Indexed, Abstracted and Cited: ISRA Journal Impact Factor, International Impact Factor Services (IIFS), Directory of Research Journals Indexing (DRJI), International Institute of Organized Research and Scientific Indexing Services, Cosmos Science Foundation (South-East Asia), Einstein Institute for Scientific Information {EISI}, Directory of Open Access Scholarly Resources,

citefactor.org journals indexing

Directory Indexing of International Research Journals

World Journal of Biology and Medical Sciences

Published by Society for Advancement of Science®

ISSN 2349-0063 (Online/Electronic)

Volume 2, Issue- 4, 1-9, October to December, 2015

WJBMS 2/04/42/2015

www.sasjournals.com

A Double Blind Peer Reviewed Journal / Referred Journal

wjbmedsc@gmail.com/wjbms.lko@gmail.com

RESEARCH PAPER

Received: 14/07/2015 Revised: 01/09/2015 Accepted: 02/09/2015

Phylogenetic Studies on Maturase K gene of *Vanda* species

Molumenla Walling, *Salam Pradeep Singh, Chitta Ranjan Deb, **Lakshmi Narayan Kakati and ***Bolin Kumar Konwar

Department of Botany, Nagaland University, Lumami-798627, Nagaland, India

***Bioinformatics Infrastructure Facility, Nagaland University, Lumami-798627, Nagaland, India**

****Department of Zoology, Nagaland University, Lumami-798627, Nagaland, India**

*****Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur-784028, Assam, India**

ABSTRACT

In this article, the Maturase K gene sequence of 13 Vanda species was retrieved from the NCBI GenBank database and performed the sequence analysis and alignment studies. Further, from the multiple sequence alignment of the Maturase K gene a phylogenetic tree was generated revealing Vanda bensonii and Vanda lioville are closely related to each other. Further, to validate the generated phylogenetic tree, we have performed a statistical clustering based on the A, T, G, C, A+T and G+C composition of the Vanda species. The result hold even true when it comes to clustering using K-means clustering based on Euclidean Distance Squared where the 2D and 3D plot indicates Vanda bensonii and Vanda lioville are closely related species.

Keywords: Maturase K, Vanda, Phylogenetic Tree, K-means and Clustering.

INTRODUCTION

Vanda is a genus in the orchid which belongs to the family Orchidaceae. Which, although not large (about fifty species), is one of the most important florally in the world. This genus and its allies are considered to be one of the most highly evolved of all orchids within Orchidaceae. The genus is very valuable and highly prized in horticulture for its showy, fragrant, long lasting, and intensely colorful flowers [**The Orchids**]. *Vanda* is widespread across the world especially in East Asia, Southeast Asia, and New Guinea, with a few species extending into Queensland and some of the islands of the western Pacific [**Flora of China**]. The genus is sometimes abbreviated as V. in the floral trade.

The name "*Vanda*" is derived from the Sanskrit [**Garay, 1972 and World Checklist of Selected Plant Families**] name for the species *Vanda roxburghii* [**Grove, 1995**]. These mostly epiphytic, but sometimes lithophytic or terrestrial orchids are distributed in India, Himalaya, SE Asia, Indonesia, the Philippines, New Guinea, southern China and northern Australia. The genus has a monopodial growth habit [**Motes and Hoffman, 1997**]. The leaves are highly variable according to their habitat environment. Leaves of this genus has a variety of shape, the leaves of some species are flat, typically broad, ovoid leaves (strap-leaves), while others have cylindrical (terete) leaves, fleshy leaves and they are adapted to dry periods [**Garay, 1972**]. The stems of these orchids vary considerably in size depending on different species [**Hogeweg, 2011**]. There are miniature plants and plants with a length of several meters. The plants can become quite massive depending on their habitat and in cultivation process. The epiphytic species possess very large and rambling aerial root systems. There is variety of many flattened flowers growing on a lateral inflorescence. Most of them show a yellow-brown color with brown markings, but some they also appear in white, green, orange, red and burgundy shades. The lip has a small spur. *Vandas* mostly bloom every few months and the flowers last for two to three weeks [**Garay, 1972**].

Among many *Vanda* orchids especially *Vanda coerulea* are endangered, and have never been common because they are usually only infrequently encountered in habitat and grow only in disturbed forest areas with high light levels. They are highly threatened and vulnerable to various habitat destruction [**Flora of China**]. The export of wild-collected specimens of the Blue Orchid that is *Vanda coerulea* and other wild *Vandas* is prohibited worldwide, as all orchids are listed on the Convention on International Trade in Endangered Species [**Hogeweg, 2011**].

In the present investigation, we have performed the sequence analysis of some Vanda species. In bioinformatics a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences [**Waterman, 1995**]. In sequence alignments of proteins, the degree of similarity between nucleotide / amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages [**Mount, 2002**]. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest [**Claverie and Notredame, 2003**] that this region has structural or functional

importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role [**Mount, 2004 and Clustal W2 FAQs**].

From the sequence alignment, we have generated the phylogetic tree of the Vanda species.

The multiple alignments of the Vanda species try to align all of the sequences in a given query set [**Ng and Henikoff, 2001 and Wang and Jiang, 1994**]. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. The multiple sequence alignment aid in establishing evolutionary relationships by constructing phylogenetic trees [**Elias Isaac 2006**]. Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences [**Durbin Richard et al., 1998 and Abu-Jamous et al., 2013**].

MATERIALS AND METHODS

The nucleotide sequence of the following *Vanda* species was retrieved from the NCBI Gene Bank Database (Shown in Table 1).

Table 1. *Vanda* species used in the present investigation.

SN	Accession No	Familiy	Genus	Species	Gene	Region
1	KF361631	Orchidaceae	*Vanda*	*alpina*	Maturase K	Chloroplast
2	KJ628992	Orchidaceae	*Vanda*	*bensonii*	Maturase K	Chloroplast
3	KJ628990	Orchidaceae	*Vanda*	*brunnea*	Maturase K	Chloroplast
4	KJ628989	Orchidaceae	*Vanda*	*coerulescens*	Maturase K	Chloroplast
5	KJ629016	Orchidaceae	*Vanda*	*denisoniana*	Maturase K	Chloroplast
6	FR832845	Orchidaceae	*Vanda*	*cristata*	Maturase K	Chloroplast
7	KJ628994	Orchidaceae	*Vanda*	*lilacina*	Maturase K	Chloroplast
8	KJ028773	Orchidaceae	*Vanda*	*flabellata*	Maturase K	Chloroplast
9	KJ628991	Orchidaceae	*Vanda*	*liovillei*	Maturase K	Chloroplast
10	KJ628996	Orchidaceae	*Vanda*	*pumila*	Maturase K	Chloroplast
11	JN004623	Orchidaceae	*Vanda*	*stangeana*	Maturase K	Chloroplast
12	KJ628993	Orchidaceae	*Vanda*	*tessellata*	Maturase K	Chloroplast
13	KJ628995	Orchidaceae	*Vanda*	*testacea*	Maturase K	Chloroplast

The nucleotide sequence of *Vanda alpine* [KF361631], *Vanda bensonii* [KJ628992], *Vanda brunnea* [KJ628990], *Vanda coerulescens* [KJ628989], *Vanda denisoniana* [KJ629016], *Vanda cristata* [FR832845],*Vanda lilacina* [KJ628994], *Vanda flabellata* [KJ028773], *Vanda liovillei* [KJ628991], *Vanda pumila* [KJ628996], *Vanda stangeana* [JN004623], *Vanda tessellata* [KJ628993] and *Vanda testacea* [KJ628995] was retrieved from the NCBI Gene Bank Database (http://www.ncbi.org/). The retrieved nucleotide sequences of these 13 Vanda species were aligned using CLC Main Workbench and clustered using neighbor Joining method and a tree was generated. The closely related were further analyzed for nucleotide composition analysis, such as the A+T content and G+C content and its phylogeny was studied.

The A, T, G, C, A+T and G+C content of the 13 *Vanda* species were further analyzed for statistical clustering using various

methods such as the K-means clustering. The K-means clustering is one of the oldest and most widely used clustering algorithms are the K-means algorithm. The K-means algorithm clusters a number of data records into K partitions based on the chosen numerical descriptors and the clustering measure. The K-means algorithm works as follows: First, assign each record to the closest centroid, using a specific cluster measure (introduced below). All the records assigned to a given centroid belong to the same cluster. Second assign *K* initial centroids, where *K* is a user-defined parameter specifying the number of clusters that should be created (each centroid corresponds to a cluster, where the centroid is the average of all the data points assigned to the cluster). The K centroids are either randomly picked records or randomly generated data points since no records have been assigned to the clusters yet. Third, update all centroids (the centroid position is set to the mean of all records assigned to the centroid). And lastly, the repetition of second and third steps until the centroids do not change (i.e. no records change cluster) or until a maximum number of iterations has occurred.

RESULTS AND DISCUSSION

Maturase K (matK) is a plant gene which encodes an introne maturase, a protein that splices introns. Amongst other maturase, this protein retains only a well conserved domain X and remnants of a reverse transcriptase domain. Universal matK primers can be used for DNA barcoding of angiosperms. In the present investigation the tree produced by neighbor joining method from the Nucleotide sequence alignment (Fig. 1) of the 13 Vanda species revealed that some of the *Vanda* species are closely related. From the phylogenetic study analysis on the matK gene of various *Vanda* species we have found out that *Vanda bensonii* and *Vanda liovillei* are very similar and closely related to each other in the matK gene which is present in the chloroplast. *Vanda pumila* and *Vanda denisoniana* are also closely related. Also, *Vanda stangeana* is also related to them. In case of *Vanda alpine* and *Vanda cristata* though they have different evolution but they are closely related to each other which are shown in Fig 2.

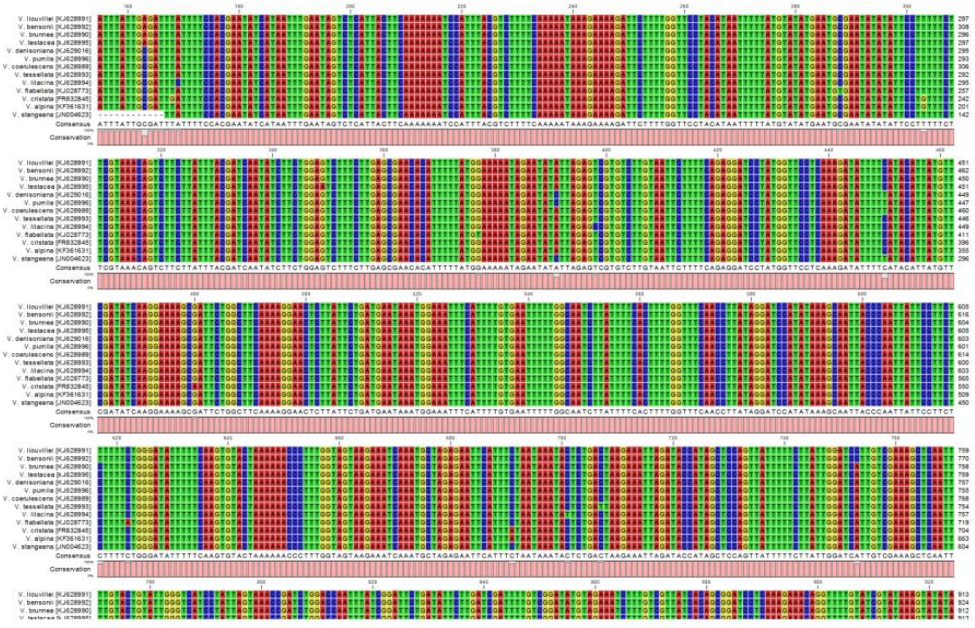

Fig. 1. Multiple sequence alignment of the 13 *Vanda* species.

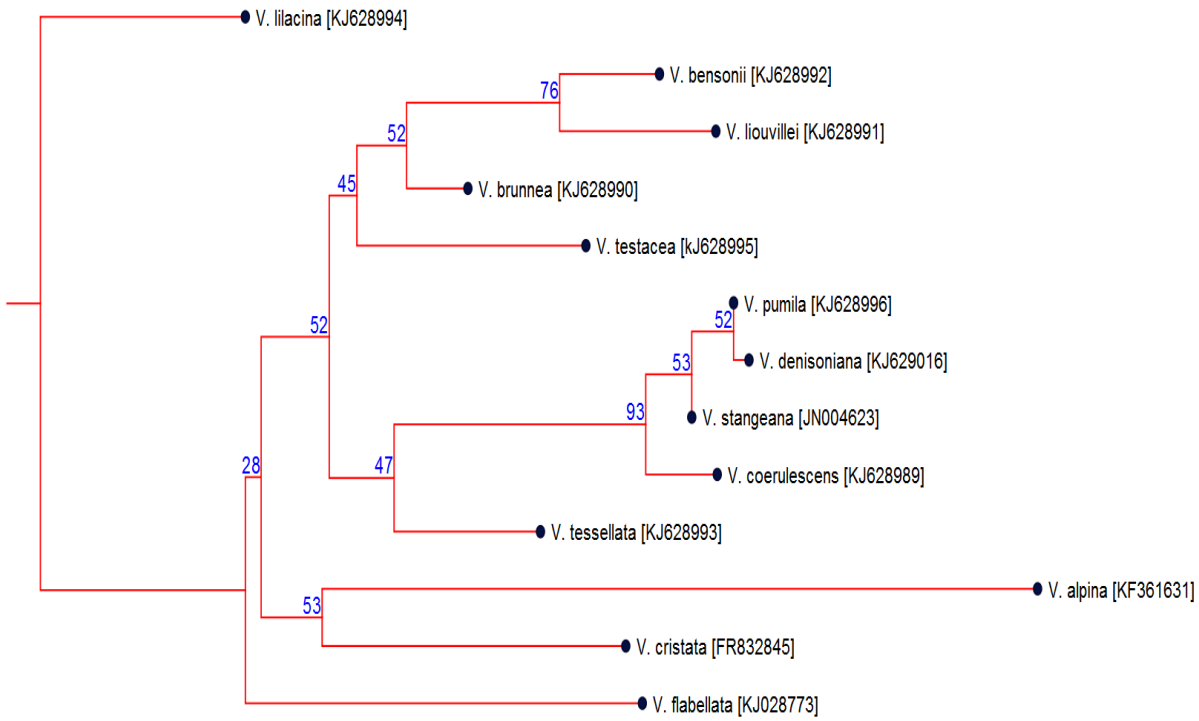

Fig. 2. Phylogenetic tree generated based on neighbor joining for the 13 *Vanda* species with a bootstrap of 100 replicates.

Further the comparative analysis on nucleotide sequence composition of A, T, G, C, A+T and G+C content is shown in Table 2 and the frequency of nucleotide distribution is shown in Table 3. Also the bar graph plot representing the variation in the nucleotide distribution of the matK gene of the various 13 species of Vanda is shown in Fig. 3.

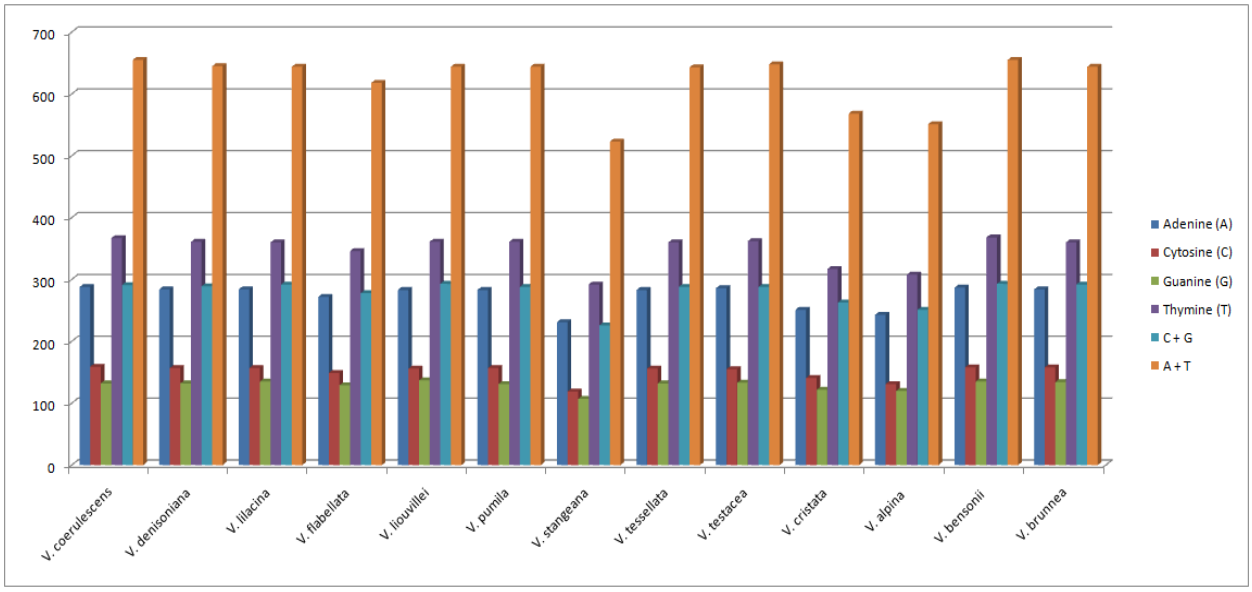

Fig. 3. Bar graph plot of the nucleotide distribution of the 13 *Vanda* species.

Table 2. Counts of nucleotide distribution of the 13 *Vanda* species.

SN	Accession	Name	Length	A	C	G	T	C+G	A+T
1	KJ628989	*V. coerulescens*	946	288	159	132	367	291	655
2	KJ6290176	*V. denisoniana*	934	284	157	132	361	289	645
3	KJ628991	*V. lilacina*	936	284	157	135	360	292	644
4	KJ028773	*V. flabellata*	896	272	149	129	346	278	618
5	KJ628991	*V. liovillei*	937	283	156	137	361	293	644
6	KJ628996	*V. pumila*	932	283	157	131	361	288	644
7	JN004623	*V. stangeana*	749	231	119	107	292	226	523
8	KJ628993	*V. tessellata*	931	283	156	132	360	288	643
9	KJ628993	*V. testacea*	936	286	155	133	362	288	648
10	FR832845	*V. cristata*	831	251	141	122	317	263	568
11	KF361631	*V. alpina*	802	243	131	120	308	251	551
12	KJ628992	*V. bensonii*	948	287	158	135	368	293	655
13	KJ628990	*V. brunnea*	936	284	158	134	360	292	644

Table 3. Frequency of nucleotide distribution of the 13 Vanda species.

SN	Accession No	Name	DNA Length	A	C	G	T	C+G	A+T
1	KJ628989	*V. coerulescens*	946	0.304	0.168	0.140	0.388	0.308	0.692
2	KJ6290176	*V. denisoniana*	934	0.304	0.168	0.141	0.387	0.309	0.691
3	KJ628991	*V. lilacina*	896	0.303	0.168	0.144	0.385	0.312	0.688
4	KJ028773	*V. flabellata*	896	0.304	0.166	0.144	0.386	0.31	0.69
5	KJ628991	*V. liovillei*	937	0.302	0.166	0.146	0.385	0.313	0.687
6	KJ628996	*V. pumila*	932	0.304	0.168	0.141	0.387	0.309	0.691
7	JN004623	*V. stangeana*	749	0.308	0.159	0.143	0.39	0.302	0.698
8	KJ628993	*V. tessellata*	931	0.304	0.168	0.142	0.387	0.309	0.691
9	KJ628993	*V. testacea*	936	0.306	0.166	0.142	0.387	0.308	0.692
10	FR832845	*V. cristata*	831	0.302	0.17	0.147	0.381	0.316	0.684
11	KF361631	*V. alpina*	802	0.303	0.163	0.15	0.384	0.313	0.687
12	KJ628992	*V. bensonii*	948	0.303	0.167	0.142	0.388	0.309	0.691
13	KJ628990	*V. brunnea*	936	0.303	0.169	0.143	0.385	0.312	0.688

Clustering or cluster analysis is about dividing data points into groups based on their similarity. The goal is to obtain a set of clusters, where the data point s in a given cluster are as similar to one another as possible and where the clusters are as distinct from one another as possible means algorithm, a density based clustering algorithm was used to calculate the relationship of the 13 *Vanda* species based on the descriptors of the A, T, G, C, A+T and G+C content. The clustering measure was set for Euclidean Distance Squared using a randomized seed of value of 4139130313. Based on the K-means clustering the 13 *Vanda* species were divided into 3 subsets which is shown in Table 4. The three dimensional depiction of the subset clustering based on K-means clustering is shown in Fig. 4. Also the 1D plot showing the clustering relationship between the 3 subset is shown in Fig. 5 where the subset differs by a margin of 1.0 point scale indicating that even these subsets are closely related.

Table 4. K-means clustering of the 13 Vanda species.

Subset Clustering		
Subset 1	**Subset 2**	**Subset 3**
V. stangeana	*V. coerulescens*	*V. lilacina*
	V. denisoniana	*V. liovillei*
	V. flabellata	*V. cristata*
	V. pumila	*V. alpina*
	V. tessellata	*V. brunnea*
	V. testacea	
	V. bensonii	

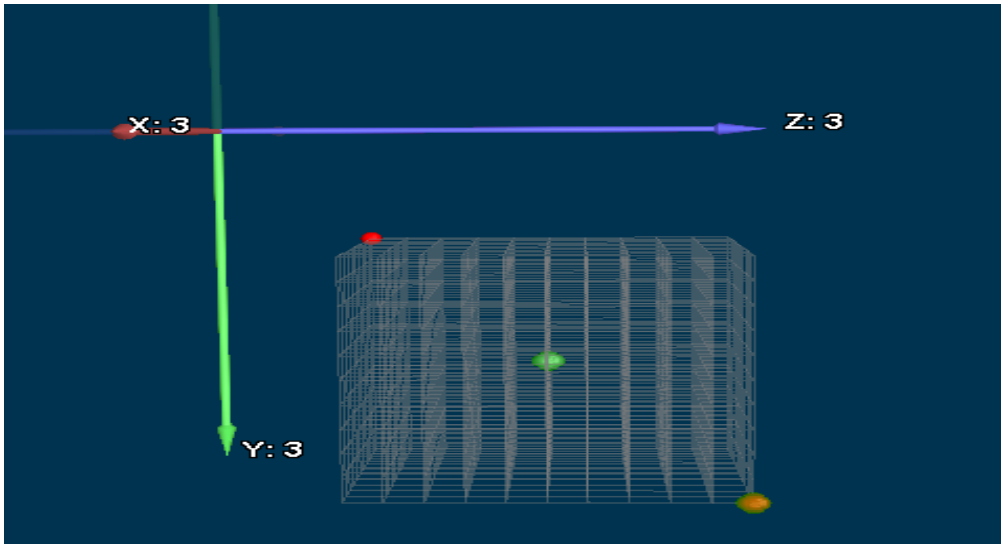

Fig. 4. K-means clustering subset calculation of the 13 *Vanda* species showing red (subset 1), green (subset 2) and yellow (subset 3).

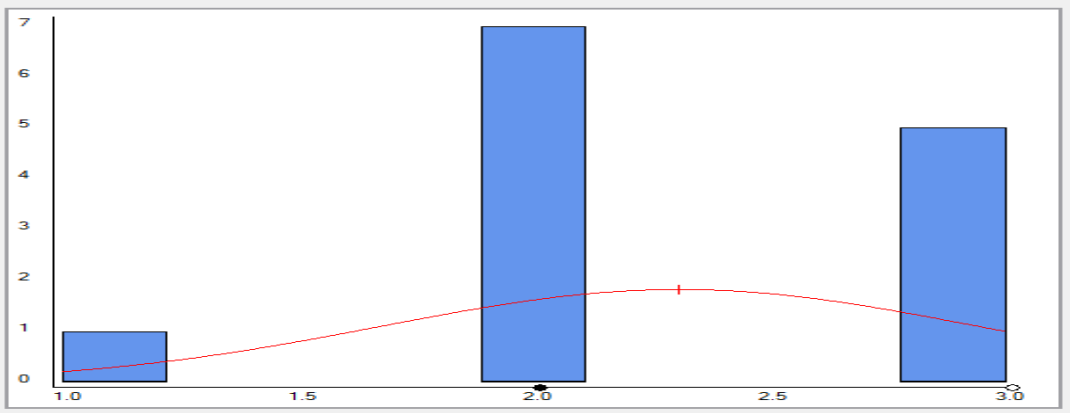

Fig. 5. 1D plot of the generated 3 subsets based on K-means clustering.

Also, the 2D depicting the relationship of *V. bensonii* and *V. liovillei* based on their A+T and G+C content is shown in Fig. 6 where there relationship indicates a mild margin of 0.002 which also confirms that these two species are closely related.

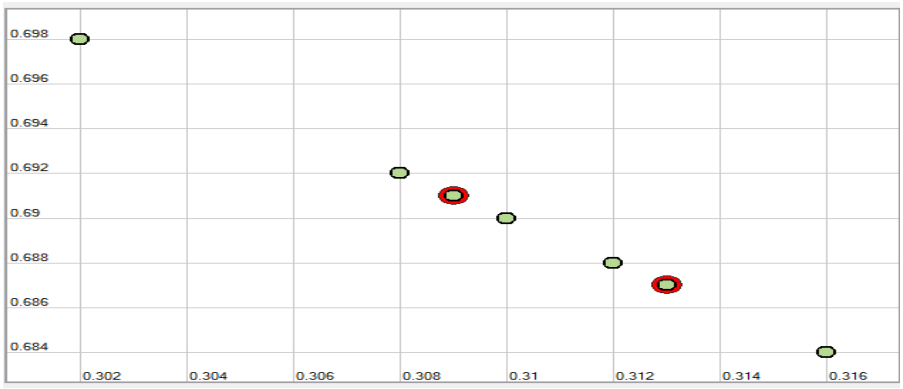

Fig. 6. Relationship between *V. bensonii and V. liovillei* based on their A+T and G+C content.

CONCLUSION

To conclude the nucleotide sequence of maturase K gene of 13 *Vanda* species was studied base on the multiple sequence alignment A, T, G, C, A+T and G+C content analysis revealing that *Vanda bensonii* and *Vanda lioville* are closely related,*Vanda pumila* and *Vanda denisoniana* are also closely related and even *Vanda alpine* and *Vanda cristata* aslo closely related to each other. More over from the statistical clustering based on K means clustering revealed *Vanda bensonii* and *Vanda liovillei* are closely related belonging to the same subset validating the generated phylogenetic tree.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi for providing Bioinformatics Infrastructure Facility at Nagaland University, Lumami, Nagaland, India.

REFERENCES

The Orchids, Natural History and Classification, Robert L. Dressler. ISBN 0-674-87526-5.

Flora of China, V 25, p 471, wan dai lan shu, *Vanda* Jones ex R. Brown, Bot. Reg. 6: ad t. 506. 1820.

Garay, L. (1972). On the systematics of the monopodial orchids, Bot. Mus. Leafl. Harvard University, 23(4): 149-212.

World Checklist of Selected Plant Families: Vanda.

Grove, D. L. (1995). Vandas and Ascocendas. Timber Press, Portland, Oregon. 241 pp.

Motes Martin, R. and Alan, L. Hoffman. (1997). Vandas, Their botany, history and culture. ISBN 0-88192-376-1.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. In Searls, David B. PLoS Computational Biology 7 (3): e1002021.

Waterman Michael, S. (1995). Introduction to Computational Biology: Sequences, Maps and Genomes. CRC Press. ISBN 0-412-99391-0.

Mount David, W. (2002). Bioinformatics: Sequence and Genome Analysis. Spring Harbor Press. ISBN 0-879-69608-7.

Claverie, J.M. and Notredame, C. (2003). Bioinformatics for Dummies. Wiley. ISBN 0-7645-1696-5.

Mount, D.M. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7.

Clustal W2 FAQs http://www.ebi.ac.uk/Tools/clustalw2/

Ng, P.C. and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.*; 11 (5):863-74.

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment". *J Comput Biol* 1 (4): 337–48.

Elias Isaac (2006). "Settling the intractability of multiple alignments. *J. Comput Biol.* 13 (7): 1323–1339. doi:10.1089/cmb.2006.13.1323. PMID 17037961.

Durbin Richard, M., Eddy Sean, R., Krogh Anders and Mitchison Graeme (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (1st ed.), Cambridge: Cambridge University Press, ISBN 0-521-62971-3.

Abu-Jamous, B., Fa, R., Roberts, D.J. and Nandi, A.K. (2013). Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments. Journal of the Royal Society Interface 10 (81): 20120990–20120990.

Corresponding author: Dr. Salam Pradeep Singh, Bioinformatics Infrastructure Facility, Nagaland University, Lumami-798627, Nagaland, India
salampradeep@gmail.com