Research Article
Araştırma Makalesi

Bartın University International Journal of Natural and Applied Sciences
JONAS, 2(2): 195-206 e-ISSN: 2667-5048
31 Aralık/December, 2019

CONTRIBUTIONS TO THE SOLUTION OF PHYLOGENETIC PROBLEM IN FABALES

Deniz Aygören Uluer[1*], Rahma Alshamrani [2]
[1] Ahi Evran University, Cicekdagi Vocational College, Department of Plant and Animal Production, 40700 Cicekdagi, KIRŞEHIR
[2] King Abdulaziz University, Department of Biological Sciences, 21589, JEDDAH

Abstract

Fabales is a cosmopolitan angiosperm order which consists of four families, Leguminosae (Fabaceae), Polygalaceae, Surianaceae and Quillajaceae. The monophyly of the order is supported strongly by several studies, although interfamilial relationships are still poorly resolved and vary between studies; a situation common in higher level phylogenetic studies of ancient, rapid radiations. In this study, we carried out simulation analyses with previously published *mat*K and *rbc*L regions. The results of our simulation analyses have shown that Fabales phylogeny can be solved and the 5,000 bp fast-evolving data type may be sufficient to resolve the Fabales phylogeny question. In our simulation analyses, while support increased as the sequence length did (up until a certain point), resolution showed mixed results. Interestingly, the accuracy of the phylogenetic trees did not improve with the increase in sequence length. Therefore, this study sounds a note of caution, with respect to interpreting the results of the "more data" approach, because the results have shown that large datasets can easily support an arbitrary root of Fabales.

Keywords: Data type, Fabales, phylogeny, sequence length, simulation.

1. Introduction

Fabales Bromhead is a cosmopolitan angiosperm order which consists of four families, Leguminosae (Fabaceae) Juss., Polygalaceae Hoffmanns. & Link, Surianaceae Arn. and Quillajaceae D. Don (APG, IV). Morphological characters supporting a close relationship of Polygalaceae, Surianaceae and *Quillaja* Molina to Leguminosae are not very extensive. For example, Leguminosae species are characterized by several clear wood anatomical features, but share relatively little with the wood anatomy characteristics of Polygalaceae, Surianaceae and Quillajaceae (Baas et al., 2000). Similarly, within Fabales, nodulation occurs only in Leguminosae specifically in most Papilionoideae, some Caesalpinioideae including most mimosoids which was originated several times independently (Soltis et al., 1995; Cannon et al., 2014), but are not found in the other three families. Unlike the polysymmetric flowers in Surianaceae and Quillajaceae, monosymmetric keel flowers are found only in two of the families within Fabales, more specifically in two tribes of Polygalaceae (Polygaleae Chodat and Xanthophylleae Chodat) and subfamily Papilionoideae DC. (Leguminosae) (Bello et al., 2010). However, this character has been accepted as an adaptation or convergent character rather than a synapomorphic character (Westerkamp, 1997), because of the developmental differences and different evolutionary modules of the keeled flowers in these two families (Bello et al., 2010; Bello et al., 2012).

Leguminosae and Polygalaceae are the largest families in the order, (Persson, 2001; Lewis et al., 2005; Lewis et al., 2013), while Quillajaceae and Surianaceae are the two species-poor families. Leguminosae is the third largest of all flowering plant families with ca. 19,500 species in 766 genera worldwide (Lewis et al., 2005; Lewis et al., 2013; LPWG, 2017). The Leguminosae have characteristic alternate, opposite, whorled, spiral or distichous leaves; the inflorescences are panicles, racemes, fascicles, spikes, heads or flowers may be solitary; flowers are predominantly actinomorphic to zygomorphic and fruits are predominantly legumes (Watson & Dallwitz, 1992). The legumes include many important species used as foods or for other purposes such as soybean, peanuts, lentils, alfalfa and clover and

***Sorumlu Yazar (Corresponding Author):**
Deniz AYGÖREN ULUER (Dr.): Ahi Evran University, Cicekdagi Vocational College, Department of Plant and Animal Production, 40700 Cicekdagi, Kirşehir, Turkey.
Work phone: +903862805500, Fax: +903862805528
E-mail: d.aygoren@ahievran.edu.tr.

Geliş (Received) : 15.08.2019
Kabul (Accepted) : 18.11.2019
Basım (Published) : 31.12.2019

Leguminosae is the second economically important family after Poaceae Barnhart (Lavin et al., 2005; LPWG, 2017). The family is also ecologically important. In African forests and the Neotropics, Leguminosae are the most species-rich family (Wang et al., 2009). Furthermore, this family contributes nitrogen-fixation via symbiotic bacteria (nodulation). These attributes of the family have led to widespread interest in its evolution and classification. According to new phylogenetic classification, the family consists of six subfamilies, Papilionoideae (501 genera, ca. 14,000 species; the great majority of them characterized by bilateral "papilionoid" flowers or "keeled" flowers *sensu* Westerkamp, 1997), Caesalpinioideae DC. (150 genera, ca. 4,400 species), Cercidoideae LPWG (13 genera, ca. 335 species), Detarioideae Burmeist. (84 genera, ca. 760 species), Dialioideae LPWG (17 genera, ca. 85 species) and Duparquetioideae LPWG (one genus, one species) (LPWG, 2017).

Polygalaceae is the second largest family in the order with a nearly cosmopolitan distribution (absent only from New Zealand, many southern Pacific islands, Antarctica and the Arctic) and ca. 1,000 species in 20 genera of herbs (e.g., *Epirixanthes* Blume, *Polygala* L.), shrubs (e.g., *Muraltia* DC.), lianas (e.g., *Securidaca* L.) and trees (e.g., *Xanthophyllum* Roxb.) (Persson, 2001). Plants in this family have characteristic alternate, opposite or whorled leaves, inflorescences are racemes or panicles (rarely flowers are solitary); flowers are actinomorphic to zygomorphic, and the fruits are capsules, samaras, drupes or berries (Eriksen & Persson, 2007). The first interfamilial subdivision of the family was carried by Chodat (1896) and he defined three tribes: Polygaleae, Moutabeae Chodat and Xanthophylleae, but later genera *Atroxima* Stapf and *Carpolobia* G. Don were placed in a new tribe, Carpolobieae Eriksen (Eriksen, 1993). Now, Polygalaceae is classified into four tribes: Carpolobieae, Moutabeae, Polygaleae and Xanthophylleae. However, aside from *Polygala* with ca. 500 species which accounts for around half the species in the family (Eriksen & Persson, 2007); *Monnina* Ruiz & Pav., *Muraltia*, *Securidaca* and *Xanthophyllum* are other species rich genera.

The three species of monogeneric Quillajaceae are distributed only in South America. These species are trees with alternate, simple leaves, cymose inflorescences, 5-merous-regular flowers and fruits that are follicles (Watson & Dallwitz, 1992; Kubitzki, 2007). Surianaceae has seven species in four genera, *Stylobasium* Desf., *Guilfoylia* F. Muell., *Cadellia* F. Muell. and *Recchia* Sessé & Moc. ex DC., and aside from one species with a pantropical distribution, an unusual distribution in Australia and Mexico (Mabberley, 1997). The Surianaceae are trees or shrubs with alternate, simple or compound leaves; the inflorescences are panicles or cymes, the flowers are regular and the fruits are berries, drupes or nuttlets (Watson & Dallwitz, 1992; Schneider, 2007).

Despite the great interest of botanists, a convincing phylogeny of the order is still not available. The monophyly of Fabales is supported strongly by several studies, although interfamilial relationships are still poorly resolved and vary between studies (e.g., Forest et al. 2007; Bello et al. 2009), a situation common in ancient, rapid radiations (Bello et al., 2009). Indeed, ancient rapid radiations have been one of the hardest problems for phylogenetic studies to resolve due to short internal branches which show a limited time span between speciation events, have weak phylogenetic signal compared to long external branches and a spurious root problem (Smith, 1994; Whitfield & Lockhart, 2007), and this type of problematic phylogenies were reported for many angiosperm clades such as Mesangiospermae (Zeng et al., 2014), eudicots (Moore et al., 2010), basal Leguminosae (LPWG, 2017) and Brassicaceae Burnett (Huang et al., 2015). Similarly, despite sampling more than 25,000 base pairs (bp) of sequence data, the phylogeny of rosids which includes order Fabales has also been problematic (Jansen et al., 2006; Wang et al., 2009). The phylogeny of the largest family of Fabales, Leguminosae, but particularly early diverging clades of the family, has also received weak and controversial results from every study (LPWG, 2017). While the "more data" approach has been seen as the ultimate solution for most phylogeny problems, yet large datasets can yield robust but inaccurate phylogenies (Jeffroy et al., 2006).

Even their main focus was not Fabales, previous studies recovering different topologies for Fabales have used different gene regions and have very different and unbalanced sampling (e.g. Savolainen et al., 2000; Kajita et al., 2001; Persson, 2001; Wojciechowski et al., 2004; Lavin et al., 2005; Bruneau et al., 2008). Not only to the putative rapid radiation of the order, but also to taxon sampling directed either above (i.e. angiosperms) or below (i.e. Leguminosae, Polygalaceae) the ordinal level of interest might have caused this phylogenetic instability (Bello et al., 2009). Nevertheless, even studies which focused on Fabales such as Forest (2004), Forest et al., (2007), Bello et al., (2009) and Bello et al., (2012) could not yield robust relationships for the order. For instance, (*rbc*L+*mat*K) and (*rbc*L+*mat*K+66 morphological characters) analyses of Bello et al., (2009) and Bello et al., (2012) respectively, yielded several low to moderately supported toplogies such as (((S+Q)L)P) and (L+P)(S+Q), however (((S+Q)L)P) was considered the most likely topology among them in both studies. Moreover, to date the DNA sequence loci used in phylogenetic reconstructions within the order have mostly been from the plastid genome.

A simulation (power) analysis is a statistical test that may help to find the amount of data needed to resolve phylogenetic problems (Whitfield & Lockhart, 2007). Simulation studies were reported as useful for systematic studies that plan to work on problematic groups (Wortley et al., 2005). In a simulation study, it is feasible to find the required sequence length, the most appropriate sequence type and the effect of combining partitions to solve a difficult phylogeny (Wortley et al., 2005). In many multigene studies, real data sets may conflict with each other, and it is not possible to predict this in simulation tests (Spinks et al., 2009); therefore, before attempting to sequence several genes, it may be time and cost-effective to estimate the required number of base pairs and types of data. Thus, in the case of Fabales, power analyses would give an estimation of the genes needed to solve this difficult phylogeny. Therefore, the aims of this study are (1) to estimate the most suitable sequence type (fast-evolving or slowly evolving), and (2) to determine the required sequence length to solve the poorly resolved interfamilial relationships of Fabales.

For these aims, the study of Wortley et al. (2005) is used as an example. Two plastid regions are employed, *mat*K and *rbc*L.

The ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit plastid gene (*rbc*L), which encodes the large subunit of RuBisCO protein, is one of the most commonly sequenced regions in high level (i.e., above family level) phylogenetic studies due to its ease of amplification and high sequence recovery rate (CBOL Plant Working Group, 2009). Indeed, several studies have reported that particularly for the ancient-rapid radiations slow-gene regions, like *rbc*L, may be more efficient because fast-evolving genes are prone to phylogenetic artefacts, such as long branch attraction (LBA), loss of phylogenetic signal and homoplasy (Felsenstein, 1978; Gribaldo & Philippe, 2002; Whitfield & Lockhart, 2007). In addition, the low discrimination capability of the region has also been questioned by several studies (e.g., Zhang et al. 2015). Some studies concluded that short internal branches may become even shorter, when slowly evolving genes are employed (e.g., Roberts *et al*., 2009). However, the possibility of the region contributing to a robust combined analysis should not be ruled out.

Intron Group II maturase *mat*K encodes a splicing-associated maturase protein. The *matK*, is one of the most rapidly evolving coding regions in plants and shows high level of species discrimination ability among angiosperms (Lahaye et al. 2008; CBOL Plant Working Group, 2009), even at the species level (Dong et al. 2012). The requirement of specific primers and reported PCR problems are the most significant drawbacks of this gene region; however, it still shows a very high level of species discrimination ability among angiosperms (Lahaye et al. 2008).

2. Materials and Methods

2.1. Taxon sampling

The dataset contained published *mat*K and *rbc*L plastid gene regions from the National Center for Biotechnology Information (NCBI/GenBank) for 27 taxa in total: 15 taxa from Leguminosae, nine taxa from Polygalaceae, one taxon from Surianaceae and the sole genus of Quillajaceae, *Quillaja*. One outgroup taxon, *Krameria ixine* L. (Zygophyllales Link) was also included to root the Fabales phylogeny. The National Center for Biotechnology Information (NCBI/GenBank) accession numbers for these previously published DNA sequences are provided in Appendix 1.
Sequences were assembled and aligned using the Geneious alignment option in Geneious Pro 4.8.4 (Kearse et al. 2012) with the automatic pairwise alignment tool and subsequently edited manually. Equivocal base calling at the beginning and end of assembled complementary strands were trimmed. All insertion and deletions (indels) were scored as missing data.

2.2. Methods to find the necessary amount and type of data to resolve Fabales phylogeny

2.2.1. Method 1: sequence simulation by Seq-Gen

Fully resolved and outgroup rooted neighbour joining (NJ) trees were created by PAUP 4.0b10 (Swofford, 2002) for *rbc*L, *mat*K and *rbc*L+*mat*K data. jModelTest 0.1 (Posada, 2008; Darriba et al., 2012) was used to find the most suitable evolutionary model for each data set. For the *rbc*L region, the most realistic model was GTR+I+G, with relative base sequences A = 0.2637, C = 0.1920, G = 0.2443, T = 0.3000, relative substitution rates [AC] = 1.8817, [AG] =

3.4498, [AT] = 0.6866, [CG] = 0.9896, [CT] = 4.8329, [GT] = 1.0000, gamma distribution shape parameter 0.5590, and proportion of invariable sites 0.5400.

For the *mat*K region, the most suitable evolutionary model was the GTR+G model with relative base sequences A = 0.3091, C = 0.1504, G = 0.1506, T = 0.3899, relative substitution rates [AC] = 1.2704, [AG] = 1.9872, [AT] = 0.2821, [CG] = 1.2551, [CT] = 1.9595, [GT] = 1.0000, gamma shape parameter = 1.2280.
Both *rbc*L and *mat*K regions were simulated by Seq-Gen v1.3.2 (Rambaut and Grass, 1997) for 3,000, 5,000, 10,000, 15,000 and 20,000 bp. For the combined data sets, GTR+I+G and GTR+I models were used to see the results of effects of both models. The *rbc*L and *mat*K sequence length, and the total sequence length for these combined datasets are shown in Table 1. For all datasets, the number of replicate matrices (n) was set to 100.

Table 1. Combined simulated data sets by Seq-Gen v1.3.2 (Rambaut and Grassly, 1997). bp=base pairs (please note that the same data sets were simulated according to GTR+I+G and GTR+I models, separately).

*rbc*L sequence length (bp)	*mat*K sequence length (bp)	Total sequence length (bp)
1,000	1,000	2,000
1,000	2,000	3,000
1,000	3,000	4,000
1,000	5,000	6,000
1,000	10,000	11,000
2,000	1,000	3,000
2,000	2,000	4,000
2,000	3,000	5,000
2,000	5,000	7,000
2,000	10,000	12,000
3,000	1,000	4,000
3,000	2,000	5,000
3,000	3,000	6,000
3,000	5,000	8,000
3,000	10,000	13,000
5,000	1,000	6,000
5,000	2,000	7,000
5,000	3,000	8,000
5,000	5,000	10,000
5,000	10,000	15,000
10,000	1,000	11,000
10,000	2,000	12,000
10,000	3,000	13,000
10,000	5,000	15,000
10,000	10,000	20,000

2.2.2. Method 2: manual sequence simulation

In this method the sequences were replicated on Geneious Pro 4.8.4 (Kearse et al. 2012) by employing the "new sequence" option. Existent *mat*K and *rbc*L sequences for each taxon were replicated from 2,000 bp to 7,000 bp by 1,000 bp increments, and *rbc*L and *mat*K were manually duplicated from the original sequences of *rbc*L and *mat*K regions for 3,000 bp, 5,000 bp, 7,000 bp, 9,000 bp, 12,000 bp and 15,000 bp. All taxa were aligned to generate the NEXUS files.

2.3. Phylogenetic analysis

For both methods, parsimony analyses were conducted using PAUP 4.0b10 (Swofford, 2002) with 1,000 replicate heuristic searches, tree-bisection-reconnection (TBR) branch- swapping, MULPARS on, holding 10 trees per replicate and saving all trees. After the parsimony searches, strict consensus trees were generated and rooted by the outgroup. FigTree 1.4.2 (Rambaut, 2009; Rambaut 2014) was used to visualize tree files.

2.4. Comparisons

Internal support, accuracy and resolution were used for comparisons. First, non-parametric bootstrap searches with 100 replicates were carried out to estimate the internal support for parsimony trees; the bootstrap supports above 90% were counted, and the percentage of these high supports were calculated. Second, accuracy was measured by the presence of the key nodes of the phylogenetic analyses of Bello et al. (2009) and Bello et al. (2012), which were a monophyletic Fabales, monophyletic Polygalaceae and Leguminosae, (Quillajaceae+Surianaceae) bifurcation and (Leguminosae (Quillajaceae+Surianaceae)) bifurcation. Lastly, consensus fork index (CFI) was estimated for the resolution. For this calculation, the formula below was used, where CI_c is the consistency fork index, B_i is the number of resolved internal branches, B_t is the number of terminal branches and R is two or three dependent on whether the tree is rooted (two) or has a basal polytomy (three) (Soltis et al., 1998; Wortley et al., 2005):
$CI_C = B_i \div (B_t - R)$

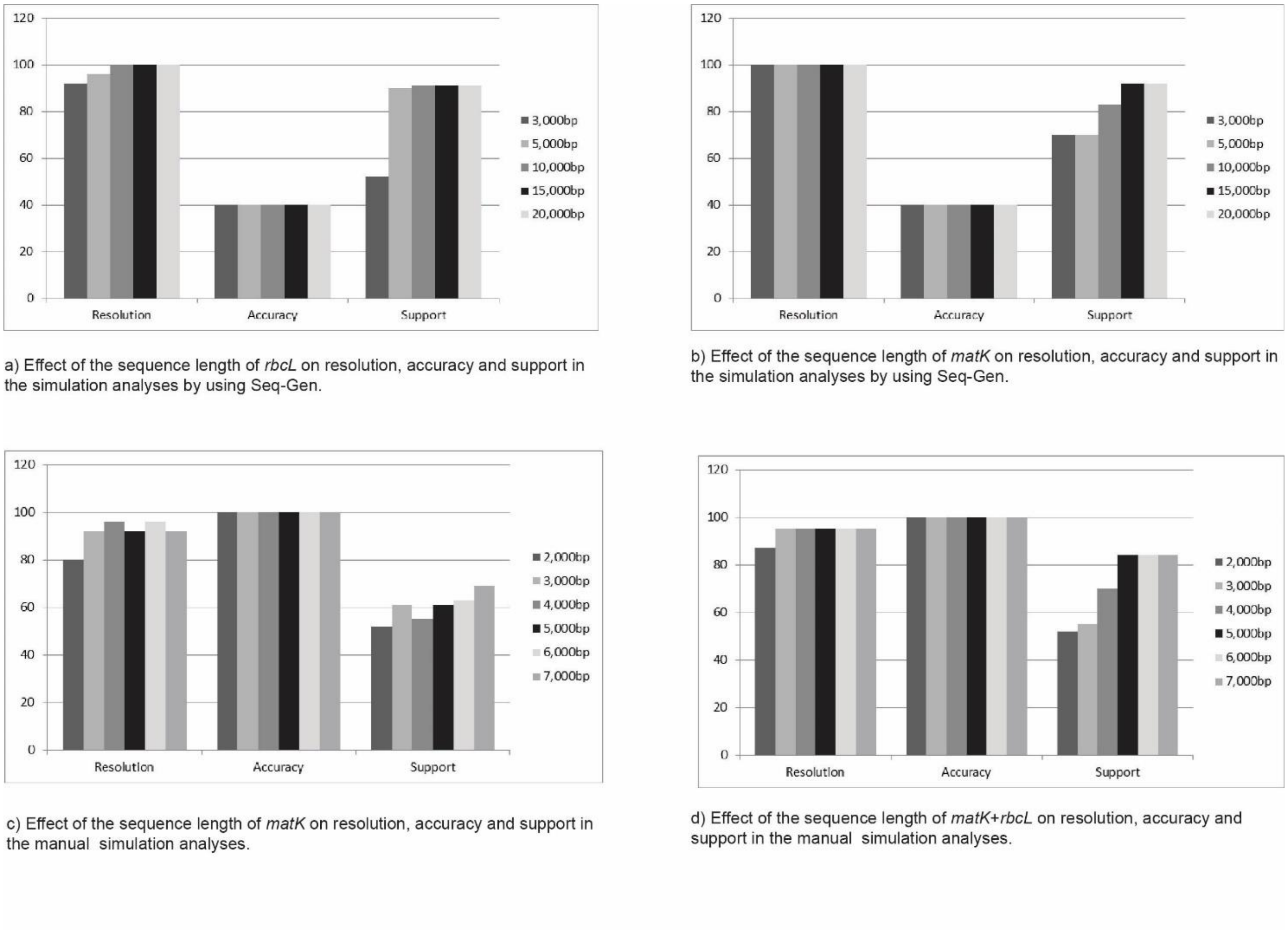
3. Results

3.1. Method 1: Sequence simulation by Seq-Gen

3.1.1. Separate analyses

In general, *rbc*L phylogenetic trees did not yield accurate relationships within Fabales (Figure 1.a). Resolution was 100% starting from 10,000 bp. However, a sharp increase was seen for the internal support after 5,000 bp.

The *mat*K region showed a similar pattern (Figure 1.b). While there were no significant differences among different data sets (i.e., different amounts of data) in terms of resolution and accuracy (the resolution was 100% and the accuracy was only 40% for all datasets), internal support increased gradually as sequence length increased.



a) Effect of the sequence length of *rbcL* on resolution, accuracy and support in the simulation analyses by using Seq-Gen.

b) Effect of the sequence length of *matK* on resolution, accuracy and support in the simulation analyses by using Seq-Gen.

c) Effect of the sequence length of *matK* on resolution, accuracy and support in the manual simulation analyses.

d) Effect of the sequence length of *matK+rbcL* on resolution, accuracy and support in the manual simulation analyses.

Figure 1:. The effect of the sequence length of *rbc*L (a), *mat*K (b and c) and *rbc*L+*mat*K (d) on resolution, accuracy and support in manual simulation analyses and simulations by Seq-Gen. The resolution was evaluated by CFI, the accuracy evaluated by the percentage of the presence of key nodes present, and the support was evaluated by the percentage of the bootstrap values greater than 90%. bp: base pairs.

3.1.2. Combined analysis

Fifty strict consensus trees that were generated according to the *mat*K model yielded polyphyletic clades, especially for the Leguminosae family, and among these trees only five of them yielded the (((QS)L)P relationship. Therefore, the results of these analyses are not included in Figure 1.

Similarly, among the trees generated according to the *rbc*L type model, five phylogenetic trees yielded monophyletic Fabales families. These were: 3,000 bp tree (2,000 bp of *rbc*L and 1,000 bp of *mat*K), 11,000 bp tree (10,000 bp of *rbc*L and 1,000 bp of *mat*K), 13,000 bp tree (10,000 bp of *rbc*L and 3,000 bp of *mat*K), 15,000 bp tree (5,000 bp of *rbc*L and 10,000 bp of *mat*K),and 20,000 bp tree (10,000 bp of *rbc*L and 10,000 bp of *mat*K). However, none of these trees yielded the most likely intrafamilial relationship for Fabales (Table 2). Furthermore, resolution, accuracy and support statistics for these trees did not show a gradual increase with the increase in sequence length. Therefore, the results of these analyses are also not included in Figure 1.

Table 2. The results of the strict consensus parsimony trees of 3,000 bp, 11,000 bp, 13,000 bp, 15,000 bp and 20,000 bp according to the simulations under *rbc*L model. L: Leguminosae, P: Polygalaceae, Q: Quillajaceae and S: Surianaceae.

Name of the analysis	Resulting topology
3,000 bp strict consensus parsimony tree with 2,000 bp of *rbc*L and 1,000 bp of *mat*K	(((QL)S)P)
11,000 bp strict consensus parsimony tree with 10,000 bp of *rbc*L and 1,000 bp of *mat*K	(((QP)S)L)
13,000 bp strict consensus parsimony tree with 10,000 bp of *rbc*L and 3,000 bp of *mat*K	((LS)(QP))
15,000 bp strict consensus parsimony tree with 5,000 bp of *rbc*L and 10,000 bp of *mat*K	((PS)(QL))
20,000 bp strict consensus parsimony tree with 10,000 bp of *rbc*L and 10,000 bp of*mat*K	(((LS)Q)P)

3.2. Method 2: manual sequence simulation

3.2.1. Separate analyses

While any of the *rbc*L simulated trees did not yield monophyletic clades, all *mat*K trees yielded monophyletic families and “correct” relationships among the families of Fabales (Figure 1.c). Additionally, for the *mat*K trees, while the support for monophyletic Leguminosae and Polygalaceae was always high (≥ 95%), the support values for (Q+S) sister relationship and ((Q+S)F) bifurcation reduced dramatically as a result of increasing the sequence lengths. For example, for the (Q+S) clade, the bootstrap values were 56% for 2,000 bp and 91% for 7,000 bp. Similarly, for the ((Q+S)L) bifurcation, the bootstrap support was 67% at 2,000 bp and 97% at 7,000 bp. In general, resolution and support increased very little with the increase in sequence length.

3.2.2. Combined analysis

Combined analysis yielded monophyletic, moderately supported trees and always a (((QS)L)P topology. Moreover, whilst the bootstrap supports were always high for the monophyletic Polygalaceae and Leguminosae, there was almost no difference for the (Q+S) and ((Q+S)L) bifurcation support values between 3,000 bp and 15,000 bp. On the other hand, whilst the accuracy was 100% for all datasets, and the resolution was almost always high (except 2,000 bp); support was gradually increased and stabilized after 5,000 bp (Figure 1.d).

4. Discussion

4.1. Interpretation of general results of the simulation study

In general, accuracy did not improve with an increase in sequence length, which contradicts the results of Wortley et al. (2005). A possible reason for the stability of the accuracy in Method 1 is that the software simulated sequences according to the given wrong phylogeny (NJ starting tree); therefore, increasing the sequence length did not improve the results. On the contrary, for Method 2, without a starting tree, all datasets yielded 100% accuracy.

Second, by using either Method 1 or Method 2, in almost all analyses, support increased by the increase of sequence length (up until a certain point). The resolution, in some cases, increased until a certain point or showed mixed results; in others, it did not change at all (e.g., the simulated *mat*K trees according to Method 2, Figure 1.c). The reason for this is the unresolved relationship of *Ceratonia siliqua* L. within Leguminosae in the 3,000 bp, 5,000 bp and 7,000 bp trees. One of the possible reasons for this pattern may be the different percentage of *rbc*L / *mat*K sequences in these analyses.

In general, manually simulated sequences yielded more accurate phylogenetic relationships, and there were a number of monophyletic groupings in the trees. However, both the NJ trees of separate analyses and the combined analyses did not yield “correct” phylogenetic relationships for Fabales families. An additional tree construction method was not employed because the aim of this study is to determine how many base pairs are enough to solve the phylogenetic relationships within Fabales, not to find the most suitable tree construction method for this problematic order. Yet, adding several thousands of base pairs did not improve the results and the phylogeny continued to produce incorrect phylogenetic relationships. This problem may be that a result of the specified tree being wrong, which would cause the software to simulate the sequences wrong as well.

On the other hand, the results of this study clearly indicate that the *mat*K type of data is more suitable to solve possible relationships within Fabales. For example, bootstrap supports for (QS) and ((QS)L) nodes dramatically increased as the sequence length increased, and it appears that just 5,000 bp is sufficient to solve the Fabales phylogeny problem. Additionally, the results of Bello et al., (2009) support this conclusion with less homoplastic character (estimated with CI) of *mat*K and higher support values than *rbc*L analyses. However, two points should be noted in this case: first, all analyses were interpreted according to previous assumptions of the phylogenetic relationships within Fabales (i.e. Bello et al., 2009, Bello et al., 2012), but in these studies the supports for (QS) and ((QS)L) forks were always low, and morphology was one of the most important reasons for determining the possible evolutionary relationship. Second, while Bello et al. (2009) and Bello et al. (2012) had the most equal taxa coverage among all studies, the others did not report a possible (((Q+S)L)P relationship (i.e. Forest., 2004). If the supposed phylogenetic inference represents the most possible evolutionary relationship for Fabales, employing a *mat*K type data set for future analyses may help to strengthen this hypothesis with “correct” and well-supported clades.

Lastly, several studies reported that, especially for ancient phylogenies, slow gene regions are more efficient; for instance, Whitfield & Kjer (2008) reported that while fast-evolving genes may help for the short internodes, these genes are more likely to be overwritten and be more homoplastic than the slowly evolving genes. Indeed, fast-evolving regions are more prone to certain artefacts such as LBA (Felsenstein, 1978) and loss of phylogenetic signal (i.e., mutational saturation) (Gribaldo & Philippe, 2002; Whitfield & Lockhart, 2007). Thus, extra caution should be taken for employing these fast-evolving regions in any phylogenetic analysis.

In summary, the results of the current study demonstrate that Fabales phylogeny may be solved successfully by employing just 5,000 bp of rapidly evolving genes (e.g., *mat*K type). This is in contrast to slowly evolving genes (e.g., *rbc*L) and some examples from ToL (Tree of Life) and IBOL (International Barcode of Life) which may not be resolved even with maximal data (e.g. whole genome sequences) in the cases of non-tree like bits of the tree such as hybridization and paralogy where many genes robustly support alternative topologies (Rokas & Carrol, 2006; Cotton & Page, 2012); this is because Fabales is not one of the cases that indicates a hard polytomy (Bello et al., 2009). Moreover, this study sounds a note of caution, with respect to interpreting the results of the “more data” approach, because, while while simulated data does not always behave like real data (Spinks et al., 2009; Schäferhoff et al., 2010), larger datasets can easily support an arbitrary root of Fabales (Jeffroy et al., 2006; Rokas & Carroll, 2006).

4.2. Future work

For Fabales a more-data approach is essential, but with the inclusion of an adequate number of taxa (Raman & Park, 2016; Pereira et al., 2017; Reddy et al., 2017) and with the “right” gene regions (i.e., fast-evolving gene regions that can be alignable across families or genes with strong phylogenetic signal), despite its drawbacks, such as homoplasy, alignment difficulties and LBA. Cases that have shown that the more data approach is useful are not rare in the literature such as Rosaceae (Zhang et al., 2017), Mesaangiospermae (Zeng et al., 2014), eudicots (Zeng et al., 2017), Brassicaeae (Huang et al., 2015), Vitaceae Juss. (Raman & Park, 2016), birds (Reddy et al., 2017), salamanders (Rodriguez et al., 2017), turtles (Pereira et al., 2017; Shaffer et al., 2017), and genus *Oryza* L. (Zou et al., 2008), because, especially in the case of rapid radiations additional data can increase the internal branch lengths; consequently, resolving the phylogeny question becomes easier. In this case, nuclear genes have been very useful in answering controversial phylogenetic questions since they are biparentlly inherited, in contrast to the plastid data for thousands of base pairs that are linked and mostly inherited maternally (Sun et al., 2015). The importance of the model of evolution was also shown by several studies for tree of life, Animalia, placental mammals and Archaea (e.g., Morgan et al., 2013; Pisani et al., 2015; Tarver et al., 2016); therefore, a complex model such as a heterogeneous model that allows sites or time periods to evolve under different models may help to reconstruct a correct root for Fabales.

References

1. **A.P.G. (2016).** An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *J. Linn. Soc. Bot.*, 181, 1-20.
2. **Baas P., Wheeler E. & Chase M. (2000).** Dicotyledonous wood anatomy and the APG system of angiosperm classification. *J. Linn. Soc., Bot*., 134, 3-17.
3. **Bello M. A., Bruneau A., Forest F. & Hawkins J. A. (2009).** Elusive relationships within order Fabales: phylogenetic analyses using *mat*K and *rbc*L sequence data. *Syst. Bot.*, 34, 102-114.
4. **Bello M. A., Hawkins, J. A. & Rudall P. J. (2010).** Floral ontogeny in Polygalaceae and its bearing on the homologies of keeled flowers in Fabales. *Int. J. Plant Sc.i*, 171, 482-498.
5. **Bello M. A., Rudall P. J. & Hawkins J. A. (2012).** Combined phylogenetic analyses reveal interfamilial relationships and patterns of floral evolution in the eudicot order Fabales. *Cladistics*, 28, 393-421.
6. **Bruneau A., Mercure M., Lewis G. P. & Herendeen P. S. (2008).** Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany*, 86, 697-718.
7. **Cannon S. B., Mckain M. R., Harkess A., Nelson M.N., Dash S., Deyholos M. K., Peng, Y. Joyce, B. Stewart Jr C. N., Rolf M. & Kutchan, T. (2014).** Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.*, 32(1), 193-210.
8. **CBOL Plant Working Group. (2009).** A DNA barcode for land plants. Proc. Natl. Acad. Sci. U.S.A. 106(31), 12794-12797.Chodat, R. 1896. Polygalaceae novae vel parum cognitae. *Bulletin de l'Herbier Boissier*, 4, 233-237.
9. **Chodat R. (1896).** Polygalaceae novae vel parum cognitae. *Bulletin de l'Herbier Boissier*, 4, 233-237.
10. **Cotton J. A. & Page R. D. (2002).** Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proc. R. Soc. B.,* 269, 1555-61.
11. **Darriba D., Taboada G. L., Doallo R. & Posada D. (2012).** jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, 9, 772-772.
12. **Dong W., Liu J., Yu J., Wang L. & Zhou S. (2012).** Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PloS One*, 7(4), e35071.
13. **Eriksen, B. (1993).** Phylogeny of the Polygalaceae and its taxonomic implications. *Plant Syst. Evol.*, 186, 33-55.
14. **Eriksen B. & Persson C. (2007).** *Polygalaceae, Families and genera of flowering plants*. In: K. Kubitski, editors. Springer, Berlin.
15. **Felsenstein J. (1978).** Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Biol.*, 27, 401-410.
16. **Forest F. (2004).** *Systematics of Fabales and Polygalaceae, with emphasis on Muraltia and the origin of the Cape flora.* Reading: University of Reading.
17. **Forest F., Chase M. W., Persson C., Crane P. R. & Hawkins J. A. (2007).** The role of biotic and abiotic factors in evolution of ant dispersal in the milkwort family (Polygalaceae). *Evolution,* 61, 1675-1694.
18. **Gribaldo S. & Philippe H. (2002).** Ancient phylogenetic relationships. *Theor. Popul. Biol.*, 61, 391-408.
19. **Huang C-H., Sun R., Hu Y., Zeng L., Zhang N., Cai L., Zhang Q., Koch M. A., Al-Shehbaz I., Edger P. P., Pires J. C., Tan D.-Y., Zhong Y. & Ma H. (2015).** Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Mol. Biol. Evol.*, 33(2), 394–412.
20. **Jansen R. K., Kaittanis C., Lee S. B., Saski C., Tomkins J., Alverson A. J. & Daniell H. (2006).** Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol. Biol.,* 6, 32.
21. **Jeffroy O., Brinkmann H., Delsuc F. & Philippe H. (2006).** Phylogenomics: the beginning of incongruence? *Trends Genet.,* 22, 225-231.
22. **Kajita T., Ohashi H., Tateishi Y., Bailey C. D. & Doyle J. J. (2001).** *rbc*L and legume phylogeny, with particular reference to Phaseoleae, Millettieae, and allies. *Syst. Bot.*, 26, 515-536.
23. **Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S. & Duran C. (2012).** Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647-1649.
24. **Kubitzki K. (2007).** *Quillajaceae. Flowering Plants· Eudicots*. Springer Berlin Heidelberg.
25. **Lahaye R., Van der Bank M., Bogarin D., Warner J., Pupulin F., Gigot G., Maurin O., Duthoit S., Barraclough T.G. & Savolainen V. (2008).** DNA barcoding the floras of biodiversity hotspots. *PNAS USA*, 105(8), 2923-2928.

26. **Lavin M., Herendeen P. S. & Wojciechowski M. F. (2005).** Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. *Syst. Biol.*, 54, 575-594.
27. **Lewis G. P. (2005).** *Legumes of the World*, Royal Botanic Gardens, Kew.
28. **Lewis G., Schrire B., Mackinder B., Rico L. & Clark R. (2013).** A 2013 linear sequence of legume genera set in a phylogenetic context-a tool for collections management and taxon sampling. *S. Afr. J. Bot.*, 89, 76-84.
29. **L.P.W.G. (2017).** A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. *Taxon*, 66 (1), 44-77.
30. **Mabberley D. J. (1997).** *The plant-book: a portable dictionary of the vascular plants*. Cambridge University Press, Cambridge.
31. **Moore M. J., Soltis P. S., Bell C. D., Burleigh G. & Soltis D. E. (2010).** Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *PNAS USA*, 107 (10), 4623-4628.
32. **Morgan C. C., Foster P. G., Webb A. E., Pisani D., Mcinerney J. O. & O'connell M. J. (2013).** Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.*, 30, 2145-56.
33. **Pereira A. G., Sterli J., Moreira F. R. & Schrago C. G. (2017).** Multilocus phylogeny and statistical biogeography clarify the evolutionary history of major lineages of turtles. *Mol. Phylogenet. Evol.*, 113, 59-66.
34. **Persson C. (2001).** Phylogenetic relationships in Polygalaceae based on plastid DNA sequences from the *trnL-F* region. *Taxon*, 763-779.
35. **Pisani D., Pett W., Dohrmann M., Feuda R., Rota-Stabelli O., Philippe H., Lartillot N. & Wörheide G. (2015).** Genomic data do not support comb jellies as the sister group to all other animals. *PNAS USA*, 112, 15402-15407.
36. **Posada D. (2008).** jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.*, 25, 1253-1256.
37. **Raman G. & Park S. (2016).** The complete chloroplast genome sequence of *Ampelopsis*: gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. *Front. Plant Sci.*, 7. 341.
38. **Rambaut A. & Grass N. C. (1997).** Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *CABIOS*, 13, 235-238.
39. **Rambaut A. (2014).** FigTree 1.4.2 software. Institute of Evolutionary Biology, University of Edinburgh.
40. **Reddy S., Kimball R. T., Pandey A., Hosner P. A., Braun M. J., Hackett S. J., Han K. L., Harshman J., Huddleston C. J., Kingston S. & Marks B. D. (2017).** Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst. Biol.*, 66 (4), 857-879.
41. **Roberts T. E., Sargis E. J. & Olson L. E. (2009).** Networks, trees, and treeshrews: assessing support and identifying conflict with multiple loci and a problematic root. *Syst. Biol*., 58, 257-70.
42. **Rodríguez A., Burgon J. D., Lyra M., Irisarri I., Baurain D., Blaustein L., Göçmen B., Künzel S., Mable B. K., Nolte A. W. & Veith M. (2017).** Inferring the shallow phylogeny of true salamanders (*Salamandra*) by multiple phylogenomic approaches. *Mol. Phylogenet. Evol.*, 115, 16-26.
43. **Rokas A. & Carroll S. B. 2006.** Bushes in the tree of life. *PLoS Biology*, 4, e352.
44. **Savolainen V., Chase M. W., Hoot S. B., Morton C. M., Soltis D. E., Bayer C., Fay M. F., De Bruijn A. Y., Sullivan S. & Qiu Y.-L. (2000).** Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.*, 49, 306-362.
45. **Schäferhoff B., Fleischmann A., Fischer E., Albach D. C., Borsch T., Heubl G. & Müller K. F. (2010).** Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evol. Biol.*, 10, 352.
46. **Schneider J. V. (2007).** *Surianaceae. In Flowering Plants· Eudicots*. Springer Berlin Heidelberg.
47. **Shaffer H. B., Mccartney-Melstad E., Near T. J., Mount G. G. & Spinks P. Q. (2017).** Phylogenomic analyses of 539 highly informative loci dates a fully resolved time tree for the major clades of living turtles (Testudines). *Mol. Phylogenet. Evol.*, 11, 7-15.
48. **Smith A. B. (1994).** Rooting molecular trees: problems and strategies. *Biol. J. Linn. Soc. Lond.*, 51, 279-292.
49. **Soltis D. E., Soltis P. S., Morgan D. R., Swensen S. M., Mullin B. C., Dowd J. M. & Martin P. G. (1995).** Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. *PNAS USA*, 92, 2647-2651.
50. **Soltis D. E., Soltis P. S., Mort M. E., Chase M. W., Savolainen V., Hoot S. B. & Morton C. M. (1998).** Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. *Syst. Biol.*, 47(1), 32-42.
51. **Spinks P. Q., Thomson R. C., Lovely G. A. & Shaffer H. B. (2009).** Assessing what is needed to resolve a molecular phylogeny: simulations and empirical data from emydid turtles. *BMC Evol. Biol.*, 9, 56.

52. **Sun M., Soltis D. E., Soltis P. S., Zhu X., Burleigh J. G. & Chen Z. (2015).** Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.*, 83, 156-166.
53. **Swofford D. (2002).** PAUP* version 4.0 b10. Phylogenetic analysis using parsimony (* and other methods). Sinauer, Sunderland, MA.
54. **Tarver J. E., Dos Reis M., Mirarab S., Moran R. J., Parker S., O'reilly J. E., King B. L., O'connell M. J., Asher R. J., Warnow T., Peterson K. J., Donoghue P. C. & Pisani D. (2016).** The interrelationships of placental mammals and the limits of phylogenetic inference. *GBE*, 8, 330-44.
55. **Wang H., Moore M. J., Soltis P. S., Bell C. D., Brockington S. F., Alexandre R., Davis C. C., Latvis M., Manchester S. R. & Soltis D. E. (2009).** Rosid radiation and the rapid rise of angiosperm-dominated forests. *PNAS USA, USA*, 106, 3853-8.
56. **Watson L. & Dallwitz M. J. (1992 onwards).** The families of flowering plants: descriptions, illustrations, identification, and information retrieval. Version: 20th July 2017.
57. **Westerkamp C. (1997).** Keel blossoms: bee flowers with adaptations against bees. *Flora: Morphologie, Geobotanik, Oekophysiologie*, 192,125-32.
58. **Whitfield J. B. & Lockhart P. J. (2007).** Deciphering ancient rapid radiations. *Trends Ecol. Evol.*, 22, 258-65.
59. **Whitfield J. B. & Kjer K. M. (2008).** Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annual Review of Entomology*, 53, 449-72.
60. **Williams T. A., Heaps S. E., Cherlin S., Nye T. M., Boys R. J. & Embley T. M. (2015).** New substitution models for rooting phylogenetic trees. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370, 20140336.
61. **Wojciechowski M. F., Lavin M. & Sanderson M. J. (2004).** A phylogeny of legumes (Leguminosae) based on analysis of the plastid *mat*K gene resolves many well-supported subclades within the family. *Am. J. Bot.*, 91, 1846-1862.
62. **Wortley A. H., Rudall P. J., Harris D. J. & Scotland R. W. (2005).** How much data are needed to resolve a difficult phylogeny? A case study in Lamiales. *Syst. Biol.*, 54, 697-709.
63. **Zeng L., Zhang Q., Sun R., Kong H., Zhang N. & Ma H. (2014).** Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.*, 5, 4956.
64. **Zeng L., Zhang N., Zhang Q., Endress P. K., Huang J. & Ma H. (2017).** Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.*, 214(3), 1338-1354.
65. **Zhang J., Chen M., Dong X., Lin R., Fan J. & Chen Z. (2015).** Evaluation of four commonly used DNA barcoding loci for Chinese medicinal plants of the family Schisandraceae. *PloS one*, *10*(5), p.e0125574.
66. **Zhang S. D., Jin J. J., Chen S. Y., Chase M. W., Soltis D. E., Li H. T., Yang J. B., Li D. Z. & Yi T. S. (2017).** Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. *New Phytol.*, 2143, 1355-1367.
67. **Zou X. H., Zhang F. M., Zhang J. G., Zang L. L., Tang L., Wang J., Sang T. & Ge S. (2008).** Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biol.*, 9, R49.

Appendix 1: Taxon sampling for the simulation study of the order Fabales.

		GenBank accessions	
Fabales		*rbc*L	*mat*K
Leguminosae	*Chamaecrista nictitans* (L.) Moench	AM234248	EU361914
	Haematoxylum brasiletto Karst.	AY904384.1	AY386905.1
	Caesalpinia calycina Benth.	AM234236	EU361899
	Cercis canadensis L.	U74188.1	AY386908
	Bauhinia syringifolia (F. Muell.) Wunderlin	AM234267	EU361878
	Cassia grandis L.f.	AM234244	JQ587551.1
	Amherstia nobilis Wall.	AM234234	AF542601
	Gleditsia triacanthos L.	Z70129	AY386849
	Petalostylis labicheoides R. Br.	AF308719	AY386895
	Lecointea peruviana Barneby.	AM234260	EU361990
	Senna alata (L.) Roxb.	U74250	EU362042
	Browneopsis ucayalina Huber	AM234233	EU361894
	Ceratonia siliqua L.	U74203	AY386852
	Cynometra mannii Oliv.	AM234231	EU361925
	Tamarindus indica L.	Z70160	EU362056
Polygalaceae	*Carpolobia alba* G. Don	AM234176	EU604053
	Securidaca retusa Benth.	EU644681	EU604029
	Xanthophyllum sp.	AJ235799	EU604044
	Comesperma esulifolium (Gand.) Prain	AM234179	EU596516
	Monnina salicifolia Ruiz & Pavon	EU644694	EU604038
	Eriandra fragrans Royen & Steenis	AM234170	EU604051
	Polygala tenella Willd.	EU644687	EU604030
	Bredemeyera floribunda Gleason	EU644699.1	EU596520.1
	Atroxima afzeliana (Oliv. Ex Chod.) Stapf	AM234175	EU604049
Quillajaceae	*Quillaja saponaria* Molina	U06822	AY386843
Surianaceae	*Suriana maritima* L.	U07680	AY386950
Outgroups			
Krameriaceae	*Krameria ixine* Lofling.	EU644679	EU604050