The application and utilization of sequence data has been found very informative in the characterization and phylogenetic relationship of different crops species. This study aimed to use bioinformatics tools to characterize the matK gene in some selected legumes with special reference to pigeon pea [cajanus cajan (L.)Millsp] matK sequence as a quarry sequence. Nucleotide and amino acid sequence of matK gene of 10 legumes were retrieved from NCBI database and analysed for homology, physiochemical properties, motifs, GC content as well as phylogenetic relationships. Results showed that the nucleotide and amino acid sequence lengths of this gene among the selected legumes differs. Its nucleotide length varied between 631-1580bp, while the amino acids sequence varied between 21 and 509 residues. P. tetragonolobus matK and C. cajan matK sequences had percentage identity of 88% while V. sativa had the lowest percentage identity of 70%. G. tomentella and P. tetragonolobus matK sequence shared the same percentage similarity of 91% with C.cajan while V. sativa had the least (78%) with C.cajan. The motif predicted were tyrosine kinase phosphorylation site, N-myristoylation site, N-glycosylation site, protein kinase phosphorylation site, casein kinase II phosphorylation site and cAMP- and –cGMP dependent protein kinase phosphorylation site. However, microbodies C-terminal targeting site was only predicted in the amino acid sequence of matK gene of P. sativum and C.cajan. Phylogenetically, two major clades were revealed with P.sativum, V.sativa, and C. arientinum matK gene sequence in clade A and matK gene sequence of P.tetragonolobus, C. cajan, G. tomentella, P.vulgaris, V.unguiculata, V. angularis and V. radiate in clade B. It showed that clade A diverged from the ancestry legume approximately 39MYA while legume sequences in clade B diverged from the ancestor about 57MYA. GC content of the nucleotide sequence of matK gene of V. sativa was highest (31.37%) with the range in the selected legume varying between 7.29%-31.37%. The secondary structure of amino acids sequence of matK gene in the selected legume revealed the alpha helix (34.14%-41.27%), extended strand (11.56%-20.99%) and random coil (39.48%-51.76%). The major domain architecture found in the amino acid sequence were single and double types. Implicitly, though maturase K gene sequences in the selected legumes differ in lengths physiochemical properties, GC content and motif. The result of this study revealed that C.cajan matK gene sequences is closely related to that of P. tetragonolobus but distant to V. unguiculata as well as P. vulgaris.

KEYWORDS: Maturase K (matK) gene, bioinformatics, phylogenetics, selected legumes, breeding

INTRODUCTION

The recent upsurge in the application and utilization of molecular/sequence data to systematic and evolutionary queries has led to significant contributions to effective classification of both plants and animals. Presently, many chloroplast, mitochondrial and nuclear genes have been utilized for studying and understanding sequence variations and evolutionary trends at the genus level (Clark et al., 1995; Hsiao et al., 1999). Before now, among the genes, sequences for the rbcl gene were frequently used and analysed by researchers in the bid to understanding plant systematics beyond the family level (Donoghue et al., 1992; Chase et al., 1993; Duval et al., 1993). However, maturase K (matk) gene, formally known as orfk has emerged as a gene of interest with potential in plant molecular systematics and evolution because of the genes’ rapid evolution at nucleotide and corresponding amino acid levels (Johnson and Soltis, 1995; Liang and Hilu, 1996; Miller et al., 2006).

Due to the rapid rates of substitution, rare presence of frame shift indels as well as few cases of premature stop codons, it has been opined that matk may not be functional in some plants (Kores et al., 2000; Whitten et al., 2000; Kugita et al., 2003; Hidulgo et al., 2004; Jankowiak et al., 2004). It has however been observed that the RNA transcripts of trnK, trnA, trnL, rps2, rpl2 and atpF require MATK for intron excision (Jenkins et al., 1997; Vogel et al., 1999). The tRNA or protein products from these genes are required for normal chloroplast function including photosynthesis,
implying that MATK has an essential function in the chloroplast, importantly as a post-transcriptional splicing factor (Michelle et al., 2007).

According to Muller et al. (2006), phylogenetic analysis of a data set composed of matK, rbcL, and trnT-F sequences from basal angiosperms demonstrated that matK contributes more parsimony informative character and significantly more phylogenetic structure on average per parsimony informative site than the highly conserved chloroplast gene rbcL. The chloroplast matK gene has two important unique features that underscore its usefulness in plant’s molecular systematics and evolution including its fast evolutionary rate. According to Johnson and Soltis (1994), Olmstead and Palmer (1994), the rate of nucleotide substitution in matK is three times higher than that of the large subunit of Rubisco (rbcL) and six fold higher than the amino acid substitution rate, which significantly presents it as a fast evolving gene. This capacity of matK gene also provides high phylogenetic signal for resolving evolutionary relationships and relatedness among plants at all taxonomic levels (Soltis and Soltis, 1998; Hili et al., 2003).

Maturase K is a chloroplast-encoding gene nested between the 5’and 3’ exons of trnK, irRNA-Lysine (Sugita et al., 1985) in the large single copy region of the chloroplast genome (Steane, 2005; Daniell et al., 2006; Turmel et al., 2006). For emphasis, maturases are enzymes that catalyse non-autocatalytic intron removal from premature RNAs. The importance attached to the leguminous family cannot be overemphasized, especially in mitigating protein deficiency in the rural population, which is more than 60% of the entire population in most sub-Saharan African countries, including Nigeria. This study aimed to use bioinformatics tools to characterize the matK gene in some selected legumes.

MATERIALS AND METHODS

Retrieval of nucleotides and amino acid sequences

The nucleotide and amino acid sequences of maturase K (matK) of C. arientinum (Chick pea), V. unguiculata (Cowpea), C. cajan (Pigeon pea), P. vulgaris (Common bean), Pisum sativum (Garden pea) and Psophocarpus tetragonolobus (winged bean), Vigna saliva (Tare vetch), Vigna radiata (mung bean) and Glycine tomentella were downloaded from the gene bank by obtaining the FASTA format from the National Centre for Biotechnology Information (NCBI, USA) database. The accession numbers of the sequences retrieved for the various legumes were noted along with the gene names, sequence length as well as the crop names were retrieved using the FASTA format option. Pair wise and multiple sequence alignments were carried out to align all retrieved sequences using MEGA 6 software as modified by Thompson et al. (2014).

Determination of percent identity and similarity (homology)

Percentage identity and similarity among the nucleotide and amino acid sequences of the retrieved maturase K (MatK) genes of the selected pulses were determined using similarity homology comparison tool for more than two sequences option of the basic alignment search tool. The nucleotide and amino acid sequence of C. cajan was used as the query sequence. This is premised on the fact that C. cajan is reported to be abiotic stress tolerant and high adaptability.

Determination of physico-chemical properties of amino acid sequences of matk gene of selected pulses

The physico-chemical properties of the MatK protein genes of the 10 leguminous species were determined using the Expert Protein Analysis System (EXPASY), (www.EXPASY.org). Protein characterization and function options, which is protparam was then selected from the tools option. The FASTA formatted amino acid sequence for sequence and physiochemical properties. Physicochemical properties of matK gene that were analysed using Expasy software are as follows Theoretical PI, molecular weight, number of amino acid residues, amino acid and atomic compositions, instability index, extinction coefficient, aliphatic index and hydropathically

Determination of predicted protein motifs and structures for MatK genes

The motifs in the amino acid sequences of MatK protein gene of the selected pulses were predicted using the protparam site (http://prosite.expasy.org/scanprosite/) FASTA formatted protein sequences were used in the scan at high sensitivity.

Prediction of secondary protein and tertiary protein structures

Prediction of motif for secondary structure was achieved using GORIV software as modified by Gamieret et al. (2015). The motif for the predicted tertiary structure (3D structure) for matk genes was obtained using the Phyre2 software (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) amino acid sequence earlier retrieved from the NCBI databases and modified by Kelley and Stemberg (2009).

Determination of start and end codons, and GC content of matK genes for legumes for the selected species

The start and end codons for the matK protein genes of the selected leguminous species (putative region) was determined using the GENSCAN software as modified by Burge (2011). The GENSCAN software was also used to determine the Guanine-Cytosine (G-C) content for each amino acid sequence of each leguminous species. http://genes.mit.edu/GENSCAN.html.

Determination of domain architecture of amino acid sequences of MatK gene in the selected legumes

The domain architecture of the amino acid sequences of MatK gene in the selected legumes was determined using the Expasy online (http://prosite.expasy.org/scanprosite/), where the amino acid of the query sequences are scanned for domain architecture.

Determination of phylogenetic and evolutionary history of matK genes

The phylogenetic analysis and evolutionary history were determined using the Molecular Evolution and Genetic Analysis (MEGA 6) software with maximum likelihood option for the construction of phylogenetic tree for the
selected legumes using their MEGA aligned retrieved nucleotide sequences from the NCBI database. The evolutionary history or pathway was traced using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) based on the Jones–Taylor-Thompson (JTT) matrix - based model. The reliability of the inferred phylogenetic tree was evaluated using the Bootstrap analysis of 1000 replications. The time of divergence or evolutionary history of MatK protein genes of the legumes was estimated based on the nucleotides percent substitution obtained per site.

RESULTS

Retrieval of nucleotide and amino acid sequences
Results obtained for sequence lengths of nucleotide and amino acid of matK gene showed that the nucleotide sequence lengths ranged from 631-1580bps while amino acids sequence lengths ranged from 21-509 residues. It was observed that nucleotide sequences of matK genes for P. tetragonolobus, G. tomentella, C. arietinum and V. sativa were the longest while P. sativum sequence was the shortest (641bps). This trend was however, observed for the amino acid sequences lengths of MatK gene of these legumes, which may have stemmed from the fact that MatK gene sequences of those legumes with longer lengths have been completely sequenced while have partial CDS.

Determination of percentage identity and similarity for amino acid and nucleotide sequences
Results on percentage identity showed that the highest identity was observed in the matK gene of P. tetragonolobus with 88% identity while the least identical species with the gene was V. sativa showing 70% identity (Table 3) taking C. cajan as the standard. On the other hand, the highest similarity in the Mat K genes was shared by G. tomentella and P. tetragonolobus (91%) The least percent similarity was however observed in V. sativa, which showed 78 percent similarity with C. cajan. Also percent identity of nucleotide sequence using C. cajan as a reference crop showed that P. tetragonolobus, G. tomentella, P. vulgaris and V. angularis had sequence identity greater than 90% while C. arietinum, V. sativa and P. sativum had percent identity with C. cajan greater than 80% (Tables 2 & 3).

Physicochemical properties
Physicochemical properties of matK protein showed that the number of amino acid residues ranged from 199-509 with G. tomentella, V. angularis, C. arietinum and P. sativum having about 500 residues of amino acids while V. unguiculata had the least 199 residues. Molecular weight for G. tomentella, V. angularis, C. arietinum and P. sativum were greater than 60000 Daltons while V. unguiculata was the lowest (24302.14Daltons). Result on theoretical PI was greater than 9.00 for all amino acid sequence of matK gene in the selected legumes. It was observed that the total –Ve and +Ve charges, total number atoms and extinction coefficient of amino acid sequence of matK gene in G. tomentella, V. angularis, C. arietinum and P. sativum were higher than other legumes investigated. However, this trend was not followed for instability index and alphatic index as these properties were high for P. tetragonolobus (80.52; 97.46), P. vulgaris, V. unguiculata, V. sativa as well as for C. cajan.

Motifs in amino acid sequence of matK gene in selected legumes
Analysis of motifs in amino acid sequence of matK gene in the 10 legumes investigated showed that there are 6 motifs revealed in G. tomentella, 6 in P. tetragonolobus, 3 in P. vulgaris, 5 in V. angularis, 6 in V. unguiculata, 5 in C. arietinum, 6 in V. radiata, 5 in V. sativa, 8 in P. sativum and 7 motifs in C. cajan. The most common motifs found in the amino acid sequences of matK gene in the selected legumes were tyrosine kinase phosphorylation site, N-myristoylation site, N-glycosylation site, casein kinase II phosphorylation site as well as cAMP- and –cGMP dependent protein kinase phosphorylation site. Conversely, microbodies C-terminal targeting site motif was observed in the amino acid sequence of matK gene of P. sativum and C. cajan (Table 5). From the table, the positions of these motifs revealed in different legumes vary though the motifs are the same.

Phylogenetic relationship and relative time of evolution of matK gene sequence in 10 legumes selected
The matK gene sequences from the 10 legumes analysed showed that there were two clades formed. Clade A had sequences of P. sativum, V. sativa and C. arietinum while clade B had P. tetragonolobus, C. cajan, G. tomentella, P. vulgaris, V. unguiculata, V. angularis as well as V. radiata. However, clade A was sub-clade into two with P. tetragonolobus, C. cajan and G. tomentella were found in sub-clade I while P. vulgaris, V. unguiculata, V. angularis and V. radiata were found in the sub-clade II (Fig. 1). The evolutionary history of matK gene sequences revealed 2 clades, which was the same as the clades in the phylogenetic tree of the gene. It showed that clade A diverged from the ancestry legume approximately 39MYA while legume sequences in clade B diverged from the ancestor about 57MYA. C. cajan and P. tetragonolobus had diverged about 35MYA probably from G. tomentella.

G-C contents and other parameters of the nucleotide sequences of matK gene in the selected 10 legumes
G-C contents analysis revealed that V. sativa had the highest G-C content (31.37), which was followed by P. sativum (30.71). It ranged from 27.29-31.37. Additionally, poly A + tail was absent in the nucleotide sequences of P. tetragonolobus, V. angularis, V. unguiculata, V. radiata as well as C. cajan. However, poly A – tail was present in all the nucleotide sequences of matK gene in the 10 legumes although they occupy varying positions. There were no initial and terminal exons as well as no peptides and coding sequences (CDS) predicted.

Secondary structure of amino acid sequences of matK gene in selected legumes
Analysis of the secondary structures of the amino acid sequences of matK gene showed that the region covered by random coil was the highest in the sequence comparing alpha helix and extended strand (Table 7). Alpha helix ranged from 31.64% - 41.27%, which for P. sativum and P. tetragonolobus while extended strand...
ranged from 11.56% (V. radiata) – 20.99% (G. tomentella). For the random coil, it ranged from 39.48% (P. tetragonolobus) – 51.76% (V. radiata).

Domain architecture of amino acid sequence of \textit{MatK gene} in selected legumes

Here we report only two types of domain architecture in the amino acid sequence of \textit{MatK gene} in the selected legumes, namely single and double domain (Table 8). However, the constraint in the above result stemmed from the fact that the coding sequences of \textit{MatK gene} in the selected legumes have not been completely sequenced, implying that those with partial CDS and having single domain architecture might have more than as reported in this paper.
Table 1: Nucleotide and amino acid sequences of Mat k gene in selected pulses

| S/N | Nucleotide sequence | Legume | Common name          | Accession number | Sequence length | Amino acid sequence | Accession number | Sequence length |
|-----|---------------------|--------|----------------------|------------------|-----------------|---------------------|------------------|-----------------|
| 1   | Cajanus cajan       | EU307315.1 | Pigeon pea           | 725              | ABZ04034.1      | 241                |                  |                 |
| 2   | Psophocarpus        | JQ669575.1 | Winged bean          | 1515             | AFC38263.1      | 504                |                  |                 |
| 3   | Glycine tomentella  | GU594697.1 | Sweet root vine      | 1518             | AF022617.1      | 505                |                  |                 |
| 4   | Phaseolus vulgaris  | EU307307.1 | Common bean          | 794              | ABZ04026.1      | 240                |                  |                 |
| 5   | Vigna angularis     | EU307332.1 | Red bean             | 770              | ABZ04051.1      | 249                |                  |                 |
| 6   | Vigna unguiculata   | EU307324.1 | Cowpea               | 747              | ABZ04043.1      | 240                |                  |                 |
| 7   | Vigna radiate       | EU307311.1 | Mung bean            | 631              | ABZ04030.1      | 199                |                  |                 |
| 8   | Cicer arientinum    | AB198876.1 | Chick pea            | 1530             | BAF46932.1      | 509                |                  |                 |
| 9   | Vicia sativa        | AF522160.1 | Tare vetch           | 1512             | AAM82152.1      | 503                |                  |                 |
| 10  | Pisum. Sativum      | EU307313.1 | Garden pea           | 641              | ABZ04032.1      | 21                 |                  |                 |

Table 2: Percent identity, similarity of amino acid sequence for mat k gene of selected Pulses

| Query Legume | E-value | % identity | % similarity | GAPS (%) |
|--------------|---------|------------|--------------|----------|
| Cajanuscajan  | 3e – 143 | 88         | 91           | 0.0      |
| Glycine tomentella | 2e – 144 | 86        | 91           | 0.0      |
| Phaseolus vulgaris | 3e – 137 | 82        | 87           | 0.0      |
| Vignaangularis   | 3e – 135 | 80        | 86           | 0.0      |
| Vignaunguiculata  | 9e – 135 | 79        | 85           | 0.0      |
| Vignaradiate     | 2e – 109 | 79        | 84           | 0.0      |
| Cicersarentinum  | 2e – 115 | 74        | 79           | 2.0      |
| Vicia sativa     | 8e – 109 | 70        | 78           | 3.0      |
| Pisum. Sativum   | 2e – 165 | 83        | 83           | 3.0      |

Table 3: Percent identity of nucleotide sequence for mat k gene of selected Pulses

| Query Legume | E-value | % identity | GAPS (%) |
|--------------|---------|------------|----------|
| Cajanuscajan  | 0.0     | 94         | 0.0      |
| Glycine tomentella | 0.0 | 93        | 0.0     |
| Phaseolus vulgaris | 0.0     | 91        | 0.0      |
| Vignaangularis   | 0.0     | 91        | 2.0      |
| Vignaunguiculata  | 0.0     | 90        | 0.0      |
| Vignaradiate     | 0.0     | 90        | 0.0      |
| Cicersarentinum  | 0.0     | 85        | 2.0      |
| Vicia sativa     | 5e – 172 | 82  | 3.0    |
| Pisum. Sativum   | 2e – 165 | 83        | 3.0      |
### Table 4: Physico-chemical properties of amino acid sequence of matk gene in selected pulses

| Pulses                     | No. of amino acid | Molecular Weight (Dalton) | Theoretical PI | Total -Ve charges | Total +Ve charges | Total No. of atom | Extinction coefficient | Instability index (II) | Alphatic index | Grand Average Hydropathicity |
|----------------------------|-------------------|--------------------------|----------------|-------------------|-------------------|-------------------|------------------------|------------------------|----------------|-------------------------------|
| Glycine tomentella         | 509               | 61199.18                 | 9.75           | 41                | 72                | 8699              | 96055                  | 50.17                  | 98.02         | -0.155                        |
| Psophocarpus tetragonolobus| 213               | 25874.87                 | 9.21           | 18                | 24                | 3654              | 40465                  | 80.52                  | 97.46         | -0.121                        |
| Phaseolus vulgaris         | 242               | 29313.01                 | 9.54           | 19                | 29                | 4160              | 46300                  | 55.94                  | 99.88         | -0.033                        |
| Vigna angularis            | 503               | 60653.82                 | 9.66           | 42                | 71                | 8618              | 93210                  | 51.15                  | 97.06         | -0.098                        |
| Vigna unguiculata          | 199               | 24302.14                 | 9.22           | 16                | 21                | 3443              | 32320                  | 48.63                  | 102.31        | 0.021                         |
| Cicer arietinatum          | 509               | 61199.18                 | 9.75           | 41                | 72                | 8699              | 96065                  | 50.17                  | 98.02         | -0.155                        |
| Vigna radiata              | 249               | 30876.85                 | 9.52           | 20                | 30                | 4377              | 49280                  | 48.25                  | 99.76         | -0.024                        |
| Vicia sativa               | 240               | 29684.49                 | 9.55           | 18                | 28                | 4210              | 47790                  | 54.69                  | 100.67        | 0.014                         |
| Pisum sativum              | 505               | 61390.78                 | 9.80           | 36                | 68                | 8732              | 99045                  | 43.86                  | 100.32        | -0.023                        |
| Cajanus cajan              | 241               | 29576.38                 | 9.01           | 19                | 24                | 4187              | 47915                  | 38.18                  | 105.10        | 0.095                         |

### Table 5: Motifs in amino acid sequence of matk gene in selected pulses

| Legumes       | Motifs                      | Position | Sequences |
|---------------|-----------------------------|----------|-----------|
| G. tomentella | (a)Tyrosine kinase phosphorylation sites | 14-20    | Rhr.Dt.L.Y |
|               | (b) N – myristoylation site  | 30-35    | GlacGH    |
|               | (c) N – glycosylation site   | 72-75    | NDSN      |
|               |                             | 227-230  | NKSs      |
|               |                             | 339-342  | NLSV      |
|               |                             | 416-419  | NISh      |
|               |                             | 416-419  | NGSA      |
|               | (d) Protein kinase (phosphorylation site) | 74-76    | SnK       |
|               |                             | 111-113  | SLR       |
|               |                             | 190-192  | TpK       |
|               |                             | 257-259  | SvK       |
|               |                             | 418-420  | SvK       |
|               |                             | 446-448  | TaR       |
|               |                             | 456-458  | SeK       |
|               | (e) Casein Kinase II Phosphorylation site | 111-114  | StrE      |
|               |                             | 257-260  | SvKD      |
|               |                             | 396-399  | SfdD      |
|               |                             | 465-468  | TeeE      |
|               | (f) cAMP- and- cGMP dependent protein kinase phosphorylation site | 420-423  | KKrs      |
|               |                             | 503-505  | NHL       |
### P. tetragonolobus

| Site Type | Sequence | Functional Domain |
|----------|----------|------------------|
| (a) Tyrosine kinase phosphorylation site | 14-20 | Khq.DILy |
| (b) N-Myristoylation site | 30-35 | GLayGH |
| (c) N – glycosylation site | 72-75 | NDSN |
| | 184-187 | NSTS |
| | 229-230 | NKSS |
| | 339-342 | NVSV |
| | 394-397 | NLSD |
| | 410-413 | NFSh |
| | 416-419 | NGST |
| (d) Protein Kinase (phosphorylation site) | 74-76 | SnK |
| | 111-113 | SIK |
| | 190-192 | TpK |
| | 257-259 | SaK |
| | 335-337 | Sik |
| | 418-420 | Stk |
| | 419-421 | TkK |
| | 446-448 | TvR |
| | 456-458 | SeK |
| (e) Casein Kinase II Phosphorylation site | 111-114 | SIKt |
| | 257-260 | SaKD |
| | 312-315 | SrpE |
| | 396-399 | SdVD |
| | 465-468 | TeeD |
| (d) cAMP – and –cGMP dependent protein phosphorylation site | 420-423 | KKkS |

### P. vulgaris

| Site Type | Sequence | Functional Domain |
|----------|----------|------------------|
| (a) Tyrosine Kinase phosphorylation site | 14-20 | RyqDLY |
| (b) N-myristoylation site | 30 – 35 | GLayGH |
| (c) cAMP – and –cGMP dependent protein kinase phosphorylation | 55-58 | KRIS |
| | 58-60 | SIR |
| | 74-76 | SnK |
| | 111-113 | SIR |
| | 117-119 | Svr |
| | 120-122 | SyK |
| | 200-202 | SkR |
| (d) N-glycosylation site | 72-75 | NDSN |
| | 227-230 | NKSS |
|                | **V. angularis**          | **V. unguiculata**          |
|----------------|---------------------------|-----------------------------|
| (a) Tyrosine kinase phosphorylation site | 14-20 Ryq.D.L.Y | 14-20 Ryq.D.I.L.Y |
| (b) N-glycosylation site | 48-51 NFSL | 30-35 GLaYAH |
|                | 72-75 NDSK | 75-78 KKnT |
|                | 227-230 NKSS | |
| (c) cAMP – and cGMP – dependent protein kinase phosphorylation site | 55-58 KRISeq | 55-58 KRISeq |
|                | 75-78 KKnT | 75-78 KKnT |
| (d) Protein kinase C phosphorylation site | 58-60 SiR | 111-113 SIR |
|                | 74-76 SKK | 165-167 SiK |
|                | 111-113 SIR | 200-202 SKR |
|                | 165-167 SiK | |
| (e) Casein Kinase II phosphorylation site | 165-168 SiKD | 165-168 SiKD |
| (f) N-glycosylation site | 72-75 NDSK | 227-230 NKSS |
| (f) Casein Kinase II Phosphorylation site | 165-168 SiKD | |
| **C. arietinum** | (a) Tyrosine kinase phosphorylation sites | 14-20 | REE.DFL.Y |
|                | (b) N – Myristoylation site              | 30-35 | GLaYSQ   |
|                | (c) cAMP – and cGMP dependent protein phosphorylation site | 38-41 | KRsS     |
|                |                                           | 421-424 | KKKs     |
|                | (d) Casein Kinase II phosphorylation site | 41-44 | SfVE     |
|                |                                           | 73-76 | SanD     |
|                |                                           | 113-116 | SLKE    |
|                |                                           | 397-400 | SdID    |
|                |                                           | 455-458 | SgsE    |
|                |                                           | 466-469 | TeeE    |
|                | (e) N – Glycosylation site               | 75-78 | NDSN     |
|                |                                           | 228-231 | NKSS    |
| **V. radiate** | (a) Tyrosine kinase phosphorylation site | 14-20 | Ryq.DiL.Y |
|                | (b) N-glycosylation site                 | 48-50 | NFSL     |
|                | (c) cAMP – and cGMP-dependent protein kinase phosphorylation site | 55-58 | KRIIS   |
|                |                                           | 75-78 | KKnT     |
|                | (d) Protein kinase C phosphorylation site | 58-60 | StR      |
|                |                                           | 74-76 | SkK      |
|                |                                           | 111-113 | SIR     |
|                |                                           | 165-167 | SIK     |
|                |                                           | 183-185 | SnR     |
|                | (e) casein kinase II phosphorylation site | 165-168 | SiKD   |
|                |                                           | 340-343 | NKSV    |
|                |                                           | 411-414 | NLSH    |
|                |                                           | 417-420 | NGSS    |
|                | (f) Protein Kinase C Phosphorylation site | 77-79 | SnK      |
|                |                                           | 113-115 | SLK     |
|                |                                           | 122-124 | SYK     |
|                |                                           | 192 – 194 | TEK   |
|                |                                           | 193 – 195 | TKK     |
|                |                                           | 265 – 267 | TLK   |
|                |                                           | 317 – 319 | TiK    |
|                |                                           | 419 – 421 | SsK   |
|                |                                           | 420 – 422 | SKK    |
|                |                                           | 447 – 449 | TVR    |
| **V. Sativa**  | (a) N – myristoylation site              | 30-35 | GLaySH   |
### SEQUENCE ANALYSIS OF MATURASE K (MATK)

|  |  |  |
|---|---|---|
| (b) N – glycosylation site | 382-387 | GQPiSK |
|   | 38-41 | NRSI |
|   | 75-78 | NDSN |
|   | 86-89 | NKSF |
|   | 337-340 | NKSV |
|   | 408-411 | NLSH |
|   | 4414-417 | NGSS |
| (c) c AMP – and cGMP – dependent protein kinase phosphorylation site | 50-53 | KKS |
|   | 224-227 | KKKS |
|   | 225-228 | KKSs |
|   | 418-421 | KKS |
| (d) Protein Kinase C phosphorylation site | 77-79 | SnK |
|   | 122-124 | SyK |
|   | 189-191 | Tik |
|   | 96-192 | TKK |
|   | 416-418 | SsK |
|   | 417-419 | SkK |
|   | 444-446 | TVR |
| (e) Casein kinase II phosphorylation site | 112-115 | SsiE |
|   | 113-116 | SSeE |
|   | 394-397 | SdFD |
|   | 452-455 | SgsE |
|   | 463-466 | TeeE |
| **P. sativum** |  |  |
| (a) Tyrosine kinase phosphorylation site | 14-20 | Rhr.Dtl.y |
| (b) N – Myristoylation site | 30-35 | GlaGH |
| (c) N-glycosylation site | 72-75 | NDSN |
|   | 227-230 | NKSS |
|   | 339-342 | NLSV |
|   | 410-413 | NFSH |
|   | 416-419 | NGSA |
| (d) Protein Kinase C Phosphorylation Site | 74-76 | SnK |
|   | 111-113 | SIR |
|   | 190-192 | TpK |
|   | 257-259 | SVK |
|   | 418-420 | SaK |
|   | 446-448 | TaR |
|   | 456-458 | SeK |
| (e) Casein Kinase II Phosphorylation | 111 – 114 | SfR |
| (f) N – glycosylation Site | 257-260 | SvKD |
|   | 396-399 | sfdD |
|   | 465-468 | TeeE |
|   | 420-423 | KKKs |
| (g) cAMP – and cGMP – dependent protein phosphorylation site |  | NHL |
| |  |  |
C. cajan
(a) Tyrosine kinase phosphorylation sites
14-20 RyqDiLY
(b) N – Myristoylation site
30-35 GLaYGH
(c) Casein Kinase II Phosphorylation
111-114 SLKE
(d) N – glycosylation site
190 – 192 TPK
252-259 SVK
418-420 SaK
446-448 TaK
456-458 SCK
111-114 StrE
257-260 SVKD
396 – 399 SdFD
465-468 TeeE
(e) cAMP – and –c6MP dependent
protein phosphorylation
420-423 KKKs
(d) Microbodies C – terminase Targeting Site
503-505 NHL

Table 6: G-C contents and other parameters of nucleotide sequence of matk gene in selected legumes

| Legumes         | G-C content (%) | Initial exons | Terminal exons | Poly A+tail | Poly A-tail | Predicted peptides | Predicted CDS |
|-----------------|-----------------|---------------|----------------|-------------|-------------|--------------------|---------------|
| G. tomentella   | 29.85           | -             | -              | 215-220     | 338-333     | -                  | -             |
| P. tetragonolobus| 27.96           | -             | -              | -           | 638-633     | -                  | -             |
| P. vulgaris     | 28.43           | -             | -              | 231-236     | 345-349     | -                  | -             |
| V. angularis    | 27.29           | -             | -              | -           | 361-356     | -                  | -             |
| V. unguiculata  | 27.99           | -             | -              | -           | 356-351     | -                  | -             |
| C. arienatum    | 30.62           | -             | -              | 745-750     | 1488-1483   | -                  | -             |
| V. radiate      | 28.02           | -             | -              | -           | 635-630     | -                  | -             |
| V. sativa       | 31.37           | -             | -              | 307-312     | 322-317     | -                  | -             |
| P. sativum      | 30.71           | -             | -              | 212-217     | 594-589     | -                  | -             |
| C. cajan        | 27.84           | -             | -              | -           | 323-318     | -                  | -             |

Table 7: Secondary structure of amino acids sequences of matk genes in selected legumes

| Legumes         | Alpha helix (%) | Extended strand (%) | Random coil (%) |
|-----------------|-----------------|---------------------|-----------------|
| G. tomentella   | 35.64           | 20.99               | 43.37           |
| P. tetragonolobus| 41.27           | 19.25               | 39.48           |
| P. vulgaris     | 34.58           | 17.92               | 47.80           |
| V. angularis    | 34.14           | 18.88               | 46.99           |
| V. unguiculata  | 37.50           | 15.83               | 46.67           |
| C. arienatum    | 35.76           | 18.66               | 45.58           |
| V. radiate      | 36.68           | 11.56               | 51.76           |
| V. sativa       | 37.57           | 15.71               | 46.72           |
| P. sativum      | 31.64           | 19.25               | 49.80           |
| C. cajan        | 39.42           | 19.25               | 46.46           |

Table 8: Domain architecture of amino acid sequence of some selected legumes
| S/N | Legume species | Domain type     | Complete/partial CDS | Start   | End     | E-value |
|-----|----------------|-----------------|----------------------|---------|---------|---------|
| 1   | C. cajan       | Double domain   | Partial CDS          | 101; 171| 112; 182| N/A     |
| 2   | P. tetragonolobus | Single Domain   | Complete CDS         | 170     | 182     | N/A     |
| 3   | G. tomentella  | Double domain   | Complete CDS         | 101; 456| 112; 475| N/A     |
| 4   | P. vulgaris    | Single Domain   | Partial CDS          | 101     | 112     | N/A     |
| 5   | V. angularis   | Single Domain   | Partial CDS          | 102     | 112     | N/A     |
| 6   | V. unguiculata | Single Domain   | Partial CDS          | 102     | 112     | N/A     |
| 7   | V. radiate     | Single Domain   | Partial CDS          | 102     | 112     | N/A     |
| 8   | C. arientinum  | Single Domain   | Complete CDS         | 457     | 476     | N/A     |
| 9   | V. sativa      | Double domain   | Complete CDS         | 186; 455| 200; 467| N/A     |
| 10  | P. sativum     | Double domain   | Partial CDS          | 171; 187| 182; 201| N/A     |
DISCUSSION

Data mined from sequenced genes have been very pivotal in molecular systematic studies. Importantly, analyses of the DNA sequences of various species provide valuable information about their taxonomy, gene make up as well as utilizations. Undoubtedly, genomic regions vary considerably in their potential phylogenetic informativeness and their contributions in resolving a given set of taxa over specified time (Hilu et al., 2014). Specifically, there are two schools of thought regarding the utilization of rapidly evolving regions as against slowly evolving regions of the genome. According to Graham and Olmstead (2000), Wang et al. (2009), Brockington et al. (2009), rapidly evolving regions will be better used for shallow evolutionary histories while slowly evolving regions for deeper epochs. Their argument was premised on the fact that multiple hits confounded by extended time scale could be significant enough to conceal phylogenetic signals and elevate homoplasy, with saturation reaching levels that can negatively impact tree structure (Graybeal, 1994; Wenzel and Siddal, 1999; Klopfstein et al., 2010; Townsend et al., 2012). The accumulation of multiple hits in rapidly evolving regions is capable of obscuring potential synapomorphies as well as results in long branch alteration (Townsend, 2007; Magallon and Sanderson, 2002). On the contrary, however, the opposing school of thought opined that rapidly evolving regions promotes effectiveness and less constrained genomic regions in deep level phylogenetics (Yang, 1998; Hilu et al., 2008; Hilu and Liang, 1997; Hilu et al., 2003; Muller et al., 2006; Worberg et al., 2007). According to Hilu et al. (2014), phylogenetic signal from rapidly evolving and un-constrained matK provides by far the most structure and accuracy, whereas slowly evolving, constrained and un-constrained genes display decreasing degrees of informativeness and tree structure. This was also the same position that Muller et al. (2006) had earlier posited that matK gene is very informative in plant systematics owing to its high phylogenetic signal when compared with other genes such as rbcL.

We report that nucleotide sequence length for *P. tetragonolobus*, *G. tomentella*, *C. arientinum* and *V. sativa* had similar sequence lengths (>1500bps) while *C. cajan*, *P. vulgaris*, and *V. unguiculata* also had similar lengths (>700bps). The same trend was observed for the amino acid sequence lengths for afore-mentioned legumes. It should also be mentioned here that the sequences of the later were only partially sequenced. However, it has been observed that variations within a family of related nucleic acids and protein sequences provide an invaluable source of information for evolution. Variations in sequence lengths in different organisms have been attributed to indels mutations that have accumulated during evolution. What this might suggest is that legumes with similar nucleotide and amino acid sequence lengths probably may have evolved at the same time or differentiated/diverged from their ancestry root almost the same time. The other likelihood is the fact that though they are legumes, their genus are not the same, which might not be unconnected with the earlier indel mutations creating evolution divergences as well as variations in sequence lengths of nucleotides and amino acids.

According to Stone et al. (2010), organisms with high percentage sequence similarity in their genes have
a similar pattern of evolution and differentiation. Sequence similarity implies that the two sequences share a common evolutionary ancestor otherwise known as homologs but should be noted that homologous sequence do not always or necessarily share significant sequence similarity. From our result, *P. tetragonolobus* and *G. tomentella* share more than 90% amino acid sequence similarity with *C. cajan* implying close relatedness. *P. vulgaris*, *V. angularis*, *V. unguiculata*, *V. radiata* and *C. arietinum* share more than 80% similarity while *V. sativa* and *P. sativum* share more than 70% sequence similarity with *C. cajan*. According to Kajita et al. (2001), if two sequences have sequence identity greater than 70%, the implication is that they have about 90% probability or more to share the same biological processes and functions. WE report nucleotide and amino acid sequence identity greater than 70% except for *P. sativum* sequence that had 70%.

This notwithstanding, what it might suggest is that matK gene found in the legumes analysed may perform similar functions and undergo the same processes. It may be recalled that Kores et al. (2000), Kujita et al. (2003), Jankowiak et al. (2004) had earlier feared that matK gene may not be functional in some plants due to rare presence of indels as well as premature stop codons. This was countered by Michelle et al. (2007) who observed that matK is involved in post-transcriptional splicing in the chloroplast. However, what this present analysis cannot infer is whether though having high percentage identity, which should have implied similar functionality, is their levels of functionality.

Protein in the same family share at least more than 30% amino acid sequence similarity with the resultant sharing of some structural characteristics (Wojciechowski et al., 2004). It thus suggest that matK gene in the respective legumes share very structural features owing to the high percentage similarity in their amino acid sequences, their genus notwithstanding. Sequence similarity off approximately 70% may suggest identical homology, functionality and very high conservation in matK gene.

The expected value (E-value) assess the significance of single pair wise alignment, which is related to the p-value. The lower the E-value, the less likely the database match is a result of random chance and thus the more significant the match is. Interestingly, E-value less than 1e-50 (E<1e-50) indicates that the match was as a result of homologous relationships. It might therefore be wise to affirm that the nucleotide and amino acid sequence identity and similarity were homologs and as such indicate strong relationship evolutionarily.

Our result on physicochemical properties of amino acid sequences of matK gene in the respective legumes showed that the higher the number of amino acids residues, the weightier; higher positively and negatively charges more number of atoms as well as extinction coefficient. Positively charged residues were greater than negatively charged residues, which implies that maturase K protein is an extracellular protein instead of an intracellular protein (Andrade et al., 1998). Gururaprasad et al. (1990) observed that instability index more than 40 implies that the matK protein is unstable in vitro. Except matK amino acid sequence of *C. cajan* that had instability index of 38.18, other matK gene of other legumes analysed had instability index more than 40. Nikhil et al. (2009) reported that the instability index is a function of the abundance of cysteine in the formation of disulphide bond in the matK protein molecule. From this report, it thus mean that excepting of matK protein of *C. cajan*, other matK proteins have low cysteine for disulphide bond formation. Proteins could either be hydrophobic or hydrophilic. In the report of kyle and Doolittle (1982), grand average hydrophaticity (GRAVY) value greater than zero indicates a relatively hydrophobic protein. Our present report suggests that matK protein is relatively hydrophobic (-0.155-0.021).

Sequence motifs re short recurring patterns in the DNA that are presumed to possess a biological functions. Usually, they indicate sequence-specific binding sites for protein in the form of enzymes (Nucleases, transcription factors, etc.). The fact that almost similar motifs were predicted on the amino acid sequence of matK gene in the 10 legumes analysed except microbodies c-terminal targeting site found in the matK gene of *P. sativum* and *C. cajan*, the differences in the positions they occupy indicate that the sequence-specific binding sites for proteins differ and might however, create functionality differences. It should be underscored that some of these positions that specifically bind to these motifs are involved in varying important processes at the RNA level including ribosome binding, mRNA processing, termination of transcription, etc. These proteins have varying initiation and termination sites.

Phylogenetically, we observed that matK gene of the different legumes were clade based on percentage identity and similarity of the sequences. The implication is that the more the sequence homology, the more probability of them to be clade together. This was evidenced on the sub-clade A of clade 1 comprising of *G.tomentella*, *C. cajan* and *P. tetragonolobus* (Figure 1a) and percentage identity and similarity of nucleotide and amino acid sequences (Tables 2 & 3) of matK gene. What this might suggest is that matK gene though coming from a common ancestor, diverge evolutionarily probably due to indel mutations. This gave rise to sequence homology and possible similar functionality and structurally characteristics. Though these different legumes fall into different genus, their gene sequences showed high homology thus being clustered together. According to Wojciechowski et al. (2004) reported that fabaceae is generally monophyletic implying that it contains clade containing an ancestral species and all its descendants. However, some are paraphyletic. Using *rbcL* gene (Kajita et al., 2001) and trnL intron (Bruneau et al., 2001; Herendeen et al., 2003) sequences, analyses of matK sequences support the monophyly of the leguminous family. What it portend is that the sequences of these genes used are might be highly homologous. Though our present result could not trace the ancestral parent, it may not negated the earlier positions on monophyletic concept.

Bruneau et al. (2008) reported that Fabaceae started their diversification approximately 60MYA while the most important clades diverged some 50MYA. Lavin et al. (2005) and Bruneau et al. (2008) reported that the age of the main cesalpinioideae clades have been estimated as being between 56 and 34MYA while the basal group of the mimosoideae was put as 44±2.6MYA. Using the matK gene sequences from the various
Additional, our result on secondary structure of the role in the evolution of plants. Base composition is a changing GC content might have played a significant is the fact that genomic adaptations associated with different legumes analysed. According to Smarda et al. (2014), the hypotheses by several authors as regards the biological impact of GC content variation in microbial and vertebrate organisms notwithstanding, the biological significance of GC content diversity in plants remains unclear due to lack of sufficiently robust genomic data. GC content showed a quadratic relationship with genome size, with the deceases in GC content in larger genomes possibly being a consequences of the higher biochemical cost of GC base synthesis. GC-rich DNA aids cell freezing and desiccation. Important to mention is the fact that genomic adaptations associated with changing GC content might have played a significant role in the evolution of plants. Base composition is a fundamental property of genomes and a strong influence on gene function and regulation (Li and Du, 2014). In higher organisms, the GC content was lower in dicot plants and highest in monocot plants (Li and Du, 2014). GC content of monocots varied between 33.6-48.9% (Smarda et al., 2014). Analysis of the GC content of the nucleotide sequences of matK gene in the selected legumes ranged from 27.29% to 30.71%. This confirms the earlier report of Li and Du (2014). Poly (A) tail is a common modification of eukaryotic mRNA and plays many fundamental roles in mRNA stability (Mangus et al., 2003; Coller and Parker, 2004). It is a long chain of adenine nucleotides that is added to the 3' end mRNA during RNA process as it increases the stability of the molecule.

The secondary structure of amino acid sequences of matK gene in selected legumes revealed that alpha helix, extended or beta strand as well as random coil. Usually, a region of secondary structure that is not an alpha helix, β-sheet, or a recognisable turn is commonly known as a coil (Mount, 2004). From the result alpha helix is the most abundant helical conformation found in globular proteins accounting for 32-38% of all the residues. Regions richer in alanine (A), glutamic acid (E), leucine (L) and methionine (M) and poorer in proline (P), glycine (G), tyrosine (Y) and serine (S) (AELM>PGYS) tend to form an alpha helix. This might mean that amino acid sequences of matK gene in the legumes were higher in AELM and lower in PGYS. Additionally, our result on secondary structure of the amino acid sequences of matK gene in the selected legumes suggests that the percentage unrecognisable regions (random coil) was higher than each of recognisable regions (alpha helix and extended strand).

Protein domains are the structural, functional, evolutionary units of the protein (Zhang et al., 2012). Usually, proteins with the similar architectures are close homologs, while different proteins possess distinct domain architectures. The implication might mean that those sequences composed of more than one single domain may have been invented by rearrangement, duplication, insertion, deletion, fusion and fission of domains (Teichmann et al., 1998; Gough, 2005; Kummerfeld and Teichmann, 2005; Fong et al., 2007; Ekman et al., 2007) This is premised on the fact simple domain architectures per protein are more often than not created de novo (Fong et al., 2007). It should be noted that the fact that some matK gene sequences of selected legumes are partially sequenced might conceal some information that may be important in resolving the differences in the legume sequences.

CONCLUSION

Expectedly, there are differences observed in the matK gene sequences in the selected legumes considering the parameters analysed. However, our results revealed some degree of identity and similarity in the sequences especially between C. cajan and P. tetragonolobus sequences. This might implicitly mean same functions in these legumes.

REFERENCES

Andrade, M. A., Donoghue, S. I and Rost, B., 1998. Adaptation of protein surfaces to sub- cellular location. Journal of Molecular Biology, 276: 517-525

Brockington, S. F., Alexandre, R, Ramdial, J., Moore Michael, J and Crawley, S, et al. 2009. Phylogeny of the Caryophyllales Sensu Lato: ’Revisiting hypotheses on pollination biology and perianth differentiation in the core Caryophyllales. International Journal of Plant Sciences 170: 627–643.

Bruneau, A., Lewis, G. P., Herendsen, P. S., Schrire, B and Mercure, M., 2008. Biographic patterns in early diverging clades of the leguminosae: In: Botany 2008. Botany without Borders Botanical Society of America. Boston.

Bruneau, A., Mercure, M., Lewis, G. P and Herendeen, P. S., 2008. Phylogenetic patterns and diversification in the caesalpinoidi legumes. Canadian Journal of Botany, 86(7): 697-718.

Chase, M. W., Solitis, D. E and Olmstead, R. G., 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Gardens 80:528-580.

Clark, L. G., Zhang, W and Wendel, J. F., 1995. A phylogeny of the grass family (Poaceae) based on ndhF sequence data. Systematic Botany. 20: 436-460.

Daniell, H., Lee, S. B., Grevich, J., Saski, C., Quesada-vargas, T., Guda, C. B., Tomkins, J and Jansen, R. K., 2006. Complete chloroplast genome sequences of Solanumbulbocastanum, Solanumlycopersicum and comparative analyses with other Solanaceae genomes. Theoretical and Applied Genetics 112: 1503–1518.
Donoghue, M. J., Olmstead, R. G., SMITH, J. F and Palmer, J. D., 1992. Phylogenetic relationships of Dipsacales based on rbcL sequences. Annals of the Missouri Botanic Garden 79: 333-345.

Duvall, M. R., Clegg, M. T., Chase, M. W., Clark, W. D., John Kress, W., Harold G. Hills., Eguiarte, L. E., Smith, J. F., Gaut, B. S., Zimmer, E. A and Learnt, G. H., 1993. Phylogenetic hypotheses for the monocotyledons constructed from RBCL sequence data. Annals of the Missouri Botanic Garden 80: 607-619.

Ekman, D., Bjorklund, A. K and Elofsson, A., 2007. Quantification of the elevated rate of domain rearrangement in metazoa. J Mol Biol 372:1337-1348.

Fong, J. H., Geer, L. Y., Panchenko, A. R and Bryant, S. H., 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol, 366:307-315.

Goldman, N., 1998. Phylogenetic information and experimental design in molecular systematics. Proceedings of the Royal Society of London Series B: Biological Sciences 265: 1779–1786.

Gough, J., 2005. Convergent evolution of domain architectures (is rare). Bioinformatics, 21:1464-1471.

Graybeal, A., 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Systematic Biology 43: 174–193.

Guruprasad, K., Reddy, B. V and Pandit, M. W., 1990. Correlation between stability of a protein and its dipeptide composition. Protein Engineering, 9: 849-856.

Hidalgo, O., Garnatje, T., Susanna, A and Mathez, J., 2004. Phylogeny of Valerianaceae based on matK and ITS markers, with reference to matK individual polymorphism. Annals of Botany 93: 283–293.

Hilu, K. W and Liang, H., 1997. The matK gene: Sequence variation and application in plant systematics. American Journal of Botany 84: 830–839.

Hilu, K. W., Black, C. M and Oza, D., 2014. Impact of Gene Molecular Evolution on Phylogenetic Reconstruction: A Case Study in the Rosids (Superorder Rosanae, Angiosperms). PLoS ONE 9(6): e99725. doi:10.1371/journal.pone.0099725.

Hilu, K. W., Black, C., Diouf, D and Burleigh, J. G., 2008. Phylogenetic signal in matK vs. trnK: A case study in early diverging eudicots (angiosperms). Molecular Phylogenetics and Evolution 48: 1120–1130.

Hilu, K. W., Borsch, T., Muller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., Chase, M. W., Powell, M. P., Alice, I. A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T. A. B., Jens, G. R., Campbell, C. S and Chatrou, I. W., 2003. Angiosperm phylogeny based on matK sequence information. American Journal of Botany 90: 1758–1776.

Hilu, K. W., Borsch, T., Muller, K., Soltis, D. E and Soltis, P. S., 2003. Angiosperm phylogeny based on matK sequence information. American Journal of Botany 90: 1758–1776. 25.

Hsiao, C., Chatterton, N. J., Asay, K. H and Jensen, K. B., 1994. Phylogenetic relationships of 10 grass species: an assessment of phylogenetic utility of the internal transcribed spacer region in nuclear ribosomal DNA in monocots. Genome 37: 112-120

Jankowiak, K., Lesicka, J., Pacak, A., Rybarczyk, A and Szewykowska-kulin’ska, Z., 2004. A comparison of group II introns of plastid tRNALysUUU genes encoding maturase protein. Cellular & Molecular Biology Letters 9: 239–251.

Jenkins, B. D., Kulhanek, D. J and Barkan, A., 1997. Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classes with distinct requirements for splicing factors. Plant Cell 9: 283–296.

Johnson, L. A and Soltis, D. E., 1994. matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. Systematic Botany 19: 143-156.

Kajita, T., Ohashi, H., Tateishi, Y., Bailey, C. D and Doyle, J. J., 2001. RbcL and legume phylogeny with particular reference to Phaseoeae.Millettiae and allies. Systematic Botany, 26: 515-536.

Klopfstein, S., Kropf, C and Quicke, D. L. J., 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Systematic Biology 59: 226–241.

Klopfstein, S., Kropf, C and Quicke, D. L. J., 2010. An evaluation of phylogenetic informativeness profiles and the molecular phylogeny of Diplazontinae (Hymenoptera, Ichneumonidae). Systematic Biology 59: 226–241.

Kores, P. J., Weston, P. H., Molvray, M and Chase, M. W., 2000. Phylogenetic relationships within the Diurideae (Orchidaceae): inferences from plastid matK DNA sequences. In K. L. Wilson and D. A. Morrison [eds.], Monocots: systematics and evolution, 449– 456. CSIRO Publishing, Collingwood, Victoria, Australia

Kugita, M., Kaneko, A., Yamamoto, Y., Takeya, Y., Matsumoto, T and Yoshinaga, K., 2003. The complete nucleotide sequence of the hornwort (Anthocerosformosae) chloroplast genome: insight into the land plants. Nucleic Acids Research 31: 716–721.

Kummerfeld, S and Teichmann, S. A., 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet., 21:25-30.
Kyle, J and Doolittle, R., 1982. A simple method for displaying the hydrophatic of a protein. Journal of Molecular Biology, 157: 105-132.

Lavin, M., Herendeen, P. S and Wojciechowski, M. F., 2005. Evolutionary rates analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. Systematic Botany, 54(4): 575-594.

Li, X-Q and Du, D., 2014. Variation, evolution and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PLoS ONE 9(2):e88339. Doi:10.1371/journal.pone.0088339.

Liang, H and Hilu, K. W., 1996. Application of the matK gene sequences to grass systematics. Canadian Journal of Botany 74: 125–134.

Magallon, S and Sanderson, M. J., 2002. Relationships among seed plants inferred from highly conserved genes: Sorting conflicting phylogenetic signals among ancient lineages. American Journal of Botany 89: 1991–2006.

Michelle, M. B and Hilu, K. W., 2007. Expression of matk: Functional and evolutionary implications. American Journal of Botany 94(8): 1402–1412.

Mount, D. M., 2004. Bioinformatics Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press.

Muller, K. F., Borsch, T and Hilu, K. W., 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: Contrasting matK, trnT-F, and rbcL in basal angiosperms. Molecular Phylogenetics and Evolution 41: 99–117.

Muller, K. F., Borsch, T and Hilu, K. W., 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting matK, trnT-F and rbcL in basal angiosperms. Molecular Phylogenetics and Evolution 41: 99–117.

Nikhil, S., Rekha, K., Sodhi, J. S and Bhalla, T. C., 2009. In silico analysis of amino acids sequence in relation to specificity and physicochemical properties of some microbial nitrilases. Journal of Proteomics and Bioinformatics, 2(4): 185-192.

Olmstead, R. G and Palmer, J. D., 1994. Chloroplast DNA systematics: a review of methods and data analysis. American Journal of Botany 81: 1205–1224.

Smarda, P., Bures, P., Horova, L., Leitch, I. J., Muccina, L., Pacini, E. Tichy, L., Grulich, V and Rotreklova, O., 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci., 111(39): E4096-102.

Soltis, D. E and Soltis, P. S., 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. Pp. 2-31 in P. S. S. Douglas E. Soltis, and J. J. Doyle., ed. Molecular Systematics of Plants II: DNA Sequencing. Kluwer Academic Publishers,

Steane, D. A., 2005. Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, Eucalyptus globulus (Myrtaceae). DNA Research 12: 215–220.

Stone, A. C., Battistuzzi, F. U., Kubatko, L. S., Perry, G. H., Trudeau, E., Lin, H and Kumar, S., 2010. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure.US National Library of medicine, 365(1556):3277-88.

Sugita, M., Shinozaki, K and Sugiyama, M., 1985. Tobacco chloroplast tRNA(Lys) (UUU) gene contains a 2.5-kilobase-pair intron: an open reading frame and a conserved boundary sequence in the intron. Proceedings of the National Academy of Sciences, USA 82: 3557–3561.

Teichmann, S. A., Park, J and Chothia, C., 1998. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. Proc Natl Acad Sci USA, 95:14658-14663.

Townsend, J. P., Su. Z and Tekle, Y. I., 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. SystBiol 61: 835–849.

Townsend, JP., 2007. Profiling phylogenetic informativeness. Systematic Biology 56: 222–231.

Turmel, M., Otis, C and Lemieux, C., 2006. The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. Molecular Biology and Evolution 23: 1324–1338.

Vogel, J., Borner, T and Hess, W., 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Research 27: 3866-3874

Wang, H., Moore, M. J., Soltis, P. S., Bell, C. D and Brockington, S. F., 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proceedings of the National Academy of Sciences 106: 3853–3858.

Wenzel, J. W and Siddall, M. E., 1999. Noise. Cladistics 15: 51–64.

Whitten, W. M., Williams, N. H and Chase, M. W., 2000. Subtribal and generic relationships of Maxillarieae (Orchidaceae) with emphasis on Stanhopeineae: combined molecular evidence. American Journal of Botany 87: 1842–1856.
Wojciechowski, M. F., Lavin, M and Sanderson, M. J., 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well supported sub-clades within the family. American Journal of Botany, 91(11): 1846-1862.

Worberg, A., Quandt, D., Barniske, A. M., Lohne, C and Hilu, K. W., 2007. Phylogeny of basal eudicots: insights from non-coding and rapidly evolving DNA. Org Divers Evol 7: 55–77.

Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. Systematic Biology 47: 125–133.