Contents lists available at UGC-CARE

International Journal of Pharmaceutical Sciences and Drug Research

[ISSN: 0975-248X; CODEN (USA): IJPSPP]

Available online at www.ijpsdronline.com

Research Article

Molecular Characterization of Maturase K Gene and Protein Prediction of *Vanda spathulata* (L.) - An Ayurinformatics Approaches

Jeline Rani J[1*], Nandagopalan V[2]

[1]*Department of Botany, Bishop Heber College, Affiliated to Bharathidasan University, Trichy, Tamil Nadu, India.*
[2]*Department of Botany, National College, Affiliated to Bharathidasan University, Trichy, Tamil Nadu, India.*

ARTICLE INFO

Article history:
Received: 06 April, 2023
Revised: 30 April, 2023
Accepted: 03 May, 2023
Published: 30 May, 2023

Keywords:
matK gene, *Vanda spathulata*, Orchid, Protein modeling

DOI:
10.25004/IJPSDR.2023.150319

ABSTRACT

Vanda spathulata (L.) Spreng. is an important plant of the orchid family of botanical and horticultural importance in the Ayurvedic medicinal system. The Maturase K (*matK*) gene of plant chloroplast is a strong marker in studying plant molecular systematics, evolution, and genetic polymorphism and is considered as the standard DNA barcode for plants. The gene has an important role in the phylogenetic reconstruction of terrestrial plants. The present study investigated the phylogenetic characterization of *matK* gene obtained from the plant species *V. spathulata* (L.) through extraction, purification, PCR amplification, and gel elution. The nucleotide BLAST search revealed similarity with 103 sequences in various closely and distantly related orchid taxa. Mega software and a real-time divergence tree discovered *Apostasia odorata* has the primitive species for all the sequences, including the *V. spathulata* which diverged about 114 million years ago. Limited gene flow was observed among the *vanda* and related taxa and the Tajima neutrality test was negative, indicating the purifying selection in the *matK* gene sequence, that evolutionary selection had played a major role rather than the neutral evolution. The species and related other species exhibited higher polymorphism at the nucleotide level. Secondary structure was predicted for the single-stranded *matK* gene of *V. spathulata* and the ancestor plant species *A. odorata* using the Zucker MFold web server and translated through a comparison of ORF with NCBI and BLASTN. The protein was modeled in SWISS MODELLER and compared between the two species. The sequence had high similarity with its ancestor gene sequence and the protein modeled showed a complex structure than the ancestor protein due to significant evolutionary changes.

Introduction

Orchid (Orchidaceae) is the largest known plant family comprising flowering plants with approximately over twenty thousand taxa[1] and is considered to be primitive among other flowering plant groups for it is estimated to be originated over 80–40 million years ago.[2] Over the millions of years, the plant family had undergone many evolutionarily significant changes; the most specific traits are high variability in the growth habitat, maximum species-level diversity, and modified flowering patterns.[3] Owing to its attractive flowers, the plant species are widely used as ornamental yet very much in medicinal systems.[4] *Vanda spathulata* (L.) Spreng. is one of the flowering plants of the Orchid family with a narrow geological distribution in South India and Sri Lanka. The endemic distribution of the plant includes zones at sea levels with partial arid habitats to high altitudes zones (over 1000 meters) such as dry deciduous forests and scrub jungles.[5] The plant, popularly known as yellow *vanda* owing to its immaculate bright yellow color and small-sized flowers, has a blooming period of 60 to 75 days, an attraction of the orchid breeders.[6] The plant is either a tetraploid or a hexaploid, i.e., the chromosome duplicates into four or six sets instead of two sets (diploid).[7] The plant is now designated as *Taprobanea spathulata.* Even though the orchid family exhibits the highest level of genetic and infraspecific diversity, in recent terms, some species of the plant group are rapidly deteriorating, attributing to the uncontrolled exploitation

***Corresponding Author:** Ms. Jeline Rani J

Address: Department of Botany, Bishop Heber College, Affiliated to Bharathidasan University, Trichy, Tamil Nadu, India.

Email ✉**:** jelineranibhc@gmail.com

Tel.: +91-9843717815

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

as ornaments and/or their habitat deterioration owing to the expansion of modern civilization; these factors in addition to the symbiote relationship with the pollinators and fungal groups are resulting in the diversity loss at an alarmingly faster rate, which is the significant property among these plants.[8] Hence, monitoring the genetic diversity at the population level, most importantly among the terrestrial orchids most vulnerable to deteriorating factors, is extensively important for the conservation of the plant taxa from extinction.

Genetic diversity is important among the plant population for the survival and conservation of genetic traits. Analysis of genetic diversity in the species gene pool and the population can facilitate understanding the evolutionary aspects such as growth pattern, germline survival capability, and ultimately the knowledge of genetic drift, mutation, and gene flow.[9,10] DNA barcoding is one of the most widely used methods for the study of genetic variation in plants. The method proved to be a useful tool in the conservation of biodiversity hotspots for the assessment of species identification and also for the endangered species of orchids in monitoring its international trade.[11] In principle, this method is the inverse of a genomics system for species identification, a DNA barcode consisting of a standardized short DNA sequence, usually 400 to 800 bp, used to efficiently recognize and characterize all plant species.[12] The PCR amplicons produced from a particular genome site of a species can be used to barcode, in other terms, identify that organism and establish a distinction between other species by comparing with a DNA barcode library. For animals, the standardized species-level barcode that had been selected is the mitochondrial gene for cytochrome c oxidase subunit 1 (CO1) which has a 600 bp sequence.[13] The genes used as the barcode for identifying plants include *matK*, rbcL, trnH-psbA, and ITS. The *matK* is a commonly used gene and it has a 2-3 times higher evolution rate than the rbcL gene. Lahaye *et al.*, 2008 studied and reported the amplification rate as 100% and a species discrimination rate of over 90% for the *matK* gene of 1667 plants.

The chloroplast encoding gene, *matK* stands for maturase K gene, is located between the 5' and 3' exons of the trnK on the large single-copy section next to the inverted repeats and encodes the tRNA- Lysine (UUU). Formerly termed orfK, this gene has approximately 1500 bp. The protein encoded by the *matK* gene acts as a maturase. As the enzyme, maturase encoded by the gene has the function of non-autocatalytic host-specific splicing of the intron from premature RNAs i.e., group II introns, thus stimulation of its innate self-splicing properties. *matK* is responsible for the post-transcriptional splicing of the RNA transcripts of the genes trnA, trnK, trnK, rpl2 rpsl2, and atp2. The proteins encoded by those genes are essential for the normal functioning of the chloroplast. Hence, *matK* is an important protein in plant metabolism. *matK* is also the most important gene for plant molecular systematics and evolutionary research. The features of this gene namely the ideal size, high nucleotide substitution rate, large proportion of variation at the nucleic acid level at first and second codon position, mutually conserved sectors, and low transition/transversion ratio signifies its role in evolutionary research.[14]

The current study aims to understand the genome level diversity and gene flow of *V. spathulata* to diagnose the current diversity level using the chloroplast gene, *matK*. The secondary structure of the protein from the gene and amino acid translation were predicted. Protein was modeled and the properties were analyzed.

Materials and Methods

Source of the Plant Material

Botanical name: *Vanda spathulata* (L.) Spreng.
Family: Orchidaceae

It is a scrambling epiphytic herb with monopodial growth and extended lengthwise maximum up 200 cm provided with a suitable supporting host. Hosts include small trees and bushes. The normal roots are vermiform, mostly ash green in color, and approximately 4 to 6 mm thick. The leaves are oblong or spoon-shaped, occurring in 2 ranks, 2 lobed at the tip, base sheathing, leathery, stake, and grow up to 6 x 1.5 cm. Flowers are yellow in color, borne in leaf-opposed racemes 6–8 in number, and are 4 cm across. Sepals are obovate, 2.5 cm, narrowing to the base. They have petals similar to sepals. 3 lobed lips; spur up to 0.5 cm and column up to 1-cm.[5]

DNA Extraction and Purification

Liquid nitrogen was used to homogenize 100 mg of leaf tissue. The powdered leaf tissue was centrifuged by using a microcentrifuge. One minute of overtaxing follows the addition of 400 microliters of buffer PL1. 10 microliters of RNase A solution were added and stirred well, followed by mixing, incubated at 65°C for 10 minutes. The lysate was again centrifuged at 11000 x g using a Nucleospin filter. The filter was discarded, and the liquid was collected. To that 350 mL of buffer, PC was added and mixed thoroughly. Centrifuged for 1-minute and flow-through was discarded after transferring the solution to the Nucleospin Plant II column. The course through fluid was disposed of after centrifugation at 11000 x g briefly and expansion of 400 mL of support PW1. A volume of 700 µL was then added, centrifuged at 11000 x g, and the course through liquid was taken out. 200 µL of PW2 was added and centrifuged at 11000 x g for 2 minutes to dry the silica layer. Cushion PE was added to the section and brooded at 65°C for 5 minutes prior to being moved to a new 1.7 mL tube. The DNA was then eluted by centrifuging the column at 11000 x g for one minute. It was kept at a temperature of 4°C to preserve the DNA.

At a concentration of 40 ng/mL of DNA, samples were incubated for one hour at 37°C with RNase A (20 mg/mL). At 37°C for 60 minutes, the sample was treated with Proteinase K (10 mg/mL). For five minutes, the DNA solution was swirled in an equal amount of saturated phenol, chloroform, and isoamyl alcohol (25:24:1). For the next 5 minutes, the tube was spun at 10,000 rpm to collect the supernatant. Then, chloroform was used to extract isoamyl alcohol (24:1). To precipitate the purified DNA, 1/10 volume of 3M sodium acetate (pH 5.6) was added along with 2.5 times (v/v) cooled ethanol (95%). A second wash with 70% ethanol eliminated any remaining salts, then the DNA was pelleted and dried under a vacuum to remove any remaining moisture. Both spectrophotometric and agarose gel analysis were used to examine the DNA's quality and amount. At room temperature, the pellet was dissolved in a small amount of TE (10:1) buffer and then kept at 20°C.

Detection of DNA

Agarose gel electrophoresis was utilized to check the virtue of the recuperated DNA. 5 µL of DNA was blended in with 1-mL of 6X gel-stacking cushion (0.25% bromophenol blue, 30% sucrose in TE cradle pH-8.0). 0.8% agarose gel ready in 0.5X Tris-Borate-EDTA cradle containing 0.5 µg/mL ethidium bromide was utilized to stack the samples for the measurement. TBE was utilized as electrophoresis support at 75 volts until the bromophenol color front arrived at the gel's lower part. When electrophoresis was complete, the image was taken utilizing a UV transilluminator (Genei) and a gel documentation framework (GDS) (Bio-Rad).

Amplification of *matK* Gene

Primers used to amplify the *MATK* genes from DNA samples of the plant studied were F: 5'-CGTACAGTACTTTTGTGTTTACGAG -3' and R: 5'- ACCC AGTCCATCTGGAAATCTTGGTTC -3'. BIOER Genepro thermal cycler was used to amplify the plant *matK* gene to its full length. The PCR enhancement combination included approximately 100 ng of genomic DNA, 20 pmol of every preliminary, 2.0 mM of $MgCl_2$, 0.225 mM of each dNTP (deoxynucleotide triphosphate) and 1U Taq DNA polymerase (Geneaid) as well as 0.2 mM of random primer and 1X PCR buffer. Using a thermal cycler, we were able to amplify the signal (BioneerMy Genie 32 Thermal Block, Global Genomics Partner). 35 cycles of PCR were performed at 95, 57, and 72°C for 1-minute each. At 72°C, the extension time lasted 10 minutes. A gel electrophoresis was used to examine the amplified products (4 l each) in 1.3% (w/v) agarose gel in 1X TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA [pH8]). The gel was then stained with ethidium bromide (1 g/mL) and electrophoresed at 60 volts for 3 hours. One Kb and 100 bp ladders were used as the molecular marker (MBI, Fermentas). A UV transilluminator was used to examine the gel and a gel documentation device was used to photograph it. The amplified product was eluted using the Gel elution kit method and stored at -20°C for further use.

Ayurinformatics - BLAST and Phylogenetic Tree Construction

Nucleotide sequences of the sequenced plant samples were compared with the NCBI Gene Bank Database using the Basic Local Alignment Search Tool (BLAST) to identify known closely related sequences (http://www.ncbi.nlm.nih.gov/blast/blast.cgi).

To construct the phylogenetic tree, the raw *matK* gene sequences of the *V. spathulata* were aligned with the MEGA-11 software (http://www.megasoftware.net). The most related *matK* sequences [retrieved from http://www.ncbi.nlm.nih.gov/blast/blast.cgi. *matK* sequences] were also selected for the divergence of the phylogenetic tree. Maximum likelihood methods and the Jukes cantor model did the evolution analysis.[15] 1000 replicates inferred the bootstrap consensus. Minimum evolution methods created the real-time divergence tree.[16] The gene flow was calculated by Qu and Takaiwa[17] methods. Haplotype diversity by Garg method[18]; Eta-Total number of mutation, Tajma relative rate test, polymorphic sites, Secondary structure[19] and protein modeling.[20]

Results

The PCR-amplified partial *matK* gene sequence (size 783bp) from *V. spathulata* was used to retrieve related and highly similar orthologue sequences from diverse species of orchids. The phylogenetic tree was constructed using the minimum evolution method with 1000 bootstraps among related orchids composed of 103 sequences derived from different taxa, which showed two clades formed from a common ancestor; the first clade gave raise to *Apostasia odorata* while the second clade diverged into two clades one leading to *C. nobilior* and *L. luddemanni*; other with many subclades containing the other orchid taxa including *V. spathulata* (Fig. 1).

The real-time divergence time tree was generated using the existing phylogenetic tree generated using the minimum evolution method. Calibration points with fixed time (22 Million Years) were set between *C. nobilior* and *L.luddemanni* nodes. The outgroup was assigned using a primitive orchid species *Apostasis odorata*.

The evolutionary history of *matK* gene sequences estimated that the divergence time of *matK* gene in *V. spathulata* and related species from *A. odorata* was estimated to be about 114 million years ago. In other Orchids such as *C. nobilior* and *L. luddemanni,* this gene was diverged from 22 million years ago from *Vanda-related* species. This gene in *Vanda* spp. and closely related species diverged from about 0.1 to 5.12 million years ago and in *V. spathulata* about 0.08 million years ago (Fig. 2).

Statistical analysis was performed to analyze the geneflow and genetic differentiation among all sequences grouped

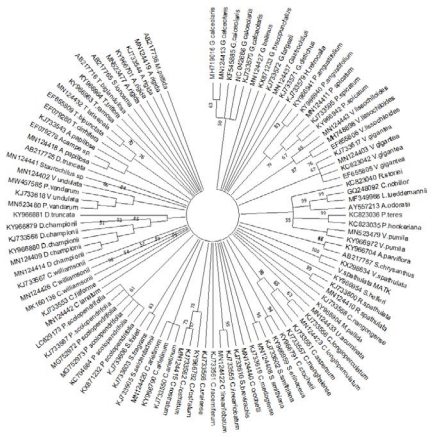

Fig. 1: phylogenetic tree constructed using minimum evolution method with 1000 bootstraps. the evolutionary distances computed with the maximum composite likelihood method. a total of 895 positions from 103 sequences were used. missing and ambiguous data were removed

separately into three clusters. Group 1 consisted of *V. spathulata* and related species with 18 sequences, group 2 consisted of all other distantly related species with 85 sequences and all the species and/or taxa with 103 sequences were considered group 3.

The genetic diversity test revealed the occurrence of slightly higher haplotype diversity and segregating sites in *Vanda spp* and its related species and other species, which was further revealed by related statistical tests as shown in Table 1. However, the overall analysis revealed the maximum diversity within the analyzed samples proceeding from different taxa of *Vanda*'s closely and distantly related species. Overall effective migration (Nm) test provided the information of Nm with 4.36, which revealed the geneflow with higher possibility at an interspecies level among *Vanda s*pp and other included taxa.

The Tajima neutrality test comparison between *vanda* and other taxa is summarized in Table 2 and represented in Fig. 3. The number of mutations and nucleotide differences was higher in *Vanda* and related taxa than in the other distantly related species. The nucleotide diversity (Pi) was also observed to be higher (0.02) than the distantly related taxa group and in the overall sequence group (0.008); indicating the Vanda spp and other related taxa had a greater number of evolutionary nucleotide diversity. The Tajima Neutrality test with a negative value (-2.09) in *Vanda* spp group indicates the existence of negative selection or purifying selection. In particular, taxa related to *Vanda* spp come under strong purifying selection, which reveals the removal of deleterious variations which leads to stabilizing selection. However, it reduces the genetic diversity.

The equality of evolutionary rate between *V. spathulata* and *Laelia lueddemannii* was performed and *A. odorata* was assigned as an outgroup species to infer Tajima relative rate. The observed divergent sites and unique differences were shown in Table 3 and apart from outgroup species, *V. spathulata* exhibited twice the amount of unique differences compared to *L. lueddemannii.* The $\chi 2$ test statistic was 4.50 ($p = 0.03389$), was used to reject the null hypothesis and accept an alternative hypothesis regarding the unique differences with equal rates between lineages. The analyzed results for the polymorphic sites of the *Vanda* species and other related species were given in Table 4. In the *matK* gene at the interspecies level, there are numerous polymorphic sites (123) in *Vanda spp* and

Table 1: genetic diversity, genetic differentiation and genetic flow of Vanda and related taxa

Genetic diversity							
Group	*Sequences*	*Segregating sites*	*Haplotypes*	*Hd*	*K*	*Pi*	*PiJC*
Group 1_V	55	121	33	0.97	11.79	0.019	0.02
Group 2_O	48	44	29	0.95	4.57	0.0076	0.007
Group 3_A	103	137	59	0.98	8.97	0.01	0.01
Genetic differentiation							
Test	χ2	Hst	Kst	Kst*	Z	Z*	Snn
Value	95	0.027	0.025	0.022	10147.15	8.88	0.23
PM test p-Value	Ns	*	***	***	***	***	Ns
Gene flow							
	Haplotype data		Sequence data				
Test	Gst		DeltaSt	GammaSt	Nst	Fst	
Gr_all	0.00392		0.00051	0.03447	0.05315	0.05420	
Nm	63.56		7	7	4.45	4.36	

Hd: Haplotype diversity, K: Average Nucleotide Differences, PiJc: Nucleotide differences with Jukes-Cantor corrections, $\chi 2$ - Chi Square Test, Hst: Haplotype-based statistics, Kst: Sequence-based statistics, Z*: Rank Statistics, Snn: Near Neighbor statistics, Gst: Differentiation of population, DeltaSt: Delta statistics, GammaSt: Gamma statistics, Nst: N statistics, Fst: Fixation Index, Nm: Effective number of migrants, PM: Permutation test with 1000 replicates. ns. not significant; # P<0.10; * P<0.05; ** P<0.01; *** P<0.001.

Table 2: Tajima neutrality test report of *vanda* and related taxa

Group	*Eta*	*K*	*Pi*	*Theta/ sequence*	*Theta site*	*Tajima D*	*Signifi-cance*
Vanda spp	140	12.50	0.02	30.59	0.05	-2.09	*
Other	57	5.51	0.008	12.84	0.01	-1.99	*
Overall	163	8.97	0.01	31.30	0.05	-2.37	**

Eta-Total number of mutations, K-Average number of nucleotide differences, Pi-Nucleotide diversity, Theta- Nucleotide differences (Watterson estimator); n.d., not determined; # p<0.10; * p<0.05; ** p<0.01; *** p<0.001.

Table 3: Tajima relative rate test for 3 sequences including *V. spathulata*, *L. lueddemanni*iand *A. odorata*

Configuration	*Count*
Identical sites in all three sequences	565
Divergent sites in all three sequences	5
Unique differences in Sequence A	22
Unique differences in Sequence B	10
Unique differences in Sequence C	60

related species groups and in overall species taxa (137) as well, however, the distantly related cluster belonging to other species resulting in lesser (53) polymorphic sites. This indicates species belonging to *Vanda* and other related species are having higher polymorphism at the nucleotide level. Parsimony informative sites (PIS) are used for distinguishing the species/taxa level and their evolutionary importance, also revealed a similar trend as the number of PIS occurrences was higher in *Vanda* spp and related species and overall species rather than other distantly related species/taxa. Despite anything to the contrary, genetic diversity is decreasing through the course of evolution, and thus purifying selection has occurred to remove the deleterious mutations in *Vanda* and closely and distantly related species.

A pairwise comparison performed between *V. spathulata* and *A. odorata* was given in Fig. 4. The aligned region of *matK* gene between *V. spathulata* and *A. odorata* exhibited an identity of 84.56%, which revealed the course of evolution and divergence of this gene from these two species. The conserved nucleotide regions were marked with an asterisk in the alignment.

The *matK* gene sequence of the *V. spathulata* and *A. odorata* was constructed into their secondary structures. The observed lowest energy ΔG (-55.04 kcal/mol at 37°C) for *V. spathulata* indicates the possibility of having more complex structural predictions associated with palindromes and another kind of inverted repeats on the primary structure of the gene. However, the medium lowest energy ΔG (-56.73 kcal/mol at 37°C) indicates the presence of complexity at the optimum level (Fig. 5). Similar analysis on *A. odorata* revealed the same, although the secondary structure at the lowest possible energy (-72.02 kcal/mol at 37°C) showed that there is no complex hairpin loop and other protrusions as that of *V. spathulata* (Fig. 6). Relative free energy between *V. spathulata* and *A. odorata* exhibited a higher level of difference and maximum free energy was required for secondary structural complexity formation in *A. odorata* (~-74 kcal/mol at 37°C) rather than *V. spathulata* (~-54 kcal/mol at 37°C). This prediction revealed that during the course of evolution, *V. spathulata* acquired complex structural modifications on its *matK* gene through the presence of extended repeats, inverted repeats, and another kind of secondary structure-related possibilities.

>matK_*V. spathulata*

GTTCCTTCTTTGCATTTATTGCGATTGATTTTC-
CACGAATATCATAATTTGAATAGTCT-
CATTACTTCAAAAAAAATCCATTTAC-
GTCTTTTCAAAAAAAAAGAAAAGATTCTTTTG-
GTTCCTACATAATTTTTATGTATATGAATGCGAATATC-
TATTCCTCTTTCTTCGTAAACAGTCTTCTTATTTAC-
GATCAATATCTTCTGGAGTCTTTCTTGAGCGAA-
CACATTTTTATGGAAAAATAGAATATCTTAGAGT-
CATGTCTTGTAATTCTTTTCAGAGGATCCTATG-
GTTTCTCAAAGATATTTTCATACATTATGTTCGATAT-
CAAGGAAAAGTGATTTTGGCTTCAAAAGGAACTCT-
TATTCTGATGCATAAATGGAAATTTCATTTT-
GTTAATTTTTGGCAATCTTATTTTCACTTTTG-
GTTTCAACCTTATAGGATCCATATAAAGCAATTA
CCCAATATTCCTTCTCTTTTCTGGGATATTTTTCA
AGTGTACTAAAAAATCCTTTGGTAGTAAGAAAT-
CAAATGCTAGAGAATTCATTTATAATAAAGACTCT-
GACTAAAGAATTTGATACCATAGCTCCAGT-
TATTTTTCTTATTGGATCATTGTCAAAAGCT-
CAATTTTGTACTGTATTAGGTCATCCTATTAGTAAAC-
CGATCTGGACAAATTTATCGGATTCTGATATTCTT-
GATCGATTTTGTCGGATATGTAGAAATCTTTGTCGT-
TATCACAGTGGATCCTCAAAGAAACAGGTTTTG-
TATCGTATAAAGTATATA

>AY557213.1 *Apostasia odorata* maturase K (*matK*) gene, partial cds; chloroplast

ACTTCCTATATCCGCTACTCCTGAAAGAGTATATT-
TATTCACTTGTTCATGATCATAGTCTTCTTTTT-

Table 4: Polymorphic sites in *V. spathulata* species, other related species group and overall species group

Group	Sequences/sites analyzed	PS	MS	SV	PIS	SV2V	PIS2V	SV3V	PIS3V	PIS4V
Vanda spp	55/597	123	474	64	59	64	43	0	15	1
Other	48/689	53	636	32	21	30	19	2	2	0
Overall	103/595	137	458	71	66	69	44	2	20	2

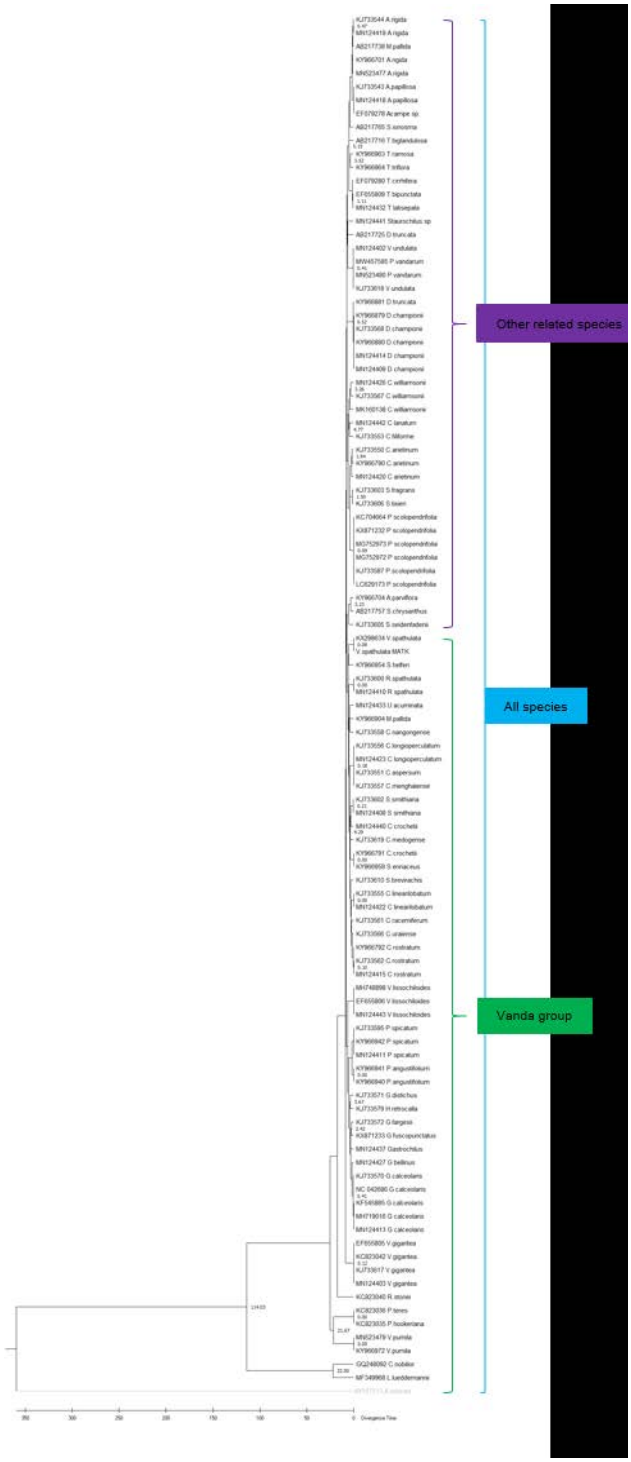

Fig. 2: Reltime divergence time tree

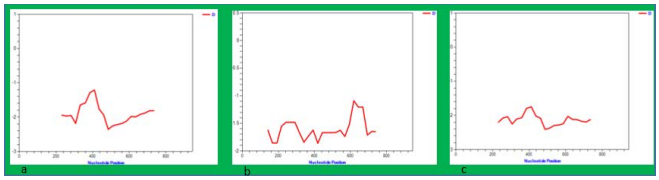

Fig. 3: a) Tajima d test for vanda spp, b) tajima d test for other spp, c) tajima d test for overall spp.

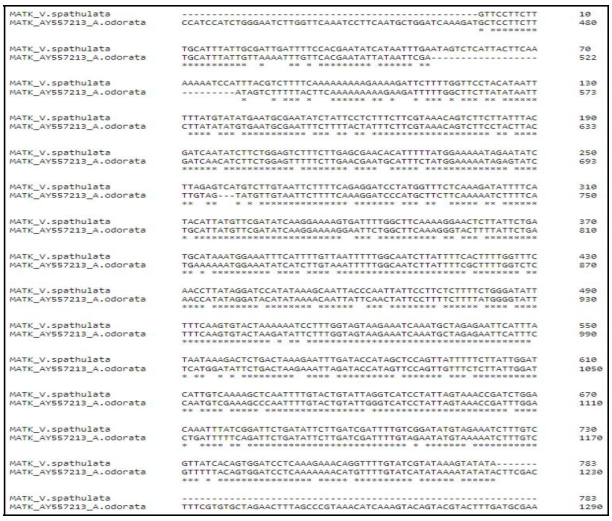

Fig. 4: Pairwise alignment of matk gene between *V. spathulata* and A. odorata

TACGAACCTGTCGAAATTTGTGGTTATGACAATA-
AATCTAGTTTAGTACTTGTGAAACGTTTAATTACTC-
GAATGTATCAACAGAATTCTTTGATTTCTTCGGT-
TAATGATTCTAACCAAAATGGATTTTGTGGTCA-
CAAGAATTCTTTTTCTTCTAATTTTGCTTTTCAAATG-
GCATCAGAAGGTTTTGGAGTCATTCTGGAAATTC-
CATT

CTCGTCGCGATTAGTTTCTTCCCCTGAAGAAATA-
AAAATACCAAAATCTCTGAATTTACGATC-
TATTCATTCAATATTTCCTTTTTTAGAGGA-
CAAATTCTTACATTTAAATTATGTATTAGGTATA-
CAAATACCCCATCCCATCCATCTGGGAATCTTG-
GTTCAAATCCTTCAATGCTGGATCAAAGAT-
GCTCCTTCTTTGCATTTATTGTTAAAATTT-
GTTCACGAATATTATAATTCGAATAGTCTTTT-
TACTTCAAAAAAAAAGAAGATTTTTGGCTTCT-
TATATAATTCTTATATATGTGAATGCGAATTTCTTT-
TACTATTTCTTCGTAAACAGTCTTCCTACTTAC-
GATCAACATCTTCTGGAGTTTTTCTTGAAC-
GAATGCATTTCTATGGAAAAATAGAGTATCTTG-
TAGTATGTTGTAATTCTTTTCAAAGGATCCCATG-
CTTCTTCAAAAATCTTTTCATGCATTATGTTCGATAT-
CAAGGAAAAGGAATTCTGGCTTCAAAGGGTACTTT-
TATTCTGATGAAAAAATGGAAATATCATCTTGTA-
AATTTTTGGCAATCTTATTTTCGCTTTTGGTCT-
CAACCATATAGGATACATATAAAACAATTATTCAAC-
TATTCCTTTTCTTTTATGGGGTATTTTTCAAGTGTAC-
TAAGATATTCTTTGGTAGTAAGAAATCAAATGC-
TAGAGAATTCATTTCTCATGGATATTCTGACTA-
AGAAATTAGATACCATAGTTCCAGTTGTTTCTCT-
TATTGGATCAATGTCGAAAGCCCAATTTTGTACT-
GTATTGGGTCATCCTATTAGTAAACCGATTTG-

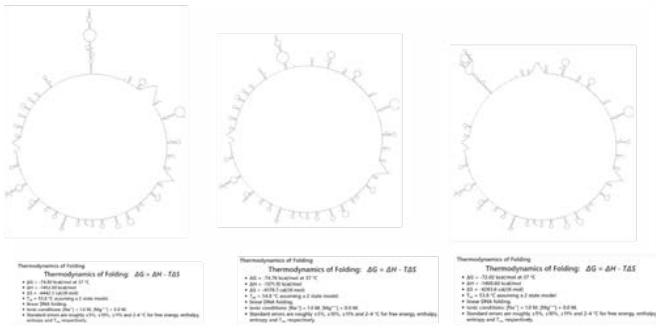

Fig. 5: Secondary structure prediction of single-stranded matk gene of *A. odorata*

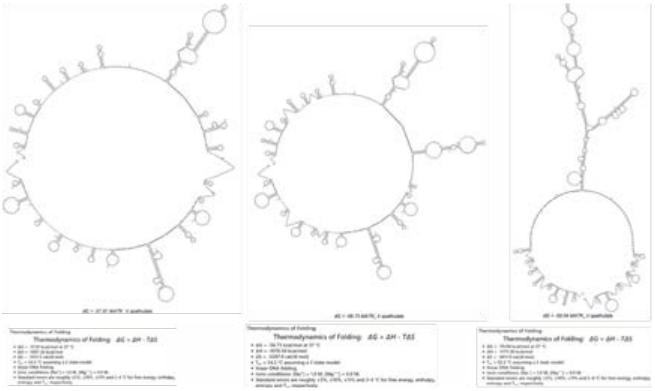

Fig. 6: secondary structure prediction of single-stranded matK gene of *V. spathulata*

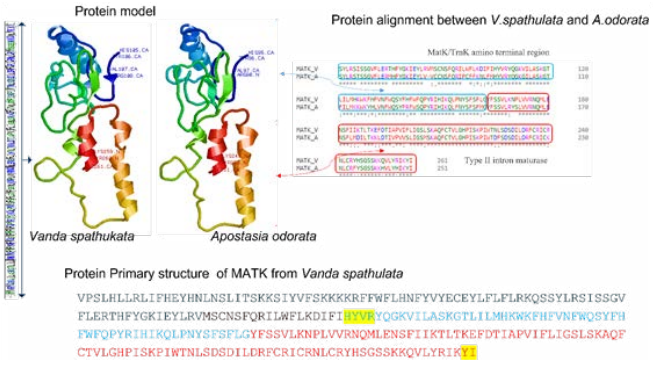

Fig. 7: Protein modeling comparison between Vanda spathulata and *A. odorata*

GACTGATTTTTCAGATTCTGATATTCTTGATC-
GATTTTGTAGAATATGTAAAAATCTTTGTCGTTTT-
TACAGTGGATCCTCAAAAAAACATGTTTTG-
TATCATATAAAATATATACTTCGACTTTCGT-
GTGCTAGAACTTTAGCCCGTAAACATCAAAG-
TACAGTACGTACTTTGATGCGAAGATTAG-
GTTCGGGATTCTTAGAAAAATTCTTTATGGAA-
GAAGAACAAGTTCTTTCTTTAATCTTCCTC-
CAAAAAATCTCTTTTTCTTTACATGGATTA-
CATAGAAAACGTATTTGGTATTTGGACATTATCTG-
TATTAATGATCTGGTGGATTCTTCATGA

FLYPLLLKEYIYSLVHDHSLLFYEPVEICGYDNKSSLV-
LVKRLI

TRMYQQNSLISSVNDSNQNGFCGHKNSFSSNFAFQMASEG-
FGVILEIPFSSRLVSSPE

EIKIPKSLNLRSIHSIFPFLEDKFLHLNYVLGIQIPHPI-
HLGILVQILQCWIKDAPSL

HLLLKFVHEYYNSNSLFTSKKKKIFGFLYNSYICECE-
FLLLFLRKQSSYLRSTSSGVF

LERMHFYGKIEYLVVCCNSFQRIPCFFKNLFMHYVRYQGK-
GILASKGTFILMKKWKYH

LVNFWQSYFRFWSQPYRIHIKQLFNYSFSFMGYFSSVL-
RYSLVVRNQMLENSFLMDIL

TKKLDTIVPVVSLIGSMSKAQFCTVLGHPISKPIWTDFS-
DSDILDRFCRICKNLCRFY

SGSSKKHVLYHIKYILRLSCARTLARKHQST-
VRTLMRRLGSGFLEKFFMEEEQVLSLI

FLQKISFSLHGLHRKRIWYLDIICINDLVDSS

The correct frame of the protein sequence was translated upon comparing the ORF with NCBI and BLASTN.
5'3' Frame 1
VPSLHLLRLIFHEYHNLNSLITSKKSIYVFSKKKKRFFW-
FLHNFYVYECEYLFLFLRKQSSYLRSISSGVFLERTHFYG-
KIEYLRVMSCNSFQRILWFLKDIFIHYVRYQGKVILAS-
KGTLILMHKWKFHFVNFWQSYFHFWFQPYRIHIKQLP-
NYSFSFLGYFSSVLKNPLVVRNQMLENSFIIKTLTKEFD-
TIAPVIFLIGSLSKAQFCTVLGHPISKPIWTNLSDSDIL-
DRFCRICRNLCRYHSGSSKKQVLYRIKYI

Translation of putative protein corresponding to its coding nucleotide region

gtt	cct	tct	ttg	cat	tta	ttg	cga	ttg	att
ttc	cac	gaa	tat	cat	aat	ttg	aat	agt	ctc
V	P	S	L	H	L	L	R	L	I
F	H	E	Y	H	N	L	N	S	L
att	act	tca	aaa	aaa	tcc	att	tac	gtc	ttt
tca	aaa	aaa	aag	aaa	aga	ttc	ttt	tgg	ttc
I	T	S	K	K	S	I	Y	V	F
S	K	K	K	K	R	F	F	W	F
cta	cat	aat	ttt	tat	gta	tat	gaa	tgc	gaa
tat	cta	ttc	ctc	ttt	ctt	cgt	aaa	cag	tct
L	H	N	F	Y	V	Y	E	C	E
Y	L	F	L	F	L	R	K	Q	S
tct	tat	tta	cga	tca	ata	tct	tct	gga	gtc
ttt	ctt	gag	cga	aca	cat	ttt	tat	gga	aaa
S	Y	L	R	S	I	S	S	G	V
F	L	E	R	T	H	F	Y	G	K
ata	gaa	tat	ctt	aga	gtc	atg	tct	tgt	aat
tct	ttt	cag	agg	atc	cta	tgg	ttt	ctc	aaa
I	E	Y	L	R	V	M	S	C	N
S	F	Q	R	I	L	W	F	L	K

Table 5: Comparison of physiochemical properties of *matK* protein of *V. spathulata* and *A. odorata*

Physico-Chemical Properties of *matK* protein	*V. spathulata*	*A. odorata*
Number of amino acids	261	251
Alpha helix	66%	54.98%
Extended strand	22.29%	16.73%
Random coil	35.67%	28.29%
Theoretical Isoelectric focusing point	9.95	9.70
Negatively charged residues (Aspartic acid + Glutamic acid)	12	12
Positively charged residues (Arginine+Lysine)	17	34
Instability Index	43.13	41.64
Aliphatic Index	97.39	88.84

```
gat  att  ttc  ata  cat  tat  gtt  cga  tat  caa
gga  aaa  gtg  att  ttg  gct  tca  aaa  gga  act
  D    I    F    I    H    Y    V    R    Y    Q
G    K    V    I    L    A    S    K    G    T
ctt  att  ctg  atg  cat  aaa  tgg  aaa  ttt  cat
ttt  gtt  aat  ttt  tgg  caa  tct  tat  ttt  cac
  L    I    L    M    H    K    W    K    F    H
F    V    N    F    W    Q    S    Y    F    H
ttt  tgg  ttt  caa  cct  tat  agg  atc  cat  ata
aag  caa  tta  ccc  aat  tat  tcc  ttc  tct  ttt
  F    W    F    Q    P    Y    R    I    H    I
K    Q    L    P    N    Y    S    F    S    F
ctg  gga  tat  ttt  tca  agt  gta  cta  aaa  aat
cct  ttg  gta  gta  aga  aat  caa  atg  cta  gag
  L    G    Y    F    S    S    V    L    K    N
P    L    V    V    R    N    Q    M    L    E
aat  tca  ttt  ata  ata  aag  act  ctg  act  aaa
gaa  ttt  gat  acc  ata  gct  cca  gtt  att  ttt
  N    S    F    I    I    K    T    L    T    K
E    F    D    T    I    A    P    V    I    F
ctt  att  gga  tca  ttg  tca  aaa  gct  caa  ttt
tgt  act  gta  tta  ggt  cat  cct  att  agt  aaa
  L    I    G    S    L    S    K    A    Q    F
C    T    V    L    G    H    P    I    S    K
ccg  atc  tgg  aca  aat  tta  tcg  gat  tct  gat
att  ctt  gat  cga  ttt  tgt  cgg  ata  tgt  aga
  P    I    W    T    N    L    S    D    S    D
I    L    D    R    F    C    R    I    C    R
aat  ctt  tgt  cgt  tat  cac  agt  gga  tcc  tca
aag  aaa  cag  gtt  ttg  tat  cgt  ata  aag  tat
  N    L    C    R    Y    H    S    G    S    S
K    K    Q    V    L    Y    R    I    K    Y
a                        t                        a
 I
```

A protein model was created using the alignment of existing protein group II intron-encoded protein LtrA: A (PDB 7d1a.1 using Swiss modeler)[20] and given in Fig. 7. There was an identity of 19.35% between the PDB file sequence and *matK* from *V. spathulata.* Further Pfam database search revealed the presence of *matK* /TrnK amino-terminal region and Type II intron maturase region on 1-58 and 85 to 157, respectively among the aligned region observed with PDB of Swiss modeler. Pairwise alignment of the selected region of *matK* proteins from these two species revealed the 77.29 Percentage of identity. However, insertion of 9 amino acids (KKSIYVFSK) in *V. spathulata,* and some conservative and semi-conservative amino acid changes were also observed. In a primary structure spanning the aligned regions of 261 amino acids between *V. spathulata* and *A. odorata* there were 35 conservative changes in amino acids such as valine to isoleucine, lysine to arginine, etc, there are 6 semi-conservative changes (eg. Lysine to Serine) and 8 non-conservative changes (eg. Phenylalanine to Serine) and others were identical amino acids. The instability between these two species of *matK* protein in the aligned region showed that the protein is slightly unstable as this aligned region of the protein exceeds the instability index limit of 40.

>*matK*_V

VPSLHLLRLIFHEYHNLNSLITSKKSIYVFSKKKKRFFW-
FLHNFYVYECEYLFLFLRKQSSYLRSISSGVFLERTHFY-
GKIEYLRVMSCNSFQRILWFLKDIFIHYVRYQGKVILASK-
GTLILMHKWKFHFVNFWQSYFHFWFQPYRIHIKQLPNYS-
FSFLGYFSSVLKNPLVVRNQMLENSFIIKTLTKEFDTIAPVI-
FLIGSLSKAQFCTVLGHPISKPIWTNLSDSDILDRFCRICRNL-
CRYHSGSSKKQVLYRIKYI

```
>AAT45854.1 maturase K, partial (chloro-
plast) [Apostasia odorata]

FLYPLLLKEYIYSLVHDHSLLFYEPVEICGYDNKSSLV-
LVKRLITRMYQQNSLISSVNDSNQNGFCGHKN

SFSSNFAFQMASEGFGVILEIPFSSRLVSSPEEIKIPK-
SLNLRSIHSIFPFLEDKFLHLNYVLGIQIPHP

IHLGILVQILQCWIKDAPSLHLLLKFVHEYYNSNSLFTSK-
KKKIFGFLYNSYICECEFLLLFLRKQSSYL

RSTSSGVFLERMHFYGKIEYLVVCCNSFQRIPCFFKNLFM-
HYVRYQGKGILASKGTFILMKKWKYHLVNF

WQSYFRFWSQPYRIHIKQLFNYSFSFMGYFSSVLRYSLV-
VRNQMLENSFLMDILTKKLDTIVPVVSLIGS

MSKAQFCTVLGHPISKPIWTDFSDSDILDRFCRICKNL-
CRFYSGSSKKHVLYHIKYILRLSCARTLARKH

QSTVRTLMRRLGSGFLEKFFMEEEQVLSLIFLQKISFSLH-
GLHRKRIWYLDIICINDLVDSS
```

The physicochemical properties of the matK protein of the *V. spathulata* and *A. odorata* were analyzed, compared, and tabulated in Table 5. In all of the parameters tested, *V. spathulata* protein values were higher than that of the *Apostasia odorata* protein. The amino acid residues were 261 and 251 with *Vanda* and *Apostasia*. The molecular

weight for *Vanda* was 31599.26 Daltons. The theoretical isoelectric (PI) focusing point for both species was more than 9.00. The total negative charge was similar in both species while the total positive charge of the *Vanda* protein was less than the Apostasia protein. The instability and aliphatic indexes were higher in *Vanda* than Apostasia. The physiochemical parameters of the *V. spathulata* protein in detail were given in Table 6.

Table 6: Physiochemical properties of matK protein of V. spathulata in detail

Number of amino acids: 261

Molecular weight: 31599.26

Theoretical pI: 9.95

Amino acid composition:

Ala (A)	3	1.1%
Arg (R)	15	5.7%
Asn (N)	11	4.2%
Asp (D)	5	1.9%
Cys (C)	6	2.3%
Gln (Q)	9	3.4%
Glu (E)	7	2.7%
Gly (G)	8	3.1%
His (H)	12	4.6%
Ile (I)	23	8.8%
Leu (L)	31	11.9%
Lys (K)	22	8.4%
Met (M)	3	1.1%
Phe (F)	27	10.3%
Pro (P)	7	2.7%
Ser (S)	27	10.3%
Thr (T)	8	3.1%
Trp (W)	6	2.3%
Tyr (Y)	17	6.5%
Val (V)	14	5.4%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 12

Total number of positively charged residues (Arg + Lys): 37

Atomic composition:

Carbon	C	1500
Hydrogen	H	2254
Nitrogen	N	378
Oxygen	O	358
Sulfur	S	9

Formula: $C_{1500}H_{2254}N_{378}O_{358}S_9$

Total number of atoms: 4499

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 58705

Abs 0.1% (=1 g/l) 1.858, assuming all pairs of Cys residues form cystines

Ext. coefficient 58330

Abs 0.1% (=1 g/l) 1.846, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is V (Val).

The estimated half-life is: 100 hours (mammalian reticulocytes, in vitro).

>20 hours (yeast, in vivo).

>10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to

be 43.13

This classifies the protein as unstable.

Aliphatic index: 97.39

Grand average of hydropathicity (GRAVY): 0.034

Discussion

Genetic diversity gives plants the essential traits to strengthen their ability to resist disease, pests and against biotic and abiotic stress. *matK* is a chloroplast gene encoding maturase K (*matK*) protein; the most widely and commonly used genetic marker for the phylogenetic analysis of plants. In previous studies, the *matK* gene sequence had been used to study the phylogenetic relationship among 36 species of the *Oncidium* orchid genus[21] and in the genetic diversity assessment of four orchid species.[3] The fragment of DNA used for this purpose should be highly recoverable, have a high proportion in species identification, and be affordable.[22] *matK* gene fragments have those properties and hence are widely used in phylogenetic analysis.

In this study, the genetic diversity of the *matK* gene of *V. spathulata* was analyzed among the other related plant taxa of the orchid family. The matK gene was easily amplified in our study by PCR. The *matK* gene of *V. spathulata* had a similarity with 103 gene sequences from different taxa of the orchid family revealed in the blast sequence search. Utilizing the minimum evolution method a high-confidence phylogenetic tree was constructed, and the divergence time tree was estimated for all those 103 sequences. The diversity in the *matK* gene of the *V. spathulata* among the closely related and distantly related taxa was determined. The gene of interest in the *vanda* species and other related species is estimated to diverge from *A. odorata* species over 114 million years ago. *A. odorata* is one of the eight primitive species of *Apostasia* genus from the orchid family; native to Eastern India, Southern China, and Malaysia (The International Plant Names Index and World Checklist of Vascular Plants 2023). The reltime divergence tree grouped the sequences from the *Vanda* species and the related taxa into *Vanda* group, *S. helferi, R. spathulata, U. acuminta, M. palida, C. nangogense, C. longioperculatum, C.aspersum, C. menghaiense, S. smithiana, C. crochetti, S. erinaceus, S. brevirachis, C. linearilobactum, C. racemiferum, C. uraiense, C. rostratum, V. lissochilloides, P. spicatum, P. angustifolium, G. distichus, H. retrocalla, G. fargesii, G. fuscopunctatus, Gastrochilus, G. calceolaris, V. gigantea, R. storiei, P. teres, P. hookeriana, V. pumila, C. nobilior, L. luddemannii*; the remaining species were grouped as other related species. Haplotypes in the chromosomes, the set of DNA variations or genetic markers, are closely linked and inherited together. They are the alleles for different types of polymorphisms, but mainly single nucleotide polymorphisms (Gard, 2021). The species-level genetic diversity monitored based on the haplotype diversity revealed a higher proportion of genetic variation and a higher possibility of gene flow within the *Vanda* and other closely related taxa groups, which was most similar to the estimated diversity observed among all the taxa. In a similar study of the South Korean terrestrial orchid plant, *Habenaria linearifolia,* the genetic diversity was comparable to other orchids observed in this study; however, the gene flow was limited.[24]

Kreitman discovered that in a natural population, evolutionary changes can be determined by the genetic variation among species while Ronald Fisher theorized natural selection as the important tool in evolutionary changes and not genetic drift. Later it was confirmed by Kimura that genetic drift is the reason for evolution changes among populations, in his Neutral theory of evolution[25], and the statistic used to test this theory is the Tajima neutrality test or D test. In the present investigation, the Tajima D test value of *V. spathulata* and related taxa was significant and negative. This is on par with the results of the Tajima test of the *matK* gene sequences of four orchid genera, *Geodorum, Dendrobium, Cymbidium,* and *Rhynchostylis*.[3] This negative D value associated with the statistical significance is an indication of the existence of low-frequency polymorphisms in excess rate among those analyzed population taxa relative to expectation and expansion of population expansion. It can be interpreted that the sequences contain rare alleles in abundance and there was a (i) selective sweep recently, a process that increased the frequency of a new beneficial mutation that was fixed in the population, (ii) a recent bottleneck (reduction in population size) followed by expansion of population and (iii) swept gene linkage (presence of the mutated gene after the sweep). Among other closely related taxa, the polymorphism rate is less. Also, the study revealed higher polymorphic sites among the *Vanda* and closely related taxa group than among the distantly related taxa group. A similar negative Tajima D result was observed among the overall taxa. In the relative rate of evolution analysis between species, *V. spathulata* and *L. lueddemannii,* the former showed greater unique differences in the observed sites than the latter.

Several conserved sites were observed in a pairwise sequence alignment of the *matK* gene sequence of *V. spathulata* with the primitive species *A. odorarta* with about 84.56% similarity between them. Higher sequence similarity may be the reason for identical homology, functionality, and the possibility of the high conservation of the matK gene.[26] Although the two sequences exhibited high similarity in alignment, the structure they gave rise to were in contrast. The secondary structure prediction of the *matK* gene sequences revealed a more complex *vanda* structure than its ancestor. Even though *A. odarata* was a