, *21*, Art.no. 241.
21. L. Shi, H. Chen, M. Jiang, L. Wang, X. Wu, L. Huang and C. Liu, "CPGAVAS2, an integrated plastome sequence annotator and analyzer", *Nucleic Acids Res.*, **2019**, *47*, W65-W73.
22. M. Lohse, O. Drechsel and R. Bock, "OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes", *Curr. Genet.*, **2007**, *52*, 267-274.
23. S. Beier, T. Thiel, T. Münch, U. Scholz and M. Mascher, "MISA-web: A web server for microsatellite prediction", *Bioinform.*, **2017**, *33*, 2583-2585.
24. K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin and I. Dubchak, "VISTA: Computational tools for comparative genomics", *Nucleic Acids Res.*, **2004**, *32*, W273-W279.
25. L. J. Huang, H. X. Yu, Z. Wang and W. B. Xu, "CPStools: A package for analyzing chloroplast genome sequences", *iMetaOmics*, **2024**, *1*, Art.no. e25.
26. P. M. Sharp and W. H. Li, "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications", *Nucleic Acids Res.*, **1987**, *15*, 1281-1295.
27. C. Y. Xiang, F. Gao, I. Jakovlić, H. P. Lei, Y. Hu, H. Zhang, H. Zou, G. T. Wang and D. Zhang, "Using PhyloSuite for molecular phylogeny and tree-based analyses", *iMeta,* **2023**, *2*, Art.no. e87.
28. K. Katoh, J. Rozewicki and K. D. Yamada, "MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization", *Brief. Bioinform.*, **2019**, *20*, 1160-1166.
29. S. Capella-Gutiérrez, J. M. Silla-Martínez and T. Gabaldón, "TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses", *Bioinform.*, **2009**, *25*, 1972-1973.
30. V. Ranwez, E. J. P. Douzery, C. Cambon, N. Chantret, and F. Delsuc, "MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons." *Mol. Biol. Evol.*, **2018**, *35*, 2582-2584.
31. L. T. Nguyen, H. A. Schmidt, A. von Haeseler and B. Q. Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies", *Mol. Biol. Evol.*, **2015**, *32*, 268-274.
32. L. Wang, X. Jiang, W. Jiao, J. Mao, W. Ye, Y. Cao, Q. Chen and Q. Song, "Pangenome analysis provides insights into legume evolution and breeding", *Nat. Genet.*, **2025**, *57*, 2052-2061.
33. M. Zhang, P. W. Fritsch and B. C. Cruz, "Phylogeny of *Caragana* (Fabaceae) based on DNA sequence data from rbcL, trnS-trnG, and ITS", *Mol. Phylogenet. Evol.*, **2009**, *50*, 547-559.