CHLOROPLAST *matK* GENE PHYLOGENY OF SOME IMPORTANT SPECIES OF PLANTS

Ayşe Gül İNCE[1] Mehmet KARACA[2] A. Naci ONUS[1] Mehmet BİLGEN[2]
[1]Akdeniz University Faculty of Agriculture Department of Horticulture, 07059 Antalya, Turkey
[2]Akdeniz University Faculty of Agriculture Department of Field Crops, 07059 Antalya, Turkey
Correspondence addressed E-mail: mkaraca@akdeniz.edu.tr

Abstract

In this study using the chloroplast *matK* DNA sequence, a chloroplast-encoded locus that has been shown to be much more variable than many other genes, from one hundred and forty two plant species belong to the families of 26 plants we conducted a study to contribute to the understanding of major evolutionary relationships among the studied plant orders, families genus and species (clades) and discussed the utilization of *matK* for molecular phylogeny. Determined genetic relationship between the species or genera is very valuable for genetic improvement studies. The chloroplast *matK* gene sequences ranging from 730 to 1545 nucleotides were downloaded from the GenBank database. These DNA sequences were aligned using Clustal W program. We employed the maximum parsimony method for phylogenetic reconstruction using PAUP* program. Trees resulting from the parsimony analyses were similar to those generated earlier using single or multiple gene analyses, but our analyses resulted in strict consensus tree providing much better resolution of relationships among major clades. We found that gymnosperms (*Pinus thunbergii, Pinus attenuata and Ginko biloba*) were different from the monocotyledons and dicotyledons. We showed that *Cynodon dactylon*, *Panicum capilare*, *Zea mays* and *Saccharum officiarum* (all are in the C_4 metabolism) were improved from a common ancestors while the other cereals *Triticum Avena*, *Hordeum*, *Oryza* and *Phalaris* were evolved from another or similar ancestors. In this study, relationships within and between Fabaceae (*Fabales*), *Rosaceae* (*Rosales*), *Moraceae* (*Uriticales*), *Cannabeceae* (*Uriticales*) and *Uriticaceae* (*Uriticales*). *Malvaceae* (*Malvales*) and *Brassicaceae* (*Brassicales*) were also discussed. Overall, our results indicated that *matK* gene provides well-defined relationships within and among the families, genus and species; therefore its sequence can be successfully used in Single Nucleotide Polymorphism (SNP) or part of the sequence as DNA fragment analysis using PCR in plant systematic.

Keywords: Bootstrap, Plant Families, Chloroplast, *matK*, Molecular Phylogeny

Bazı Önemli Bitki Türlerinin Kloroplast *matK* Geni Filogenisi

Özet

Bu çalışmada, genomda bulunan genlerin çoğundan daha fazla varyasyon gösteren ve kloroplasta bulunan *matK* geninin moleküler filogeni çalışmalarında kullanımı araştırılmıştır. Çalışma da kloroplast *matK* geni DNA sekansları kullanılarak 26 farklı familyaya ait 142 bitki türünde, takım, familya, cins ve türler arasındaki evrimsel ilişkilerin belirlenebilmesi amaçlanmıştır. Cins ya da türler arasındaki genetik akrabalıkların belirlenmesi modern ve geleneksel ıslah metodları kullanılarak gerçekleştirilecek genetik ilerlemeler için çok önemlidir. 730-1545 nükleotid dizilimli kloroplast *matK* geni sekansı Gen Bankası'ndan alınmış ve Clustal W programı kullanılarak bu DNA sekanslarının sekansta bulanan baz içerikleri sıralanmıştır. "PAUP*" programı kullanılarak "Maksimum parsimony" metoduyla filogenetik ilişkiler belirlenmiştir. Parsimony analizinden elde edilen filogeni sonuçları daha önceden yapılmış olan tekli ve çoklu gen analizleri sonuçlarıyla benzerlik gösterdiği gibi elde edilen sonuçlar önemli türlerin akrabalıklarını belirlemede de daha iyi sonuçlar vermiştir. Analiz sonuçları açık tohumlu bitki türlerinin (*Pinus thunbergii, Pinus attenuata and Ginko biloba*) monokotiledonlar ve dikotiledonlardan oluşan kapalı tohumlu bitki türlerinden oldukça farklı olduğunu göstermiştir. Ayrıca C_4 metabolizmasına sahip bitkilerden *Cynodon dactylon*, *Panicum capilare*, *Zea mays* ve *Saccharum officiarum*'un ortak atadan gelmelerine karşın C_3 metabolizmasına sahip *Triticum, Avena*, *Hordeum*, *Oryza* ve *Phalaris* bitkilerinde farklı ya da benzer atadan yayıldıkları tespit edilmiştir. Bu çalışmada ayrıca Fabaceae (*Fabales*), *Rosaceae* (*Rosales*), *Moraceae* (*Uriticales*), *Cannabeceae* (*Uriticales*) and *Uriticaceae* (*Uriticales*). *Malvaceae* (*Malvales*) ve *Brassicaceae* (*Brassicales*) arasındaki veya içerisindeki ilişkiler tartışılmıştır. Genel olarak alınan sonuçlar *matK* gen sekansı Tek Dizi Polimorfizmi (TDP) çalışmalarında, veya Polimeraz Zincir Reaksiyonu, (PZR) analizleriyle familya, cins ve türlerin kendi içlerinde ve türler arasındaki ilişkileri en iyi şekilde belirlenmesini sağlayabileceği gibi bitki sistematiğinde de başarıyla kullanılabileceğini göstermiştir.

Anahtar Kelimeler: Bootstrap, Bitki Familyaları, Kloroplast, *matK*, Moleküler Filogeni

1. Introduction

Recent advances in DNA sequencing technologies and molecular biology enable us to characterize genomes of organisms and now many ongoing genome projects for various species are providing valuable insights into their biology and utilizations. The application of molecular biology information to systematic and evolution has resulted in significant contributions to plant systematics and in the emergence of molecular systematics as a solid interdisciplinary field (Mort *et al.*, 2001).

Nucleotide sequence variability in chloroplast DNA (cpDNA or plastid DNA) at inter-(between families or genus) and intra-specific level (within species or varieties) has been surveyed primarily in order to analyze the phylogenetic relationships and plant identification studies (Tamura *et al.*, 2004).

The *matK* gene, a chloroplast genome encoded locus located within the intron of the chloroplast gene *trnK,* encodes a maturase on the large single-copy section adjacent to the inverted repeat of every plant families, has high rates of substitution compared to other chloroplast genes and its DNA sequence is one of the least conserved plastid genes; therefore, has been effectively used in plant evolution and addresses the phylogenetic questions in various taxonomic levels (Ito *et al*., 1999; Fuse and Tamura, 2000).

The *matK* gene has several advantages in comparison to other genes including the organelle genome genes. First of all the *matK* gene evolves approximately three times faster than the widely used plastid genes *rbcL* and *atpB*. It is in the chloroplast genome and in many cases it is maternally inherited. This gene has a reasonable size, high rate of substitution, large proportion of variation at the first and the second codon positions, low transition-transversion ratio, and the presence of mutationally conserved sectors. Research has shown that the variations at nucleic acid (DNA) and amino acid levels evenly distributed throughout the entire gene, and the 5' region of the *matK* gene appears to have more variation than the 3' region in many monocotyledons and dicotyledons. Because of these unique characteristics, *matK* gene sequences (at both nucleic acids and amino acid sequence levels) have been used successfully to resolve family and even species level relationships (Steele and Vigalys, 1994; Brochmann *et al.*, 1998; Koch *et al.*, 2001; Tamura *et al.*, 2004).

In this article, we report the results of phylogenetic analyses of chloroplast *matK* gene sequences from 142 plant species belong to families of 26 plants and 22 orders. Relationships within and between monocotyledons, dicotyledons and gymnosperms were discussed. This information may facilitate the utilization of the genetic resource in wild germplasm and provide an important basis for addressing the many intriguing questions involving the biogeography and genome evolution studies. Also determined genetic relationships may provide valuable information for both conventional and modern plant breeding studies.

2. Material and Methods

A total of 142 *matK* sequences were downloaded from GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/index.html). These *matK* DNA sequences were then aligned using the Clustal W program (Thompson *et al.*, 1994). Result of the alignments showed that there were variable numbers of indels in *matK* gene. All gap characters were scored as missing data rather than a fifth character. Sequences ranging from 730 to 1545 bp in length provided a data set of 2089 bp after alignment.

Phylogenetic analyses of the sequence data were conducted using the parsimony method using *Petroselinum crispum matK* DNA sequence as reference. The sequence data were also analyzed with a neighbor-joining (NJ) and Unweighted Pair Group Mean Average UPGMA methods as implemented in PAUP* 4.0 (Swofford, 2002). The level of support for branches of the phylogenetic trees was evaluated with the bootstrap analysis (Felsenstein, 1985) to verify the length of

the branches based on 100 replicates, using the branch-and-bound search and the bootstrap support for each clade was estimated based on 100 replicates.

3. Results and Discussion

Recent advances in molecular biology and DNA sequencing techniques enable scientists to characterize the genomes of organisms and now ongoing various genome projects for various species are providing valuable information into their taxonomy, gene makeup and utilizations. In this study we conducted nucleotide sequence polymorphisms of the chloroplast *matK* gene, for 26 families from monocotyledons, dicotyledons and gymnosperms to assess the degree and pattern of inter-specific and intra-specific differences. Bootstrap support values for all of those families in phylogenetic tree were strong (80–100 %) indicating the resolution and reliability of the results.

3.1. *Monocotyledons*

Results clearly indicated that gymnosperms (*Pinus thunbergii, Pinus attenuata and Ginko biloba*) were different from the monocotyledons and dicotyledons as shown in Figure 1 and Figure 2 as expected. The gymnosperms were placed at the base of the UPGMA tree (Figure 1) and they were genetically distal from the other species (Figure 2). Monocotyledons consisted of *Alliaceae*, *Agavaceae*, *Iridaceae*, *Bromoliaceae*, *Liliaceae*, *Orchidaceae*, *Zingibaraceae* and *Poaceae* families. In the *Poaceae* family, there were clear identification between the cold season cereals and summer season cereals. The sequence polymorphisms resulted from the DNA sequence indels or substitutions of the *matK* gene indicated that *Cynodon dactylon*, *Panicum capilare*, *Zea mays* and *Saccharum officiarum* (all are in the C_4 metabolism) were evolved from a common ancestor while other cereals *Triticum Avena*, *Hordeum*, *Oryza* and *Phalaris* were evolved from another or similar ancestor. In the monocotyledon group, we also observed another subgroup consisting of *Zingiber*, *Lilium*, *Ananas*, *Yucca*, *Iris*, *Allium* and *Brassia*. These observed relationships within the monocotyledons were strongly supported (100 % bootstrap value) and were like nearly all previous molecular analyses (Fuse and Tamura, 2000).

The level of support for branches of the phylogenetic trees was evaluated with the bootstrap analysis and NJ method (Figure 2). Within the monocotyledons the highest genetic difference was observed between *Ananas ananassoides* (representing the water-conserving mode of photosynthesis known as crassulacean acid metabolism (CAM)) and *Avena sativa* while the most closed relationship was observed among the *Oryza* spp. Analyses also clearly differentiated the perennial monocotyledons from the annual ones. Within the monocotyledon group *Ananas ananassoides* (*Bromeliaceae*) has long been regarded as an isolated and natural group but its taxonomic classification is still incomplete. Results indicated that *Ananas ananassoides* was not within the large order *Poales* but it was related to *Poales* (Crayn *et al.*, 2004).

The bootstrap value of the monocotyledon branch was 100 % indicating clear differentiation of monocotyledon plants from dicotyledons and gymnosperms. Observed similarities also indicated that there are more common sets of ortholog genes at across cereals than other plant species in the monocotyledon as it was also stated by Kilel (2004).

3.2. *Dicotyledons*

Based on the *matK* DNA sequences, dicotyledons were divided into several subgroups (Figure 1 and 2), consisting of *Brassicaceae, Fabaceae, Urticaceae, Chenopodoceae, Malvaceae, Rosaceae, Lamiaceae, Oleaceae, Theaceae, Vitaceae, Hamameliaceae, Asteraceae, Umbellifera* families. Within the dicotyledons, while the highest genetic differences were observed in the *Vicia* genus, the most related genus was *Malus* spp. Genetic differences were greater in dicotyledons when compared to that of the cotyledons.

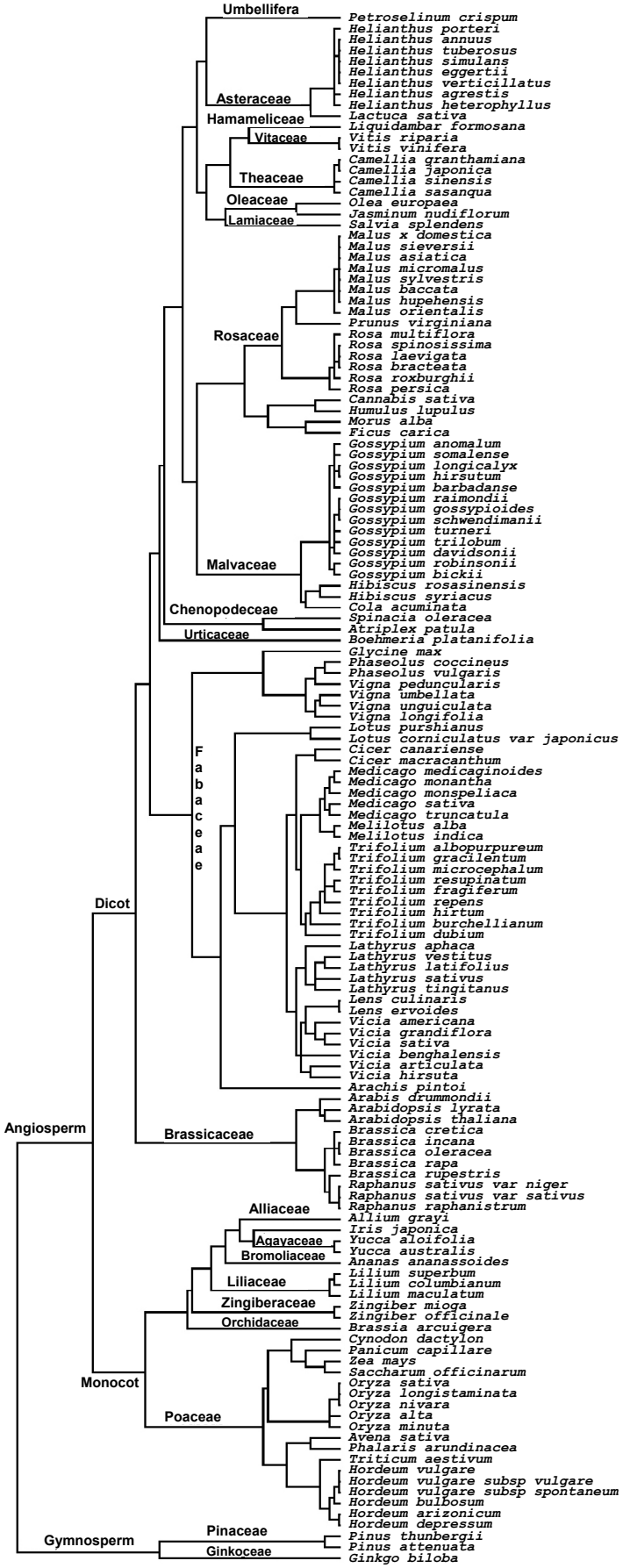

Figure 1. Most parsimonious UPGMA tree for 142 species of plants by the maximum parsimony method based on nucleotide sequences of the *matK* gene.

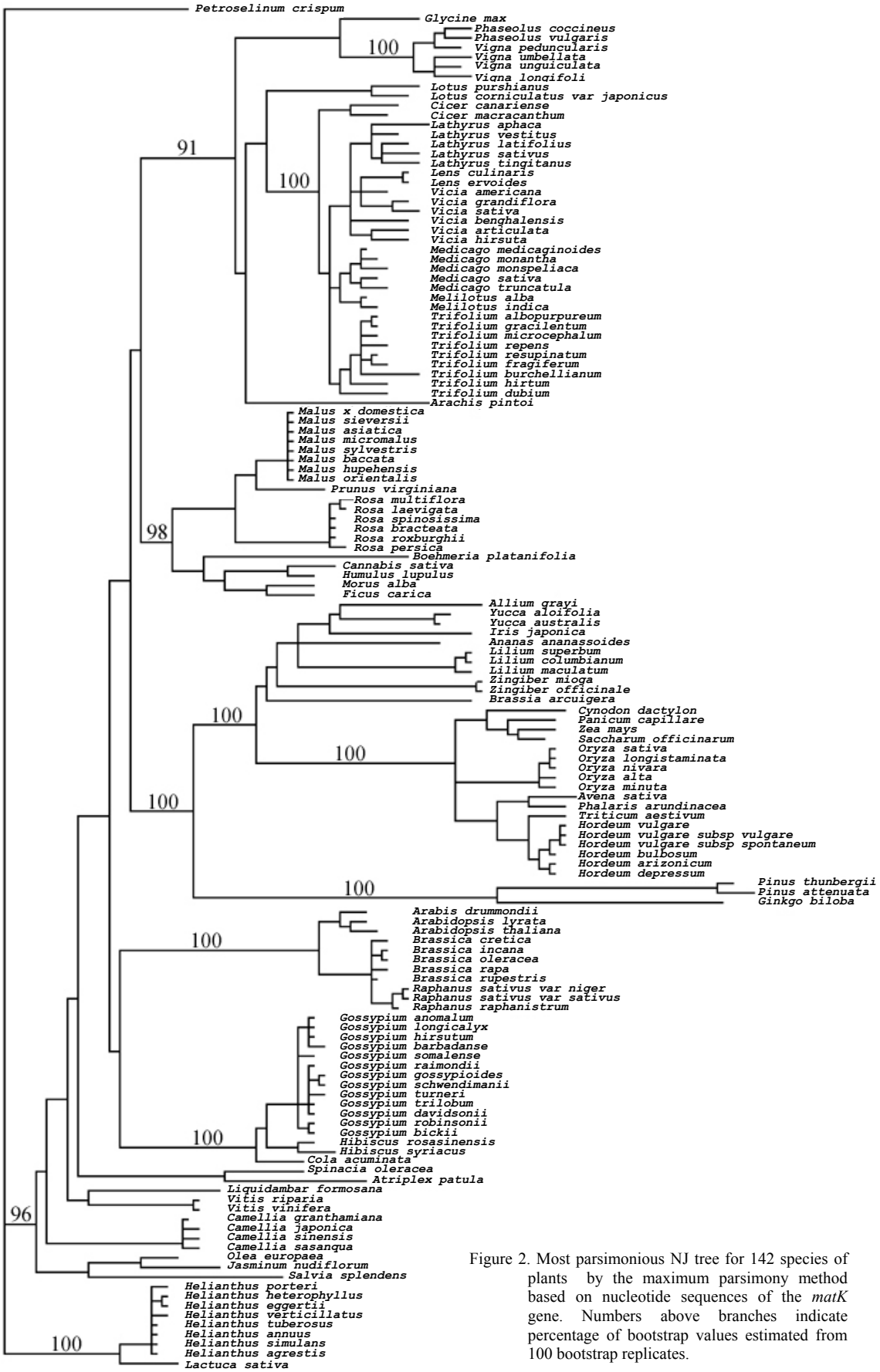

Figure 2. Most parsimonious NJ tree for 142 species of plants by the maximum parsimony method based on nucleotide sequences of the *matK* gene. Numbers above branches indicate percentage of bootstrap values estimated from 100 bootstrap replicates.

Results in this study clearly demonstrated that all the dicotyledons studied in this article were in the C_3 metabolism, differing from the monocotyledons which consisted of CAM, C_3 and C_4 metabolisms. Since the numbers of plant species studied were greater numbers in dicotyledons than monocotyledons, further studies are required to confirm this finding using a larger sample of monocotyledons and dicotyledons.

4. Conclusions

Inference of relationships from DNA sequences as well as proteins of known function to DNA or proteins of unknown function that are structurally similar can be accomplished through the comparative analysis. Using this information plant species that have not been fully sequenced can be compared on whole genome level using chloroplast genomes. This is an important aspect in the quest to decipher more the plant characteristics for ensured food security in the developing economies. In this study we showed that a taxonomically difficult group could be resolved using *matK* DNA sequences. Determined genetic relationship between the species or genera is very valuable for genetic improvement studies. The *matK* DNA sequence could also be utilized in Single Nucleotide Polymorphism (SNP) or in PCR studies. In order to utilize the *matK* we designed several primer pairs from *matK* sequences and used them in several plant species including *Vicia sativa*, *Capsicum annuum*, *Gossypium hirsutum* and some *Salvia* spp. These *matK* based primer pairs showed a high degree of polymorphisms within and between the studied plant species. These results indicated that the *matK* sequence was a valuable tool in further genetic studies.

References

Brochmann, C., Xiang, Q. Y., Brunsfeld, S. J., Soltis, D. E. and Soltis, P. S. 1998. Molecular evidence for polyploidy origins in Saxifraga (*Saxifragaceae*): the narrow arctic endemic *S. svalbardensis* and its widespread allies. Amer. J. Botany, 85: 135–143.

Crayn, D. M., Winter, K. and Smith, J. A. C. 2004. Multiple origins of crassulacean acid metabolism and the epiphytic habit in the Neotropical family *Bromeliaceae*. PNAS, 101: 3703–3708.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39: 783–791.

Fuse, S. and Tamura, M. N. 2000. A phylogenetic analysis of the plastid *matK* gene with emphasis on *Melanthiaceae* sensu lato. Plant Biol., 2: 415–427.

Ito, M., Kawamoto, A., Kita, Y., Yukawa, T. and Kurita, S. 1999. Phylogenetic relationships of *Amaryllidaceae* based on *matK* sequence data. J. Plant Res., 112: 207–216.

Kilel, B. 2004. Comparative analysis and relationships of six important crop species chloroplast genomes using whole genome web-based informatics tools. Afr. Biotech., 3: 210-214.

Koch, M., Haubold, B. and Mitchell-Olds, T. 2001. Molecular Systematics of the *Brassicaceae*: Evidence from coding plastidic *matK* and nuclear *chs* sequences. Amer. J. Botany, 88: 534–544.

Mort, M. E., Soltis, D. E., Soltis, P. S., Francisco-Ortega, J. and Santos-Guerra, A. 2001. Phylogenetic relationships and evolution of *Crassulaceae* inferred from *matK* sequence data. Amer. J. Botany, 88: 76–91.

Steele, K. P. and Vilgalys, R. 1994. Phylogenetic analysis of *Polemoniaceae* using nucleotide sequences of the plastid gene *matK*. Systematic Botany, 19: 126–142.

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland.

Tamura, M. N., Yamashita, J., Fuse, S. and Haraguchi, M. 2004. Molecular phylogeny of monocotyledons inferred from combined analysis of plastid *matK* and *rbcL* gene sequences. Journal of Plant Research, 117: 109–120.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research, 22: 4673-4680.