Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Phylogenetic analysis of chloroplast *mat*K gene from Zingiberaceae for plant DNA barcoding

Dhivya Selvaraj[1], Rajeev Kumar Sarma[1] and Ramalingam Sathishkumar [1,*]

[1]School of Biotechnology and Genetic Engineering, Plant Genetic Engineering Laboratory, Bharathiar University, Coimbatore, India; Ramalingam Sathishkumar* - Email: sathishkumar_ram@hotmail.com; * Corresponding author

received May 08, 2008; revised June 18, 2008; accepted July 20, 2008; published September 08, 2008

Abstract:
MaturaseK gene (*MatK)* of chloroplast is highly conserved in plant systematics which is involved in Group II intron splicing. The size of the gene is 1500 bp in length, located with in the intron of *trn*K. In the present study, *mat*K gene from Zingiberaceae was taken for the analysis of variants, parsimony site, patterns, transition/tranversion rates and phylogeny. The family of Zingiberaceae comprises 47 genera with medicinal values. The *mat*K gene sequence have been obtained from genbank and used for the analysis. The sequence alignments were performed by Clustal X, transition/transversion rates were predicted by MEGA and phylogenetic analyses were carried out by PHYLIP package. The result indicates that the Zingiberaceae genus *Afromonum*, *Alpinia*, *Globba*, *Curcuma* and *Zingiber* shows polyphylogeny. The overall variants between the species are 24% and transition/transversion rate is 1.54. Phylogenetic tree was designed to identify the ideal regions that could be used for defining the inter and intera-generic relationships. From this study it could be concluded that the *mat*K gene is a good candidate for DNA barcoding of plant family Zingiberaceae.

Keywords: *maturaseK*; Zingiberaceae; transition/transversion; phylogenetic tree; consistency index; retention index; MEGA; PHYLIP

Background:
In DNA barcoding, a short DNA sequence is used as a molecular marker for identifying the diversity that exists among plant and animal species. An internal transcribed spacer (ITS) region of nuclear ribosomal cistron is the most commonly used sequence locus for plant molecular systematic investigations **[1]**. Many chloroplast, mitochondrial and nuclear genes have been utilized for studying sequence variation at genus level. Among these genes *rbcL* gene sequence have been analysed by various workers to address plant systematics **[2]**. The *mat*K gene of chloroplast is 1500 bp long, located within the intron of the *trn*K and codes for maturase like protein, which is involved in Group II intron splicing. The two exons of the *trn*K gene that flank the *mat*K were lost, leaving the gene intact in the event of splicing. The gene contains high substitution rates within the species and is emerging as potential candidate to study plant systematics and evolution **[3]**. A homology search for this gene indicates that the 102 aa at the carboxyl terminus are structurally related to some regions of maturase-like polypeptide and this might be involved in splicing of group II introns. It is another emerging gene with potential contribution to plant molecular systematics and evolution **[4]**. The *mat*K-*trn*K gene complex is commonly used for plant evolution studies and addresses the solution for various taxonomic levels **[5]**. The *mat*K gene has ideal size, high rate of substitution, large proportion of variation at nucleic acid level at first and second codon position, low transition/transversion ratio and the presence of mutationally conserved sectors. These features of *mat*K gene are exploited to resolve family and species level relationships. Polymorphism of chloroplast DNA especially *trn*K, *mat*K and intergenic *trnL - trnF* regions has been used to study the phylogeny of various plants **[6]**. The position of *mat*K in the *trn*K gene was determined by comparing with a *mat*K sequence of *Trillium* **[7]**. This data was used to identify molecular markers, which was used for identifying species of these taxa and also to provide the valuable information for both conventional and molecular plant breeding studies **[8]**.

The objective of this study is to evaluate generic, species variation and phylogenetic relationships of Zingiberaceae family by using the chloroplast *mat*K gene sequences available from genbank. Zingiberaceae is a family of flowering plants comprising 47 genera and about 1000 species. This family contains many traditional medicinal plants to cure various disorders. The study will also address the issues of optimal number of nucleotides essential for exploring phylogenesis and the consequences of utilizing different segments of gene as barcodes.

Methodology:
Data collection
The entire coding region of *mat*K sequences of 101 different species belonging to eight taxa of Zingiberaceae. Generic and species information were obtained from taxonomy database of National Centre for Biotechnology information (NCBI) **[9]**.

Sequence analysis
The data analysis was done for the two grouped datasets. One set includes all the plant species of Zingiberaceae for which the sequences are available in genbank to find the interspecies

ISSN 0973-2063
Bioinformation 3(1): 24-27 (2008)

variation. Another dataset includes various genera of Zingiberaceae like *Aframomum*, *Alpinia*, *Curcuma*, *Globba*, *Hedychium* and *Zingiber* to find intergeneric variation. Multiple sequence alignment was performed by using Clustal X, which is offline software that performs optimum alignment for sequence. Alignments were not complicated due to the occurrence of indels and were not included in data analysis **[10]**. Aligned sequences were edited by using the software Bioedit (Biological sequence alignment editor) **[11]**.

Phylogenetic analysis

The basic sequence statistics including nucleotide frequencies, transition/transversion (ns/nv) ratio and variability in different regions of sequences were computed by Molecular Evolutionary Genetics Analysis (MEGA) **[12]**. The sequence data was analyzed by neighbor-joining method using NEIGHBOR program Phylogeny Inference Package (PHYLIP) **[13]** and Unweighed Pair Group Mean Average (UPGMA) methods by using MEGA. Distances were calculated using DNADIST program of PHYLIP. Bootstrapping and decay analysis were performed by NJ plot. Parsimony analysis and various clades were determined by MEGA.

Results and discussion:

Advancement in molecular biology and DNA sequencing techniques has enabled to characterize the genomes of various organisms rapidly. Analyses of the DNA sequences of various species are providing valuable information about their taxonomy, gene makeup and utilizations. In this study, DNA sequence polymorphism of the chloroplast gene *mat*K of Zingiberaceae family was assessed to know the inter-specific and intra- specific differences.

Combined analysis

Multiple sequence alignment shows that, there are variable numbers of Indels in the gene *mat*K. The alignment of *mat*K gene of combined nucleotide sequence shows 497 variable sites and 251 parsimony sites, the overall mean distance is 0.027. The percentage of variants ranges from 0.38 to 8.85. The combined tree show three groups and they are as follows: **Group I** has three clusters A, B and C Figure 2. The cluster A has two clades, taxa *Alpinia* with 10 species and *Plagiostachys* with one clade (a), shows 100% identity and clade (b) consisting of *Elingera* and *Vanoverberghia* shows 99% identity and the taxa *Hornstedtia* exist as monotaxa but it is more closer to clade (b). The cluster B has two clades consisting of *Elettarriopsis, Paramomum* and *Afarmomum. Aframomum* is the largest African genus of the *Zingiberaceae* family that contains about 70 species. They are found in tropical forests and Savannahs. *Renealmia* is grouped into the same clade in the taxon. *Elettarriopsis* and *Paramomum* have similar features and are highly conserved, which shows 100% similarity with no variants. Previous studies shows that the unique taxonomic position of the disputed genus of *Paramomum* and *Elettariopsis* by morphological characters of flowers. Both the genera have evolved from the core clade of *Amomum* through inflorescence and flower diversification **[14]**. The taxa *Afarmomum* and *Renealmia* show 98% similarity with 2% of variants. The cluster C has a single clade, *Burbidgea* and *Riedelia* which shows 100% identity. The genus *Pleuranthodium* and *Siliquamomum* show 35% identity with higher range of variants. **Group II** *Caulokaempferia* and *Globba* show no variants which represent 100% identity, indicating absence of any variants. Taxon *Rhynchanthus* exist as monoclade. **Group III** has two branches; each one with many sub branches. Branch I has three clades and three monoclade in which the taxa *Hedychium* and *Hitchenia* exists as single clade representing 98 percent identity and 2 percent variants. Second clade has taxa *Zingiber* and *Scaphochlamys*, which are highly similar and do not show any variants and the third clade has *Curcuma* and *Stahlianthus* showing one percent variants and are highly conserved. Branch II has three clades A, B and C. A has *Boesenbergia* and *Curcumorpha* which comes under the same clade but it shows sixty percentage of variants. The clade B possesses the genus, *Cautleya* and *Cornukaempferia* which represents 72% of variants with minimal identity and clade C has two sub clades in which the taxa *Haniffia* and *Keampferia* are highly similar with no variance. Similarly the genus *Hemiorchis* and *Gagnepainia* represents the same.

Analysis of individual taxon

The genus *Aframomum* consists of 9 species and it is one of the smaller genus in Zingiberaceae family. Gene sequence for *mat*K is available for only 4 species in the database. The species variation rate is 0.96 and the transition/transversion range is 1.32. The species *A. daneiellii* and *A. sceptrum* shows the branch length of 100. *Alpinia* is the largest, most wide spread and taxonomically complex genus in Zingiberaceae with 230 species consisting of 84 species and are available in taxonomy database. The coding region for *mat*K is available for 10 species in the Genbank and the rest of the sequence has not yet been characterized. The interspecies relationship of this genus shows three percent of variations and 1.52 percent of transition/transversion ratio. The phylogenetic tree consists of 14 informative sites and the overall mean distance is 0.027, transition/tranversion ratio 1.054. The genus *Curcuma* has 37 species and most of the species has potential medicinal value. This genus shows 0.77% of variance and transition/ transversion rate is 0.445 %.

Like *Alpinia, Globba* is another largest genus from Ginger family containing 84 species among which coding region for *matK* is available for 26 species at present in Genbank and they are distributed from eastern Himalayas to South China and from Indochina to Malaysia. Phylogeny for 26 species was studied using MEGA. The analysis shows that the tree has 81 most parsimonious site (length = 110) as shown in Figure 2. The consistency index is 0.789474, the retention index is 0.913669, and the composite index is 0.813996 for all sites and parsimony-informative sites. There are total of 1291 positions in the final dataset, out of which 40 were parsimony informative. Among all genera *Globba* shows largest of 8.85% of variance and decrease in transition and transversion rate of 0.92.

ISSN 0973-2063
Bioinformation 3(1): 24-27 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

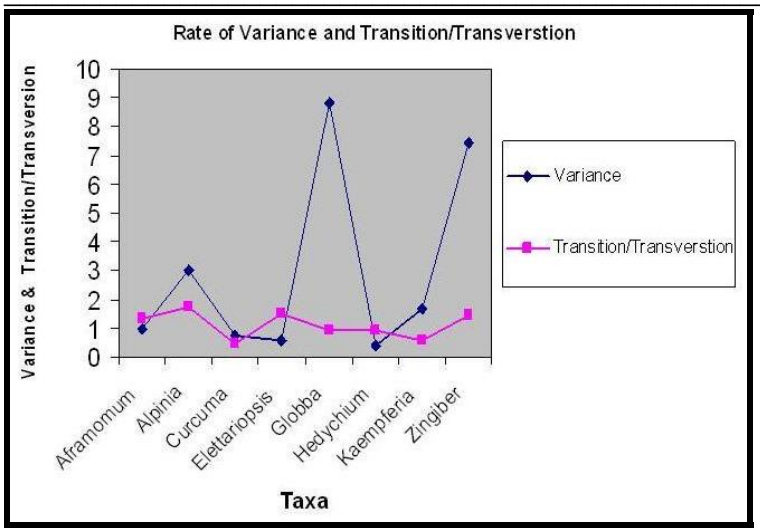

Figure 1: Comparative sequence variation among taxa representing different taxonomic hierarchy using the genbank sequences of *mat*k coding region. The x-axis represents the taxon of *Zingiberacea*; the y-axis represents the variance and transition/transversion ratio for the respective taxon.

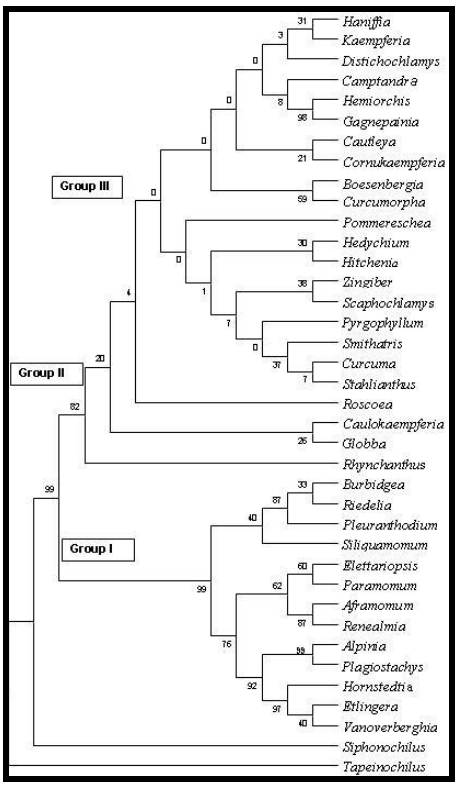

Figure 2: Combined Phylogenetic Tree of Family Zingiberaceae. (Evolutionary relationships of 107 taxa were inferred using the Maximum Parsimony method Tree. 1 out of 1054 most parsimonious trees (length = 219) is shown. The consistency index is 0.794521 (0.691781), the retention index is 0.916201 (0.916201), and the composite index is 0.727941 (0.633810) for all sites and parsimony-informative sites .The MP tree was obtained using the Close-Neighbor-Interchange algorithm with search level in which the initial trees were obtained with the random addition of sequences (10 replicates). All positions containing gaps and missing data were eliminated from the dataset. There were a total of 434 positions in the final dataset, out of which 78 were parsimony informative).

ISSN 0973-2063
Bioinformation 3(1): 24-27 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Hedychium consist of 37 species and phylogenic analysis was carried out for 7 species whose sequence is available. The bootstrap consensus tree is inferred from 500 replicates taken to represent the evolutionary history of the taxa. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates which are collapsed. There are total of 1180 positions in the final dataset, out of which 4 were parsimony informative. The evolutionary relationship of eight taxa is as follows, the most parsimonious trees represent the branch length of 30. The consistency Index is 0.727273, the retention index is 0.750000, and the composite index is 0.675000 for all the sites and parsimony-informative sites. There are total of 1547 positions in the final dataset, out of which 8 were parsimony informative.

Genus *Zingiber* contains 26 species, coding region of *matK* is available for 8 species, and it shows 7.44 percent of intra-generic variation with 1.76 transitions and transversion rates. *Z. officinale* is commonly known as ginger and is closely related with *Zinger gramineum*. The consistency index is 0.590361, the retention index is 0.750000 and the composite index is 0.624724 for all sites and parsimony informative sites. There are total of 1478 positions in the final dataset, out of which 118 were parsimony informative which is shown in Table 1 (see supplementary material) and Figure 2.

Conclusion:

Phylogenetic analysis complements and often outperforms similarity searches, identifying variants, patterns and transition/transversion rate in nucleotide sequence, when addressing sequence identity, especially the reference database does hold high matches in the *mat*K gene. A portable software Molecular Evolutionary Genetics Analysis (MEGA) framework for qualified identification of nucleotide sequences of Zingiberaceae family is provided with inter and intra species relationship. From the combined tree analysis shows that group I have higher boot strap values making the evolutionary sense between the genus of Zingiberaceae family. Thus, from this study it can be suggested that *matK* gene is a good candidate for DNA barcoding of Zingiberaceae family members. It can be also concluded that barcodes for distinguishing the zingiberaceae family members could be selected from the nucleotide positions between 115 to 130, 680 to 690 and 1455 to 1465 of the *mat*K gene.

References:

[01] W. John-kress and J. Kenneth, *Proceedings of National Academy of Sciences*, 8369: 837 (2005) [PMID: 15928076]
[02] M. W. Chase *et al.*, *Annals of the Missouri Botanic Garden*, 80: 528 (1993)
[03] C. Notredame, *Journal of Molecular Biology*, 205: 217 (2000) [PMID: 10964570]
[04] W. Khidir and L. Hongping, *American Journal of Botany*, 830 (1997)
[05] M. Ito and A. Kawamoto, *Journal of Plant Research*, 207: 216 (1999)
[06] K. Wolfe, *Proceedings of the National Academy of Science*, 9054: 9058 (1987) [PMID:3480529]
[07] K. Osaloo and F. Utech, *Journal of Plant Research*, 35: 49 (1999)
[08] L. Pedersen, *Plant Systematics and Evolution*, 239: 258 (2004)
[09] www.ncbi.nlm.nih.gov/GenBank
[10] J. Thompson, *Nucleic Acid Research*, 22: 4673 (1994) [PMID: 7984417]
[11] www.mbio.ncsu.edu/BioEdit
[12] S. Kumar and K. Tamura, *Briefings in Bioinformatics*, 150: 163 (2004) [PMID: 15260895]
[13] J. Felsenstein, *Evolution*, 39: 783 (1985)
[14] B. Efron and E. Halloran, *National Academy of Sciences*, 13429: 34 (1996) [PMID: 8917608]

Edited by P. Kangueane

Citation: Selvaraj *et al.*, Bioinformation 3(1): 24-27 (2008)

Supplementary material

Genus	% of Variance	No. of parsimony site	Overall distance mean	Transition/Transversion ratio
Aframomum	0.96	3	0.04	1.32
Alpinia	3.00	26	0.09	1.733
Curcuma	0.77	18	0.02	0.445
Elettariopsis	0.59	0	0.03	1.54
Globba	8.85	46	0.012	0.929
Hedychium	0.38	4	0.02	0.941
Kaempferia	1.67	4	0.09	0.591
Zingiber	7.44	8	0.06	1.476

Table 1: Transition/transversion ratios of the 8 taxa from Genbank