INTERNATIONAL JOURNAL OF SCIENCE AND NATURE
© 2004 - 2014 Society for Science and Nature (SFSN). All rights reserved
www.scienceandnature.org

PHYLOGENETIC AND STRUCTURAL SCRUTINY OF *matK* GENE FROM WHEAT REPRESENTING POACEAE FAMILY FOR DNA BARCODING

[1]*Geetika Jethra, [2]Mishra A. K., [3]Pandey, P.S., [1]Sharda Choudhary & [2]Chandrasekharan H.
[1]National Research Centre on Seed Spices, Tabiji, Ajmer-305206
[2]Unit of Simulation and Informatics, Indian Agricultural Research Institute, New Delhi-110012
[3]Krishi Anusandhan Bhawan, Indian Council of Agricultural Research, New Delhi-110012
Corresponding Authors email: gjethra08@gmail.com

ABSTRACT

A Maturase-encoding gene (*matK*) is located within the intron of the chloroplast *trn*K gene of wheat and rice. It is highly conserved in plant systematic having ~1500bp length. It is known to be involved in Group II intron splicing. In the present study, phylogenetic tree and motifs were determined to identify the ideal regions that could be used for defining inter and intra-family relationships among *matK*. Intra-family gene sequences of 21 species for Poaceae family and inter-family gene sequences of 17 important families have been analysed to determine relationships on the basis of *matK*. The scrutiny was performed by MEME, ClustalW and MEGA5. It determine motifs, variants, parsimony site, patterns, transition/ transversion rates and phylogeny for the above two datasets. The protein sequences of *Triticum aestivum* and *Oryza sativa* representing Poaceae were retrieved from UniProt database to determine secondary and tertiary structure furthermore representing the comparative analysis using MODELLER 9.10, PHYRE2, MUSTER, LOMETS, and PHYMOL. Results indicate that while in the intra-family (among Poaceae members) there is polyphylogeny whereas in the inter-family there subsist a phylogeny. It is suggested that the *mat*K gene can be a good candidate for DNA barcoding for Poaceae plant family.

KEYWORDS: DNA barcoding, Maturase K, Poaceae, Phylogenetic scrutiny, transition/ transversion.

INTRODUCTION

The rush in different applications of molecular and cellular biology, knowledge to systematic and evolutionary relationships has resulted in considerable assistance to both plant and animal biology and in the emergence of solid interdisciplinary and innovative field as ‘molecular systematic.’ DNA barcoding, a novel concept that has recently emerged, is characterized as a short DNA sequence used as a molecular marker for identifying diversity that exists among plant and animal species [1,2]. The mitochondrial cytochrome *c* oxidase subunit1 (*CO1*) gene was selected to be a DNA barcode for animal species [3]. The difficulty in selecting specific genes to be a plant barcode is due to the imperfection of any gene from either the chloroplast, mitochondrial, or nuclear genomes[1]. The *mat*K gene, formerly known as *orf*K, has emerged as an additional gene with prospective contributions to plant molecular systematics and evolution[4-7]. The *matK* gene is approx. 1500 base pairs (bp) and ~519 amino acids (aa) which codes for Maturase protein. It is positioned within the intron region of chloroplast gene *trn*K on the large single-copy section adjacent to the inverted repeat. *MatK* is very useful in DNA barcoding for the identification of plant families [8-10].

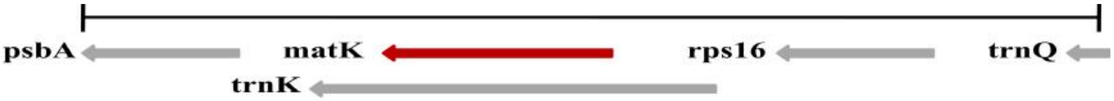

FIGURE 1: Positioning of *matK*-*trnK* gene complex

The *mat*K-*trn*K gene complex (Fig 1) is commonly used for various plant evolutionary relationships and has helped finding answers for various taxonomic studies [11]. 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[12]. The following features of *mat*K gene are exploited to resolve family and species level relationships. This information was utilized to identify and study both molecular and conventional plant breeding studies[13]. Due to highly economic and ecological important plants, Poaceae (also known as grass family) is the most widely attracted plant family in all morphological, molecular phylogenetic studies and structural analysis[14]. Poaceae is one of the largest families in monocots comprising of 700 genera and about 11000 species that inhabit temperate-to-arctic regions throughout the world[15]. Members of this family are cosmopolitan in distribution. Of all the species around 900 species are present in India. The present study emphasizes on three major objectives. First is to determine the individual region of *matK* that best represents the entire gene for Poaceae family and other conserved regions of the gene. Secondly, to predict various physiochemical properties, secondary structure using Protparam (http://web.expasy.org/protparam/), GORIV (http://npsa-pbil.ibcp.fr/cgibin/ npsa_automat. pl?page= /NPSA/ npsa_gor4.html) and tertiary structure using different online tools SWISS-MODEL (http://swissmodel. expasy.org/),LOMETS (http://zhanglab.ccmb.med. umich. edu/LOMETS/) and MUSTER(http://zhanglab. ccmb.med. umich. edu/ MUSTER/). Third objective was to align 3D structures of 2 main species (*Triticum aestivum* and *Oryza sativa*) representing the Poaceae family using PHYMOL standalone tool. The last most significant goal is to determine and evaluate generic and species similarities, variation and phylogenetic relationships of Poaceae family. This is done by using 21 *matK* gene sequences which are available in the GenBank database of NCBI.

MATERIAL & METHODOLOGY

Data Retrieval and Collection

Complete coding regions of gene sequences representing *matK* for 21 different species (*Zea mays, Oryza sativa, Brachypodium distachyon, Lolium perenne, Sorghum bicolor, Triticum aestivum, Coixlacryma-jobi, Dendrocalamus latiflorus, Anomochloa marantoidea, Hordeum vulgare, Festucaarundinacea, Panicum virgatum, Rhynchoryza subulata, Phyllostachyspropinqua, Leersiatisserantii, Saccharum hybrid cultivar, Agrostisstolonifera, Ferrocalamusrimosivaginus, Bambusaemeiensis, Acidosasapurpurea, Indocalamuslongiauritus*)of Paocea family were retrieved. Generic and species information along with sequences were obtained from taxonomic and GenBank database of National Centre for Biotechnology information (NCBI). Also fully annotated, non redundant and complete protein sequences for all the above species were obtained from UniProt database (http://www. uniprot.org/).

Motif Prediction

The motifs or the highly conserved region in groups of related DNA or protein sequences of *matK* which represent the entire *Poaceae* family are predicted with MEME and MAST (http://meme.sdsc.edu/meme/ intro. html)

Physiochemical properties and Secondary Structure Prediction

The physiochemical properties like molecular weight, theoretical isoelectric point, extinction coefficients, instability index, grand average of hydropathicity (GRAVY) etc. were determined and tabulated for *matK* protein sequences derived from 21 genera using Protparam. Also the secondary structures for *matK* (Maturaes K) protein of *Triticum aestivum* and *Oryza sativa* were predicted using GORIV.

Tertiary Structure Prediction, Validation and Alignment

The tertiary structure for *matK* (Maturase K) protein of *Triticum aestivum* and *Oryza sativa* were predicted using LOMETS (Local Meta-Threading-Server), MUSTER [16,17] and PHYRE2 (http://www.sbg.bio. ic.ac.uk/ phyre2/ html/page. cgi?id= index).The predicted models were validated using Structural Analysis and Verification Server (SAVES) (http://nihserver.mbi.ucla.edu/SAVES/). The alignment for best 2 predicted structures were analysed and visualized using PHYMOL standalone server.

Sequence and Phylogenetic analysis for *matK*

Data analysis using CLUSTALW, a Multiple Sequence Alignment (MSA) tool for all the 21 plant species present in *Poaceae* was performed for which the complete sequences are available in GenBank to find the interspecies variation and amino acid composition. Aligned sequences were edited by using Bioedit (Biological sequence alignment editor) (www.mbio. ncsu. edu/ BioEdit).

Phylogenetic analysis was completed using Maximum Parsimony and Neighbour joining methods. This was through MEGA5 and phylip3.2 [2]. The analysis was done for two grouped datasets. One set includes all the 21 plant species belonging to Poaceae to uncover the interspecies variation. Another dataset includes major families which also show the presence of *matK* gene to find relationship between them.

RESULTS & DISCUSSION

Motif Prediction

Term Motif refers to a set of highly conserved region or short, recurring patterns in DNA that are presumed to have biological function or contiguous secondary structure elements that either have a particular functional significance or define a portion of an independently folded domain [18]. Dataset of 21 different species representing Poaceae shows 3 predominant motifs. These motifs have a length of 50 bp and lie approx. between 300-375, 1100-1200 and 1325-1375bp respectively (Fig 2).

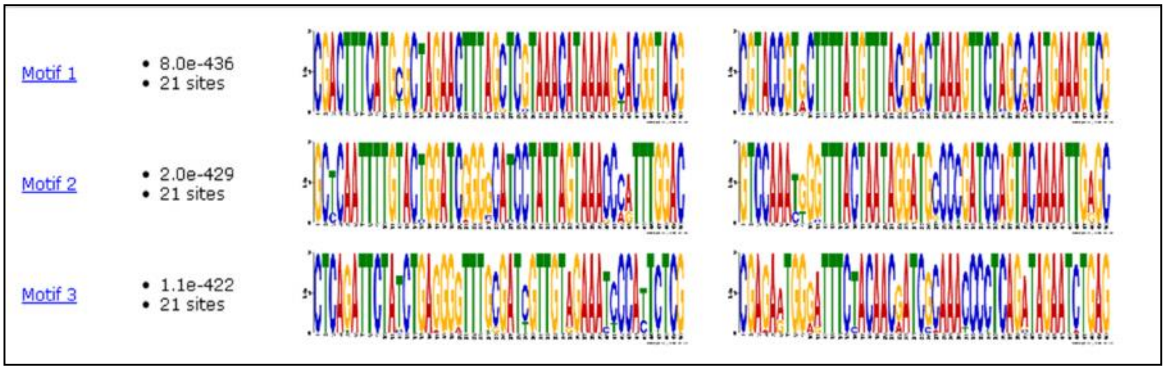

FIGURE 2: 3 motifs predicted for *matK* protein for Paoceae family

The regular pattern were also predicted for the above 3 motifs.

Motif 1:
CGACTTTCATG[CT]GCTAGAACTTTAGCTCGTAAA
CATAAAAG[CT]ACGGTACG
Motif 2:GCTCAATTTTGTACTGGATCGGGGCATCCTATT
AGTAAACC[CA][AG]TTTGGAC
Motif 3:CTCAGATTCTATCTGAGGGGTTTGCGAT[CT]GTT
GTAGAAATCCCA[TC]TCTCG

Physiochemical properties and Secondary Structure Prediction

Physiochemical properties of completely sequenced *matK* protein sequences belonging to 19 genera for Poaceae were deduced using Protparam (Table 1). The aliphatic index of a protein is the relative volume occupied by its aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. The aliphatic index for *matK* protein for all the 19 species lies between 87 and 96. The instability index provides an estimate stability of a protein in vitro. The protein whose instability index is smaller than 40 is predicted as stable. This analysis showed that *matK* protein for considered genera are unstable in vitro. The grand average of hydropathy (GRAVY) value for a protein is the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence.

TABLE1: Species, Mol. Wt. (Molecular Weight), Theo.PI (Theoretical Iso-electric point), Instability Index, GAVY (Grand average of hydropathicity).

S.No.	Species	Mol. Wt.	Theo. PI	Aliphatic Index	Instability Index	GRAVY
1.	*Zea mays*	61148.8	9.27	90.41	45.61	-0.124
2.	*Oryza sativa*	62332.6	9.44	92.39	46.55	-0.128
3.	*Brachypodium distachyon*	61212.1	9.49	89.06	47.96	-0.154
4.	*Lolium perenne*	61337.3	9.39	87.18	44.70	-0.158
5.	*Sorghum bicolour*	61539.4	9.41	89.86	47.91	-0.158
6.	*Triticum aestivum*	61392.4	9.42	88.12	46.20	-0.175
7.	*Coixlachryma-jobi*	62766.8	9.36	92.62	45.49	-0.109
8.	*Dendrocalamus latiflorus*	61330.3	9.57	88.47	46.25	-0.191
9.	*Hordeum vulgare*	61353.2	9.42	88.88	46.56	-0.154
10.	*Anomochloama rantoidea*	61252.4	9.63	95.53	46.78	-0.106
11.	*Panicum virgatum*	61289.2	9.45	90.04	43.56	-0.132
12.	*Rhynchoryzasubulata*	61488.5	9.62	90.04	45.73	-0.161
13.	*Phyllostachyspropinqua*	61228.1	9.55	88.85	45.30	-0.187
14.	*Leersiatisserantii*	61582.6	9.58	88.71	46.24	-0.152
15.	*Agrostisstolonifera*	61532.5	9.59	88.12	44.12	-0.163
16.	*Ferrocalamusrimosivaginus*	61362.4	9.59	88.85	44.81	-0.174
17.	*Bambusaemeiensis*	61465.4	9.54	88.47	45.53	-0.194
18.	*Acidosasapurpurea*	61180.1	9.59	88.85	45.83	-0.188
19.	*Indocalamuslongiauritus*	61228.1	9.55	88.85	45.30	-0.187

The secondary structure was predicted for only *Triticum aestivum* (Fig 3.a) and *Oryza sativa* (Fig 3.b), represent Poaceae family using GOR IV and ASSPS2 algorithm.

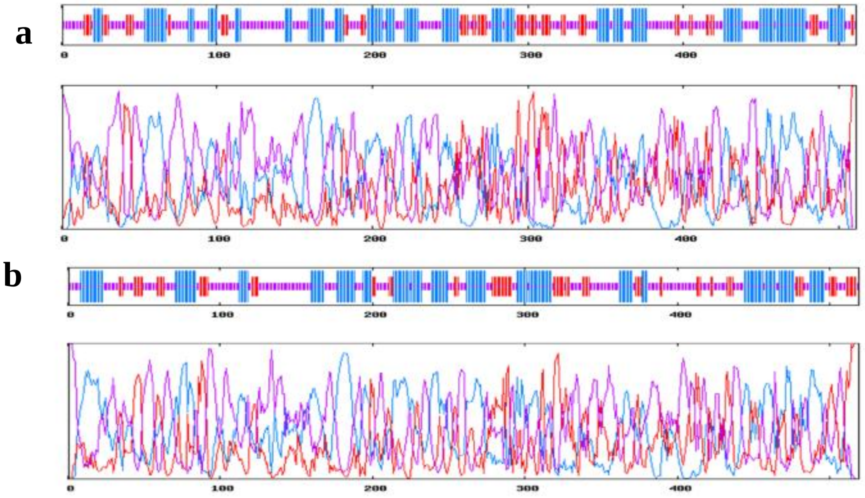

FIGURE 3: Predicted Secondary Structure (a) *Triticum asetivum* (b) *Oryza sativa*
[Blue: Helix, Red: Extended strand, Magenta: Coil]

Tertiary Structure Prediction, Validation and Alignment

Tertiary structures for above two proteins were predicted using homology modelling (MODELLER9.10). Since the templates with good sequence identity for Maturase K, were not available, they were modelled based on threading method using LOMETS (Local Meta-Threading-Server), MUSTER and PHYRE2.

LOMETS a quick and automated protein prediction tool gave tertiary structures and spatial constraints using nine state-of-the-art threading programs run in a cluster. It predicts the structure with 7% more accuracy. MUSTER (MUlti-Sources ThreadER) is also a protein threading algorithm to identify the template structures from PDB library. It generates sequence-template alignments by combining sequence profile-profile alignment with multiple structure information.

PHYRE2 (Protein Homology/analogy Recognition Engine) also incorporates an *ab initio* folding simulation called Poing, to model regions of proteins with no detectable homology to known structures.

SAVES, was used to validation 3D structure, it provides a package of different structure validation tools (PROCHECK, WHAT_CHECK, ERRAT, VERIFY_3D and PROVE). According to this analysis, the residues of predicted structures were within the allowed regions of Ramachandran plot. This indicates that both the predicted models are of good quality and they can be used for further studies.

The above validated structures for *matK* proteins of *Triticum aestivum* and *Oryza sativa* were superimposed using PyMOL (Fig 4). The superimposition shows that most part of the 2 structures is super imposable, showing structural similarity.

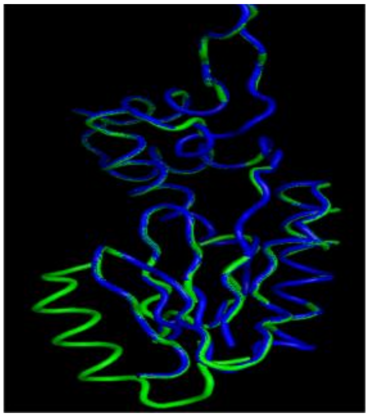

FIGURE 4: Superimposed structure of Wheat (Blue) and Rice (Green)

Sequence and Phylogenetic analysis for *matK*

ClustalW an online multiple sequence alignment (MSA) tool showed that there are variable numbers of Indels in the *matK* gene. MSA for all 21 *matK* nucleotide sequences showed 432 variable sites, 233 parsimony sites, 1116 conserved sites and 199 singletons having overall mean distance of 0.070. The pair-wise distance ranges from 0.000 to 0.132 (**Fig 5**).

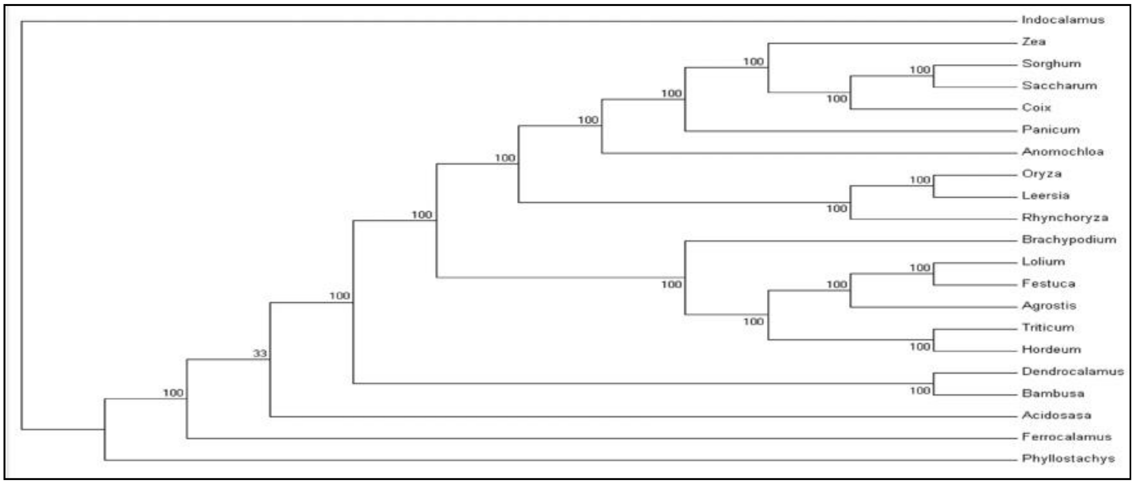

FIGURE 5: Phylogenetic Tree of Family Poaceae (Evolutionary relationships of 21 members were inferred using the Maximum Parsimony method Tree. 1 out of 3 most parsimonious trees (length = 686) is shown. The consistency index is 0.771137 (0.665245), the retention index is (0.794503), and the composite index is (0.612670) for all sites and parsimony-informative sites (in parentheses). 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 1531 positions in the final dataset.

The phylogenetic analysis of 21 species shows that the tree has consistency index of 0.771137 (0.665245), retention index (0.794503), and composite index (0.612670) for all sites and parsimony-informative sites (in parentheses). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates). The MP tree was obtained using the Close-Neighbor-Interchange algorithm with search level 1 in which the initial trees were obtained with the random addition of sequences (10 replicates). All positions containing gaps and missing data were eliminated. There final dataset includes a total of 323 positions. The estimated Maximum Likelihood Transition/ Transversion bias (*R*) was estimated to be 1.17. Substitution pattern and rates were estimated under the Kimura (1980) 2-parameter model [19]. The total branch length of optimal tree is 0.43796017.

Phylogenetic analysis was also performed for 17 different families to predict interfamily relationship on the basis of *matK* gene (Fig 6). MSA for all 17 genes showed 1106 variable sites, 638 parsimony sites, 448 conserved sites and 446 singletons.

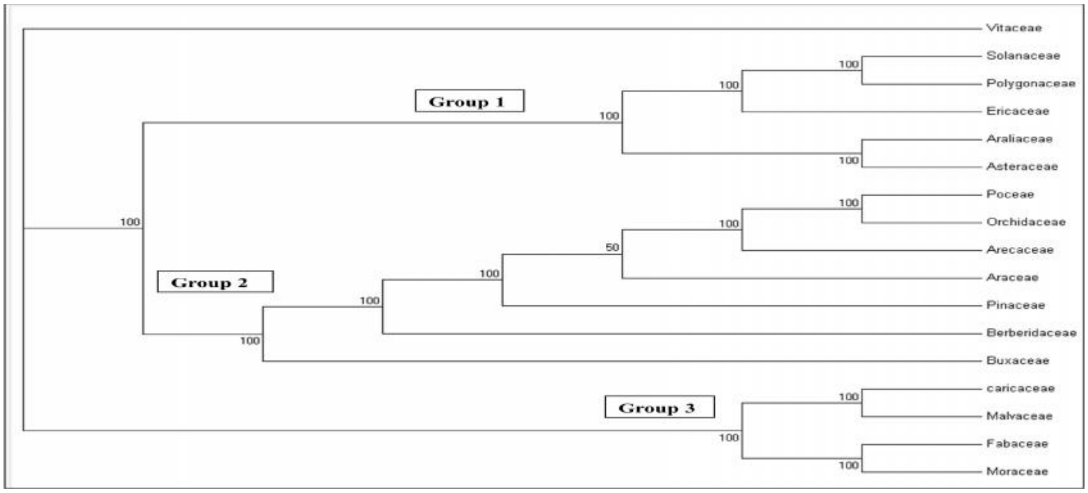

FIGURE 6: Phylogenetic Tree of 17 major families (Evolutionary relationships of 17 members were inferred using the Maximum Parsimony method Tree. 1 out of 2 most parsimonious trees (length = 686) is shown. The consistency index is 0.588280 (0.481922), the retention index is 0.397344 (0.397344), and the composite index is 0.233749 (0.191489) for all sites and parsimony-informative sites (in parentheses). 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 1531 positions in the final dataset.

The combined tree showed three groups:

- Group 1 has Solanaceae, Polygonaceae which belongs to clade A and Araliaceae, Asteraceae comprise of clade B. Ericaceae also exists in the similar group but is more closer to clad A.
- Group 2 consists of the remaining families including Poaceae which show 100% identity with Orchidaceae.
- Group 3 also has two clades A and B consisting of Caricaceae, Malvaceae and Fabaceae, Moraceae respectively having 100% identity.

CONCLUSION

matK gene has been used in addressing systematic questions in various families; its potential application to plant systematics is the recent area of research. In the present study, the plant systematic for Poaceae family has been proved with the help of structural comparison and phylogenetic analysis. The structural comparison of *matK* gene for *Triticum aestivum* and *Oryza sativa* has shown that there is a deviation in the overall structure of matK gene but the domain region are superimposing. The phylogenetic analysis of Poaceae family (intrafamily: 21 species) and interfamily: 17 families were performed using MEGA5. Combined tree analysis showed that group I has higher boot strap values making the evolutionary sense between the genus of Poaceae family. Thus, from this study it can be suggested that *matK* gene is a good candidate for DNA barcoding for Poaceae family members.

ACKNOWLEDGMENT

The authors are thankful to Dr. T. Madan Mohan, Department of Biotechnology (DBT), New Delhi (INDIA) for providing support through capacity build-up funds and guidance for counting the research project.

REFERENCES

[1]. J. Yu, J. Xue, SL. Zhou, New universal *matK* primers for DNA barcoding angiosperms, Journal of Systematics and Evolution 49 (2011)176–181.

[2]. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Molecular Biology and Evolution 28 (2011) 2731–2739.

[3]. PDN. Hebert, A. Cywinska, SL. Ball, JR. DeWaard, Biological identifications through DNA barcodes, Proceedings of the Royal Society B: Biological Sciences 270 (2003) 313–321.

[4]. LA. Johnson, and DE. Soltis, *mat*K DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str., Systematic Botany19 (1994-1995) 143-156.

[5]. KP. Steele and R. Vilgalys, Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the Plastid gene *mat*K, Systematic Botany19 (1994) 126-142.

[6]. H. Liang, and KW. Hilu, Application of the *mat*K gene sequences to grass systematics, Canadian Journal of Botany74 (1996)125-134.

[7]. PA. Gadek, PG. Wilson and CJ. Quinn, In press. Phylogenetic reconstruction in Myrtaceae using *mat*K, with particular reference to the position of 41 *Psiloxylon*and *Heteropyxis*, Australian Systematic Botany.

[8]. YL. Qiu, J. Lee, F. Bernasconi-Quadroni, DE. Soltis, PS. Soltis, M. Zanis, EA. Zimmer, Z. Chen, V. Savolainen and MW. Chase, The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes, Nature 402(1999) 404–407.

[9]. XX. Li, ZK. Zhou, The higher-level phylogeny of monocots based on *matK*, *rbcL*and 18S rDNA sequences, Acta Phytotaxonomica Sinica 45 (2007) 113–133.

[10]. X. Gao, YP. Zhu, BC. Wu, YM. Zhao, JQ. Chen, YY. Hang, Phylogeny of Dioscorea sect. Stenophora based on chloroplast *matK*, *rbcL*and *trnL-F* sequences, Journal of Systematics And Evolution 46 (2008) 315–321.

[11]. M. Ito and A. Kawamoto, Phylogenetic Relationships of Amaryllidaceae Based on *matK* Sequence Data, Journal of Plant Research, (1999) 207-216.

[12]. KW. Hilu and H. Liang, The *mat*K Gene: Sequence Variation and Application in Plant Systematics, American Journal of Botany 84 (1997) 830-839.

[13]. L. Pedersen, Phylogenetic analysis of the subfamily Alpinioideae (Zingiberaceae) with special emphasis on *Etlingera* Giseke, based on nuclear and plastid DNA, Plant Systematics and Evolution (2004) 239-258.

[14]. W Zhang, Phylogeny of the Grass Family (Poaceae) from rpl16 Intron Sequence Data, Molecular Phylogenetics and Evolution Sci Verse 15 (2000) 135–146.

[15]. L. Watson and MJ. Dallwitz, The Grass Genera of the World C.A.B International Cambgidge (1994)

[16]. S. Wu, Y. Zhang, LOMETS: A local meta-threading-server for protein structure prediction, Nucleic Acids Research 35 (2007) 3375-3382.

[17]. S. Wu, Y. Zhang. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Structure, Function, and Bioinformatics 72 (2008) 547-556.

[18]. P D'haeseleer, Nature Publishing Group http://www.nature.com/naturebiotechnology, (2006)

[19]. D. Selvaraj, RK. Sarma and R. Sathishkumar, Phylogenetic analysis of chloroplast *mat*K genefrom Zingiberaceae for plant DNA barcoding, Bioinformation 3 (2008)24-27