UNIVERSIDADE FEDERAL DO PAMPA

CAMPUS SÃO GABRIEL

PROGRAMA DE PÓS- GRADUAÇÃO *STRICTO SENSU* EM CIÊNCIAS BIOLÓGICAS

ANÁLISE *IN SILICO* DE RELAÇÕES FILOGENÉTICAS DE PLANTAS BASEADAS NO GENE *GA20ox*

DISSERTAÇÃO DE MESTRADO

LILIAN DE OLIVEIRA MACHADO

SÃO GABRIEL, RS, BRASIL

2012

ANÁLISE *IN SILICO* DE RELAÇÕES FILOGENÉTICAS DE PLANTAS BASEADAS NO GENE *GA20ox*

Lilian de Oliveira Machado

Dissertação apresentada ao programa de Pós-Graduação *Stricto sensu* em Ciências Biológicas da Universidade Federal do Pampa (UNIPAMPA,RS), como requisito parcial para obtenção do Título de Mestre em Ciências Biológicas.

Orientador: Prof. Dr. Valdir Marcos Stefenon

São Gabriel, RS, Brasil

2012

MACHADO, Lilian de Oliveira

Análise *In Silico* de Relações Filogenéticas de Plantas Baseadas no Gene *GA20ox* / Lilian de Oliveira Machado.

23folhas;

Dissertação (Mestrado) Universidade Federal do Pampa, 2012.
Orientação: Prof. Dr. Valdir Marcos Stefenon.

1.Sistemática e Ecologia. 2. Ecologia Molecular. 3.Filogenia de plantas. I. Stefenon, Valdir Marcos.
II. Doutor

Lilian de Oliveira Machado

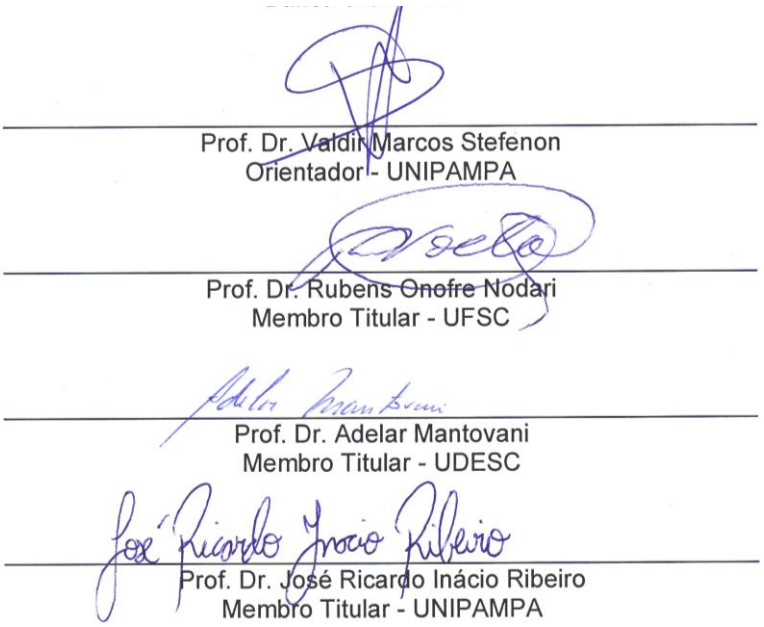
ANÁLISE *IN SILICO* DE RELAÇÕES FILOGENÉTICAS DE PLANTAS BASEADAS NO GENE *GA20ox*

Dissertação apresentada ao programa de Pós-Graduação *Stricto sensu* em Ciências Biológicas da Universidade Federal do Pampa (UNIPAMPA,RS), como requisito parcial para obtenção do Título de Mestre em Ciências Biológicas.

Área de concentração: Genética.

Dissertação defendida e aprovada em 14 de dezembro de 2012.

Banca Examinadora:



Prof. Dr. Valdir Marcos Stefenon
Orientador - UNIPAMPA

Prof. Dr. Rubens Onofre Nodari
Membro Titular - UFSC

Prof. Dr. Adelar Mantovani
Membro Titular - UDESC

Prof. Dr. José Ricardo Inácio Ribeiro
Membro Titular - UNIPAMPA

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por todas as oportunidades que me ofertou, pela força e coragem nos momentos mais difíceis.

Agradeço aos meus queridos pais, Altivo e Saionara, por todo carinho, apoio, compreensão, confiança e amor incondicional. A minha vó amada, Vera!!!! Muito obrigada por tudo que sempre fez por mim, pelo apoio e amor incondicional. Obrigada de coração, por tudo. Vocês são tudo pra mim! A minha irmã Daniele e as minhas primas amadas Quérolla e Ketelen pelo amor, carinho, compreensão, em todos os momentos.

Ao meu orientador prof. Dr. Valdir Marcos Stefenon por ter acreditado no meu potencial, até mais do que eu mesma. Muito obrigada pelos seus ensinamentos que foram de imensa importância nesses dois anos. Professor Valdir, se eu for 1/3 do que o senhor é, serei a profissional e a pessoa mais feliz desse mundo. Tenho imensa admiração pelo senhor tanto profissionalmente quanto pessoalmente. Pois, além de um excelente Professor é também um excelente Pai e Marido. Tive muita sorte de ter sido sua orientada. Muito obrigada por contribuir com o meu crescimento profissional e pessoal.

À Carmine Hister por toda ajuda no CIPBiotec, muito obrigada!!

À meu querido pai pelo concerto do fio que escapou de uma das partes do sequenciador.

Aos Professores Juliano, Zeca, Igor e Marília, muito obrigada por toda ajuda.

Aos meus queridos amigos de mestrado Ângela, Cris, Jô, Stefânia, Tiago, Thiago, Lucio, Leici e Alex pela amizade, e por estarem sempre dispostos a me ajudar e em especial a Ana, muito obrigada.

Ao pessoal do laboratório principalmente ao Tiaguinho, a Jô, a Ana, a Nanda, a Nathana, a Rayssa, a Manu, a Deise e a Natiéle, obrigada pela ajuda, incentivo e amizade!!!

Às minhas *friends* amadas, Afnan, Grazi, Letícia, Mari e Vivi pelo carinho, amizade, confiança e por todo apoio e força que me deram. Amo vocês!!!!

A todos que estiveram torcendo por mim, mesmo que em pensamento.

À comissão examinadora desta dissertação pela disponibilidade e contribuições.

Agradeço à UNIPAMPA e ao PPGCB pela possibilidade de realização deste curso, a FAPERGS pelo apoio financeiro e a CAPES pela bolsa de mestrado.

”Nenhum trabalho de qualidade pode ser
feito sem concentração e auto sacrifício,
esforço e dúvida.”
Max Beerbohm

RESUMO

Há inúmeros bancos de dados genéticos distribuídos por países da Europa e Ásia, mas todos trocam informações 24 horas por dia com o NCBI, considerado o mais importante. Dados *on-line* são submetidos e consultados nesses bancos remotos. Atualmente existe no NCBI aproximadamente 127 milhões de sequências de DNA, possibilitando a realização de inúmeros estudos *in silico*, incluindo estudos de sistemática. Os genes plastidiais são amplamente utilizados nas análises filogenéticas devido a sua eficácia, enquanto os genes nucleares têm sido pouco explorados nesse âmbito, deixando um vasto campo para estudos preliminares. Neste trabalho investigou-se a eficácia do gene nuclear *GA20ox1* na resolução das relações filogenéticas em eudicotiledôneas. Sequências do gene *rbcL* foram utilizadas para comparação. Todas as sequências foram obtidas diretamente no *GenBank*. A eficácia deste gene foi avaliada determinando-se a proporção de transversões/transições, número de caracteres parcimônio-informativos, índice de retenção (IR) e índice de consistência (IC) e através da construção de árvores filogenéticas. A proporção de mutações de transição/transversão e proporção de caracteres parcimônio-informativos do gene *GA20ox1* sugere alta capacidade informativa, apesar de os valores de IR e IC serem relativamente baixos. A árvore filogenética baseada no gene *GA20ox1* foi parcialmente congruente com as reconstruções filogenéticas baseadas nas sequências *rbcL* e com a árvore reconhecida pelo APG III. As incongruências observadas refletem os diferentes caminhos evolutivos seguidos pelos diferentes grupos de plantas com relação a este gene. Os resultados obtidos revelaram suficiente sinal filogenético no gene *GA20ox1* para agrupar espécies em nível de família, sendo útil para resolver as relações filogenéticas em níveis taxonômicos inferiores e aprimorar nossos conhecimentos sobre a evolução das plantas com flores.

Palavras-chave: banco de dados genético, gene nuclear, gene de cloroplasto, sistemática de plantas.

ABSTRACT

There are many genetic databases distributed across countries in Europe and Asia, but all share information 24 hours a day with the NCBI, considered the most important. Online data are submitted and consulted in these remote banks. Currently there are approximately 127 million DNA sequences in the NCBI, enabling the performance of several *in silico* studies, including studies on systematics. The plastid genes are most commonly used in phylogenetic analyzes due to its efficacy, while nuclear genes have been little exploited in this context, leaving a vast field for preliminary studies. This study investigated the efficacy of the nuclear gene *GA20ox1* in resolving phylogenetic relationships within the eudicots. Sequences of the *rbcL* gene were employed for comparison. All sequences were obtained directly from the GenBank. The efficacy of this gene was evaluated by determining the rate of transversions/transitions, proportion of parsimony-informative characters, retention index (RI) and consistency index (CI) and through the construction of phylogenetic trees. The rate of transversions/transitions mutations and proportion of parsimony-informative characters of the *GA20ox1* gene suggest high informative capacity, although the RI and CI values were relatively low. The phylogenetic tree based on the *GA20ox* gene was partially congruent with the phylogenetic reconstruction based on the *rbcL* sequences and with the tree recognized by the APG III. The observed incongruence reflect the different evolutionary ways followed by the different groups of plants concerning this gene. The results obtained revealed enough phylogenetic signal in the *GA20ox1* gene to group species at family level, being useful to resolve the phylogenetic relationships at lower taxonomic levels and increase our knowledge about the evolution of flowering plants.

Key-words: genetic database, nuclear gene, chloroplast gene, plant systematics.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E UNIDADES

Ácido Desoxirribonucleico - DNA

Angiosperm Phylogeny Group - APG

DNA Data Bank of Japan - DDJB

European Molecular Biology Laboratory - EMBL

Estados Unidos da América - EUA

Gene Giberelina 20 oxidase1 - *GA20ox1*

Genetic sequence database - GenBank

Giberelina - GA_n

International Nucleotide Sequence Database Colaboration - INSDC

Intron plastidial que codifica o gene matK - *trnK*

Mega base - Mb

National Center for Biotechnology Information - NCBI

Pares de Base - pb

Sequence Analysis Package – GCG

Subunidade maior da enzima fotossintética RuBisCO - *rbcL*

APRESENTAÇÃO

No item **INTRODUÇÃO**, apresenta um texto integrador sobre os temas trabalhados nesta dissertação.

A metodologia realizada e os resultados obtidos que fazem parte desta dissertação estão apresentados sob a forma de um manuscrito, os qual se encontra no item **MANUSCRITO**. Nesse item constam as seções: Material e Métodos, Resultados, Discussão e Referências Bibliográficas.

O item **CONCLUSÕES** encontra-se no final desta dissertação, e apresenta interpretações e comentários gerais sobre os resultados dos manuscritos presentes neste trabalho.

Às **REFERÊNCIAS** referem-se somente às citações que aparecem nos itens **INTRODUÇÃO** e **CONCLUSÕES** desta dissertação.

SUMÁRIO

1 INTRODUÇÃO

1.1 Bases de dados moleculares para análises *in silico*

As bases de dados moleculares são ferramentas importantes, pois os dados produzidos em todo o mundo podem ser visualizados por toda comunidade científica de forma acessível e rápida. A primeira base de dados de biologia molecular surgiu na década de 1960 devido aos avanços tecnológicos que permitiram que os computadores se tornassem ferramentas importantes na Biologia Molecular. Dayhoff *et al.* (1965) construíram um catálogo contendo todas as sequências de proteínas conhecidas até a data porém a quantidade de informações era limitada a 1 mega base (Mb) (Baxevanis e Ouellette, 2001).

A partir do final da década de 1970 surgiram pacotes de programas para a análise de sequências de nucleotídeos e de proteínas, como o Staden (Staden, 1977), o Pustell (Pustell e Kafatos, 1982) e o *sequence analysis package* (GCG) (Devereux *et al.*, 1984). Além desses programas, foram criadas bases de dados públicas servindo de repositórios para as sequências e resultados de análises das mesmas, como o *National Center for Biotechnology Information* (NCBI)/*Genetic sequence database* (*GenBank*) (GENBANK, 2012), *European Molecular Biology Laboratory* (EMBL, 2012), DNA *Data Bank of Japan* (DDBJ, 2012) e Swiss-Prot (SWISS-PROT, 2012) hoje parte do UniProt (UNIPROT, 2012). Essas bases de dados possibilitam a submissão individual de sequências de DNA e trocam informações entre si diariamente (Stoesser *et al.*, 2002).

Atualmente o NCBI é a maior base de dados público disponível, contando com aproximadamente 127 milhões de sequências (dados de dezembro de 2012) o que tem possibilitado a realização de inúmeros trabalhos *in silico.* Os genes nucleares apresentam algumas características, como a elevada taxa de evolução de sequência, a existência de múltiplos *loci* independentes e a herança de genes nucleares biparentais que tornam esses genes uma alternativa atrativa na avaliação das relações filogenéticas.

1.2 Sistemática Filogenética

Ligada à teoria da evolução, a sistemática filogenética, inicialmente proposta por Willi Hennig em sua obra de 1950, posteriormente traduzida para o inglês e ampliada (Hennig, 1966), é considerada o paradigma contemporâneo no campo da taxonomia e sistemática biológica (Schuh & Brower, 2009). A sistemática filogenética também conhecida como cladística é uma ferramenta importante para o entendimento da diversidade orgânica e para a reconstrução criteriosa de cenários histórico-evolutivos. Segundo Santos (2008) O refinamento da sistemática filogenética em relação à taxonomia evolutiva estava em discriminar caracteres plesiomórficos de apomórficos, e estabelecer relações de parentesco apenas a partir do compartilhamento dessas apomorfias, a fim de distinguir as homologias das convergências e determinar os grupos monofiléticos.

Nesses estudos, o grupo externo é utilizado como referência para polarizar as transformações dos caracteres. Para isso são utilizados no mínimo dois táxons que são escolhidos a partir de premissas básicas: os táxons escolhidos não devem fazer parte do grupo de interesse (grupo interno) e o grupo interno deve ser monofilético, ou seja, possua um ancestral comum (Maddison *et al.*, 1983).

Em 1998 foi criado nos Estados Unidos da América um grupo de pesquisadores interessados em resolver os problemas relacionados à classificação taxonômica das angiospermas, o Angiosperm Phylogeny Group (APG). Esse grupo utiliza a sistemática filogenética para gerar árvores filogenéticas que sejam capazes de refinar as dúvidas taxonômicas existentes para este grupo de plantas. A primeira árvore filogenética elaborada por este grupo foi construída com base em sequências de alguns genes como *rbcL* e *atpB*. Porém, a classificação proposta pelo grupo se alterou ao longo dos anos em alguns aspectos e nomenclaturas e já ocorreram duas atualizações de suas classificações. A primeira atualização ficou conhecida como sistema APGII (2003) e a segunda atualização ficou conhecida como sistema APGIII (2009). Esses dois últimos sistemas também foram baseadas em genes plastidiais. Contudo, genes nucleares também podem

apresentar importantes informações relacionadas à evolução das plantas, ou seja, sobre a sistemática filogenética deste grupo.

1.3 Os genes *GA20ox1* e *rbcL*

As enzimas *GA20ox* são responsáveis pela conversão da giberelina GA_{12} para a forma bioativa GA_4. Elas são codificadas por pequenas famílias de genes, com quatro cópias no genoma nuclear das plantas com flores (Hedden e Phillips, 2000, Huerta *et al.*, 2009, Lee e Zeevaart, 2007) e são reguladas por giberelinas bioativas através de mecanismos de respostas negativas e por influência de diversos fatores ambientais como luz, temperatura e de desenvolvimento da planta (Hedden e Kamiya, 1997, Kamiya e García-Martínez,1999, Lee e Zeevaart, 2007). Muitas vezes esses níveis de giberelina encontram-se em concentrações limitantes, mostrando que a regulação dessa via biossintética é um fator importante no controle do desenvolvimento de plantas (Coles *et al.*, 1999). A principal enzima desta família é codificada pelo gene *GA20ox1* e foi utilizada na produção de plantas transgênicas de *Arabidopsis thaliana* (Coles *et al.*, 1999; Rieu *et al.*, 2008) e *Populus tremula* (Erikson *et al.*, 2000).

O *rbcL* é um gene plastidial responsável pela codificação da unidade maior da enzima ribulose 1,5-bifosfato-carboxilase-oxigenase (RuBisCO). Essa é a proteína mais abundante nas folhas das plantas e é um componente indispensável no metabolismo do carbono (DUVALL *et al.*, 1993). Essa substância é responsável pela conversão de dióxido de carbono em água e carboidratos. O *rbcL* é um gene que possui baixas taxas de mutação. Esse gene apresenta baixas taxas de substituições nucleotídicas sinônimas em comparação com os genes nucleares e sua restrição funcional reduz as taxas evolucionárias das substituições não-sinônimas (Wolfe, Li e Sharp, 1987). Estas características tornam este gene um dos preferidos pelos sistematas na resolução das relações filogenética de grandes grupos como famílias e ordens (Duvall *et al.*, 1993; Hasebe *et al.*, 1994; Savolainen *et al.*, 2000, Heenan *et al.*, 2012).

Neste estudo, as relações filogenéticas entre representantes das eudicotiledôneas foram analisadas *in silico*, através da sistemática filogenética e baseada em sequências de DNA do gene *GA20ox1* retiradas do *GenBank*. As relações filogenéticas obtidas a partir deste gene foram comparadas com a árvore filogenética produzida com base em sequências do gene plastidial *rbcL*, de forma a avaliar a eficácia do gene nuclear *GA20ox1* na resolução das relações filogenéticas das plantas com flores.

2 OBJETIVO

- Investigar a eficácia do gene nuclear *GA20ox1* na resolução das relações filogenéticas entre representantes das eudicotiledôneas utilizando sequências nucleotídicas depositadas no *GenBank*.

3 MANUSCRITO

Title: Phylogenetic signal of the nuclear gene *GA20ox* in the core eudicots[1]

Authours: L. O. Machado, D. S. Schröder, N. M. Oliveira, V. M. Stefenon*

Universidade Federal do Pampa, Nucleus of Genomics and Molecular Ecology, Interdisciplinary Center of Biotechnological Research, Av. Antonio Trilha 1847, 97300-000, São Gabriel, RS, Brazil.

Running Title: *GA20ox* phylogeny of core eudicots

*Corresponding Author: Phone +55-55-32326075; Email, valdirstefenon@unipampa.edu.br

[1]Manuscrito submetido para publicação em "Plant Systematics and Evolution"

ABSTRACT:

Phylogenetic relationships among flowering plants have been dependent on nucleotide sequences of chloroplast and nuclear ribosomal genes and the analysis of sequences of nuclear genes may be an alternative to improve the resolution of these relationships. This study investigated the efficacy of the nuclear gene *GA20ox1* in resolving the phylogeny of plant species of the core eudicots group. DNA sequences of the gene *rbcL* were also employed for comparison of the phylogenetic relationships based on nuclear and plastid genes. All sequences were obtained from the GenBank. The ratio of transition to transversion mutations in the sequences of the *GA20ox1* gene equaled 1.5. The percentage of parsimony informative characters of the *GA20ox1* sequences were higher than of the *rbcL* sequences while the consistency and retention indexes were lower than the values observed for the *rbcL* gene. In the phylogenetic trees, the *GA20ox1* sequences provided enough phylogenetic signal to resolve relationships at the family levels within the core eudicots with high support. The *GA20ox1* phylogenetic reconstruction was partially congruent with the phylogenetic reconstruction based on the *rbcL* sequences and with the tree recognized by the APG III. Additionally to resolving phylogenetic relationships among plants, studies based on this gene may also highlight the different evolutionary pathways followed by the land plant groups.

Keywords: Nuclear gene, flowering plants, systematics, gibberellin

INTRODUCTION

The large amount of DNA sequences generated in the last decade for an increasing number of different species has made possible to refine the phylogenetic relationships among flowering plants and enabled the generation of well-resolved classifications of this group (APG 2009). Despite this progress, some undefined or weak supported relationships still remain and additional molecular data are needed to increase support for these relationships. Thus, further progress in plant phylogenetics depends, in part, on identifying more phylogenetically useful loci.

In general, plant molecular phylogenetics has been very dependent on nucleotide sequences of chloroplast (cpDNA) and nuclear ribosomal (rDNA) genes, but the importance of using additional new nuclear sequences has been proposed as an alternative to improve the resolution of phylogenetic relationships (Prychitko & Moore 1997; Qiu *et al.*, 1999; Soltis *et al.*, 1999; Small et al., 2004).

Chaw *et al.* (2000) studied seed plants phylogeny using nuclear (*nuSSU* rDNA), mitochondrial (*mtSSU* rDNA) and chloroplast (*rbcL*) DNA sequences and obtained congruent topologies for individual gene trees as well as for the combined dataset of three genes, all with high bootstrap support. Soltis *et al.* (2000) reported a well resolved and highly supported topology of the angiosperms phylogeny by combining chloroplast and nuclear sequence data sets (*rbcL, matK* and *18S* rDNA) and suggested that most of the remaining large-scale phylogenetic questions could best be addressed by sequencing additional genes, without the need of adding more taxa in the analysis.

Although species phylogenies can be inferred from a single gene tree, it is advantageous to validate this estimation using independent gene trees. However, the identification of novel phylogenetic markers is not an easy task. Phylogenetic markers must reach at least 500 bp in length and should neither be much conserved nor much variable. Since introns and intergenic regions are usually excessively variable to be used as informative markers above intra-specific level, the evaluation of exon regions may be a valuable alternative for phylogenetic reconstructions above the species level (e.g.

Belink *et al.*, 2012). In addition, attributes as the elevated rate of sequence evolution, existence of multiple independent loci and the biparental inheritance make nuclear genes a very attractive alternative for estimating species trees (Small *et al.*, 2004). Following this rationality, a phylogenetic investigation based on nuclear genes should start by selecting the candidate gene for a preliminary study (Small *et al.*, 2004), in which the utility of such DNA sequence can be evaluated.

The Gibberellin 20-oxidase (*GA20ox*) is a low-copy nuclear gene (about four copies in flowering plants genome) related to the activation of the plant growth hormone gibberellin (Hedden and Phillips 2000) and seems to be relatively conserved across the plants (Huerta *et al.*, 2009). Although there is no grounds to expect that any particular gene will be universally useful at any given phylogenetic depth (Small *et al.*, 2004), the *GA20ox1* gene may match the main features needed by a phylogenetic useful nuclear gene. The aim of the present study was to investigate the efficacy of the nuclear gene *GA20ox1* in resolving phylogenetic relationships within the core eudicots group, using nucleotide sequences deposited in the GenBank.

MATERIAL AND METHODS

Initially an exhaustive search was performed in the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/genbank/) in order to find all sequences deposited for the gene *GA20ox1*. Sequences of 43 species representing 18 eudicots families were recorded in the GenBank (Table 1). Sequences of the *GA20ox1* gene from the moss *Physcomitrella patens* (Funariaceae), the lycophyte *Selaginella moellendorffii* (Selaginellaceae) and the monocot *Oryza sativa* (Poaceae) were also included in the analysis as outgroups. For three species (*Rosa wichurana*, *Brassica rapa* and *Datura ferox*) just partial sequences of the gene were deposited in the GenBank, sizing 884, 428 and 344 bp respectively. The length of the other sequences ranged from 1011 to 1953 bp (Table 1). Sequences were aligned using the software Muscle (Edgar 2004).

Table 1: Nucleotide content (%), sequence size (in base pairs) and the gene ID (NCBI) of the analysed sequences of the *GA20ox1* gene.

Family /Species	%T	%C	%A	%G	Size	Gene ID
AMARANTHACEAE						
Spinacia oleracea	26.0	21.3	31.7	20.9	1313	1144389
APIACEAE						
Daucus carota	27.4	21.7	30.5	20.4	1380	50428326
ASTERACEAE						
Chrysanthemum morifolium	25.7	22.5	30.9	21.0	1131	190192209
Helianthus annuus	26.8	19.9	32.6	20.7	1384	187455573
Lactuca sativa	26.9	21.0	31.9	20.2	1440	4164140
BRASSICACEAE						
Arabidopsis lyrata	25.8	25.9	28.4	19.9	1153	297799453
Arabidopsis thaliana	26.1	25.3	28.0	20.6	1134	145344084
Brassica rapa	25.2	29.7	24.5	20.6	428	27804386
CARYOPHYLACEAE						
Dianthus caryophyllus	26.9	19.1	32.4	21.5	1407	189409354
CHENOPODIACEAE						
Beta vulgaris	28.7	18.5	32.9	19.9	1382	115361479
CONVOLVULACEAE						
Ipomoea nil	22.2	28.2	28.8	20.8	1494	303303655
CUCURBITACEAE						
Cucurbita maxima	28.6	21.7	28.7	20.9	1442	27124555
EUPHORBIACEAE						
Ricinus communis	27.0	21.0	30.9	21.1	1137	255539616
FABACEAE						
Acacia mangium	27.1	20.3	30.9	21.8	1328	160623442
Glycine max	27.9	21.0	28.2	22.8	1038	351723112
Medicago truncatula	30.9	16.7	30.0	22.5	1011	357444170
Phaseolus vulgaris	26.8	23.8	29.0	20.4	1402	2108431
Pisum sativum	30.1	16.5	34.5	18.9	1381	1848145
FAGACEAE						
Castanea mollissima	25.9	22.2	31.0	20.9	1222	365176181
Fagus sylvatica	25.8	20.6	33.2	20.5	1469	18496056
LINDERNIACEAE						
Torenia fournieri	24.3	23.7	29.3	22.7	1340	323098311
MALVACEAE						
Gossypium hirsutum	27.1	21.5	29.6	21.8	1363	222875433
POLYGONACEAE						
Rumex palustris	24.0	24.6	28.3	23.1	1200	109729786
ROSACEAE						
Fragaria ananassa	23.8	22.5	28.8	24.9	1401	77632795
Prunus dulci	25.2	23.0	29.1	22.7	1177	390013401
Pyrus communis	23.2	24.1	29.2	23.5	1179	333440996
Rosa wichurana	23.6	24.3	26.7	25.3	884	256772627
RUTACEAE						
Citrus sinensis x Poncirus trifoliata	27.0	18.3	32.4	22.4	1561	8919864
SALICACEAE						
Populus alba	26.9	23.1	29.2	20.8	1158	34013373
Populus nigra	27.5	22.6	29.5	20.4	1158	28316357
Populus simonii	27.2	21.3	29.9	21.6	1279	233142141

Populus tomentosa	26.9	23.2	29.6	20.3	1158	40233166
Populus tremula x P. tremuloides	26.5	19.9	33.1	20.5	1584	15384972
Populus trichocarpa	27.8	20.9	31.3	19.9	1158	224063946
Populus trichocarpa x P. deltoides	27.3	22.2	30.6	19.9	1485	118489892
SOLANACEAE						
Capsicum annuum	30.5	20.4	30.3	18.8	1662	326581982
Datura ferox	26.2	19.8	30.8	23.3	344	17225008
Nicotiana tabacum	27.4	21.6	30.9	20.2	1140	30519872
Nicotiana sylvestris	28.2	20.8	32.3	18.7	1724	20149238
Solanum dulcamara	26.7	19.9	33.4	19.9	1475	7328336
Solanum lycopersicum	29.9	19.4	30.9	19.7	1267	350538088
Solanum tuberosum	29.2	17.8	33.0	20.0	1506	10800973
VITACEAE						
Vitis vinifera	24.8	23.9	27.1	24.2	1149	99032730
OUTGROUP						
Physcomitrella patens	26.9	26.0	23.2	23.9	1953	168012844
Selaginella moellendorffii	23.5	26.9	22.9	26.7	1295	159902522
Oryza sativa	17.0	29.9	20.9	32.3	1149	115462222

In order to characterize the *GA20ox1* sequence, the transition/transversion matrix (using the maximum likelihood method) was estimated using Mega 5.05 (Tamura *et al.*, 2011). The number of parsimony informative sites, the consistency index (CI) and the retention index (RI) were computed using PAUP 4.0 (Swofford 1998), with the Tree-Bisection-Regrafting (TBR) algorithm, all characters unordered and gaps treated as missing data.

The phylogenetic relationship among species was analyzed using the maximum likelihood and a Bayesian inference approaches. The maximum likelihood tree was built in Mega 5.05 (Tamura *et al.*, 2011) using the GTR+G mutation model with rate variation among sites modeled by a discrete gamma distribution with four categories (Tavaré 1986), as determined through the Akaike Information Criterion (AIC) in the software jModelTest 0.1.1 (Posada 2008). A bootstrap analysis with 500 replicates was employed to assess the internal support of the groups.

The Bayesian analysis was performed in MrBayes (Huelsenbeck and Ronquist 2001) using the default settings of the program corresponding to the GTR+G+Γ model and estimating base frequencies from the data. The Markov chain was run for 400000 generations sampling every 100th tree. A 'heated' chain was run in order to reduce the risk of the first chain getting fixed on a local optimum by discarding the first 20000 generations as a burn-in period. Two independent analyzes were performed.

Aiming to compare the phylogenetic signal of the *GA20ox1* gene with the phylogenetic relationship retrieved with plastid genes traditionally employed in such studies with the parsimony method sequences of the chloroplast region *rbcL* deposited in the GenBank were recorded for each of the 43 species representing the same taxa or at least the same genera recorded for the former gene. *rbcL* sequences were aligned and analyzed using the maximum likelihood approach as described above. The number of parsimony informative sites, the CI and the RI were computed as for the *GA20ox1* sequences.

RESULTS AND DISCUSSION

Sequence variability and informative capacity of the *GA20ox1* gene

Transversions are considered the more reliable type of mutations in constructing phylogenies (Quicke 1993) and are therefore, an important aspect in the informative capacity of DNA regions for phylogenetic analyses. Concerning the patterns of nucleotide substitution, 39.27% of the mutations retained in the *GA20ox1* sequences evaluated are transversions, with a ratio of transition to transversion equal to 1.5 (Table 2). This amount of tranversional mutations seems to guarantee a useful phylogenetic signal in this gene, making it a promissory tool for highlighting phylogenetic relationships among flowering plants.

Table 2: Maximum Likelihood Estimate of the Pattern of Nucleotide Substitution. Substitution pattern and rates were estimated under the General Time Reversible model using a discrete Gamma distribution to model evolutionary rate differences among sites (GTR+G). Rates of different transitional substitutions are shown in bold and those of transversional substitutions are shown in italics.

	A	**T**	**C**	**G**
A	-	*5.35*	*5.31*	**12.18**
T	*6.04*	-	**14.31**	*2.82*
C	*7.22*	**17.27**	-	*4.47*
G	**16.97**	*3.48*	*4.58*	-

For the parsimony analysis of the *GA20ox1* sequences (tree not shown), 1195 out of 3174 (37.65%) characters were parsimony informative, generating a consensus tree with consistency index CI = 0.342 and retention index RI = 0.367. For the *rbcL* sequences, 658 out of 2910 (22.61%) characters were parsimony informative, generating a consensus tree with consistency index CI = 0.553 and retention index RI = 0.528.

In a large analysis of the angiosperms phylogeny, Soltis *et al.* (2000) evaluated a combined dataset of three genes (*atpB*, *rbcL* and 18S *rDNA*) and obtained a parsimony tree with CI = 0.12 and RI = 0.59. Evaluating the informative capacity of *matK* and *trnK*5' regions for phylogenetic studies of the early diverging eudicots, Hilu *et al.* (2008) found 64% and 55% of parsimony informative characters respectively, with CI = CI = 0.36 and RI = 0.44 for the *matK* sequences and CI = 0.39 and RI = 0.47 for the *trnK*5'. Although the proportion of parsimony informative characters of the *GA20ox1* sequences is lower than those observed for *matK* and *trnK*5' regions, the consistency and retention indexes, two measures of the homoplasy in the data set, are just slightly lower, suggesting equivalent reliability of the *GA20ox1* sequences for phylogenetic analysis, despite the low values.

Phylogenetic signal of the *GA20ox1* sequences

For both approaches of phylogenetic inference, the maximum likelihood (Figure 1) and Bayesian (Figure 2) approaches, the *GA20ox1* sequences provided enough phylogenetic signal to resolve relationships at the family levels within the core eudicots with high support based on bootstrap (>99%) and posterior probability (> 0.96). At this level, the *GA20ox1* phylogenetic reconstruction was congruent with the phylogenetic reconstruction based on the *rbcL* sequences (Figure 3) and with the tree recognized by the APG III (APG 2009). Even the species with partial sequences of the *GA20ox1* were placed within their respective botanical families.

For both, *rbcL* and *GA20ox1* analyses, the nine families from superorder rosiid formed a monophyletic group, while the six families from superorder asteriid are recovered as a paraphyletic group (Figures 1 and 2). Concerning the four families from order Charyophyllales, the *rbcL* based phylogeny recovers a monophyletic group, while the *GA20ox1* based tree represent this families as a paraphyletic group in both, the maximum likelihood (Figure 1) and Bayesian (Figure 2) analyses.

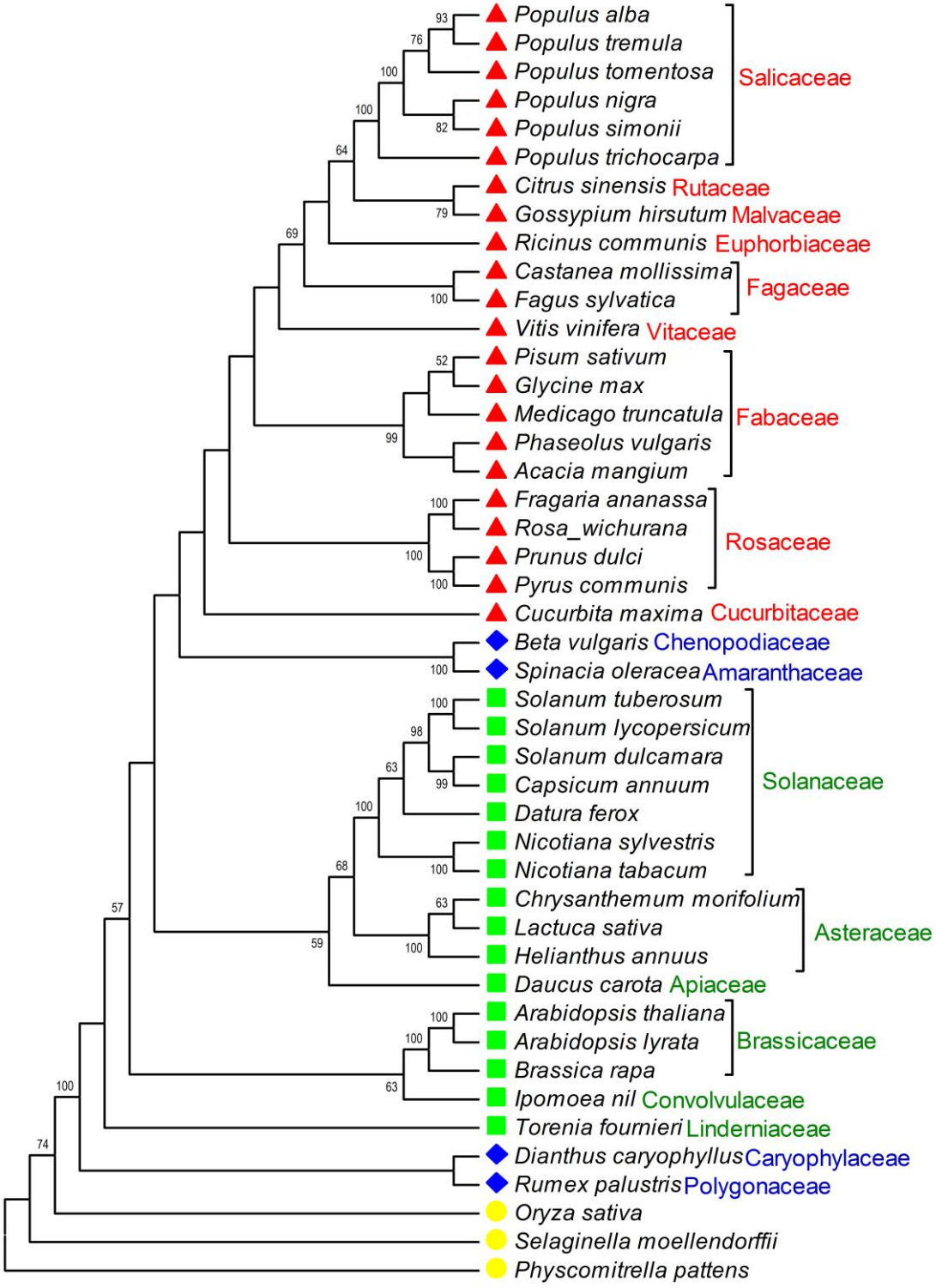


Figure 1: Maximum likelihood phylogenetic tree for species of the core eudicot based on the *GA20ox1* sequences. Species from the superorder rosiid are preceded by a red triangle, from superorder asteriid are preceded by a green square and from order Charyophyllales are preceded by a blue diamond. Outgroup species are preceded by a yellow circle. The corresponding families are given after the species names. Values at the nodes represent the bootstrap support.

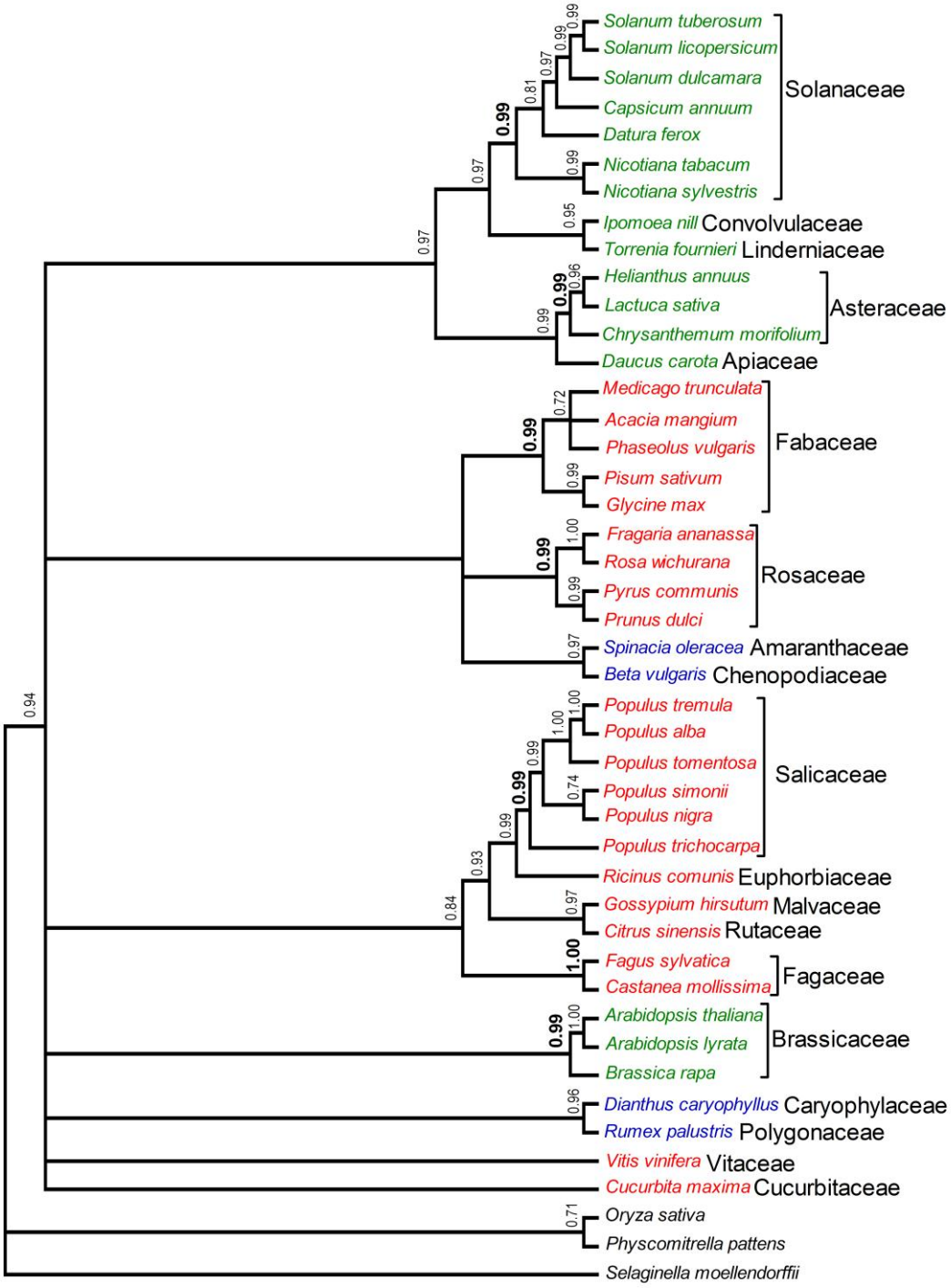


Figure 2: Bayesian majority-rule consensus tree for the *GA20ox1* sequences. Values of posterior probability are presented over the branches. Species from the superorder rosiid are typed in red, from superorder asteriid are typed in green and from order Charyophyllales are typed in blue. Outgroup species are typed in black. The corresponding families are given after the species names. Values at the nodes represent the posterior probability support. Posterior probabilities of family clades are presented in bold.

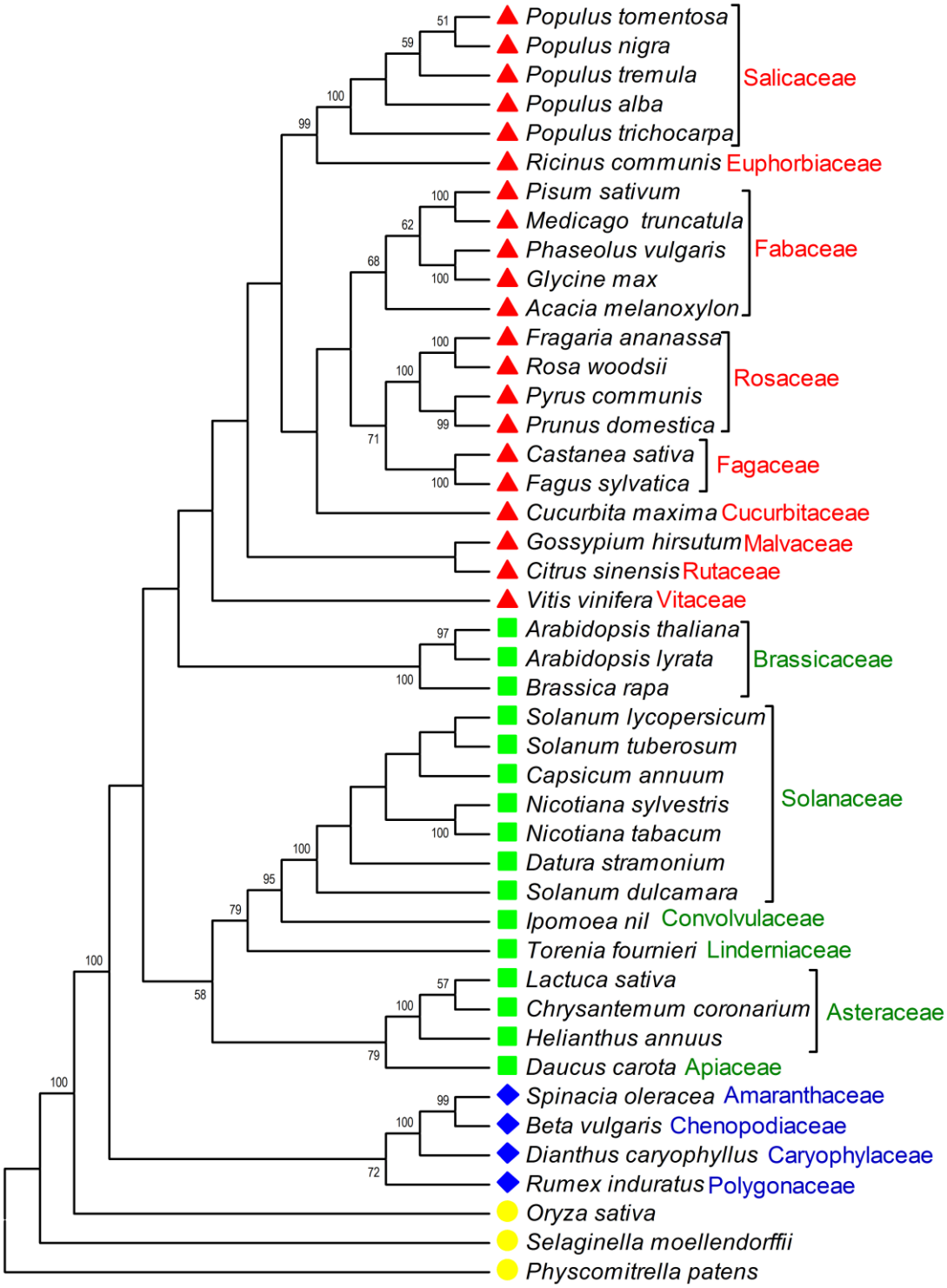


Figure 3: Maximum likelihood phylogenetic tree for species of the core eudicot based on the *rbcL* sequences. Species from the superorder rosid are preceded by a red triangle, from superorder asteriid are preceded by a green square and from order Charyophyllales are preceded by a blue diamond. Outgroup species are preceded by a yellow circle. The corresponding families are given after the species names. Values at the nodes represent the bootstrap support.

Considering that a gene tree is the phylogeny of a particular DNA sequence, viewing the alleles themselves as the operational taxonomic units (OTUs) and not the evolutionary pathway of a group of OTUs (Avise 1989), the present analysis of the *GA20ox1* reflects the phylogenetic relationships of this gene across the recorded species.

The *GA20ox1* gene codifies the enzyme GA20-oxidase 1, which is involved in the gibberellic acid metabolism by catalyzing the stepwise conversion of the C_{20} gibberellins, GA_{12}/GA_{53}, by three successive oxidations to GA_9/GA_{20}, which are the immediate precursors of the active gibberellins, GA_4 and GA_1, respectively (Eriksson *et al.*, 2000). Considering the different patterns of plant growth across species and the regulatory role of gibberellins in this physiological feature, phylogenetic studies based on this gene may reveal patterns of lineage sorting related to plant evolution of this characteristic. In this sense, the comparative analysis of different gene-phylogenies may help researchers to understand the evolution of the land plants, as demonstrated by Willyard *et al.* (2009) for ponderosa pines.

CONCLUDING REMARKS

The present results demonstrated that the *GA20ox1* sequences revealed high phylogenetic signal for grouping species within their respective families with high support. Although much of the needed effort in phylogenetics is towards resolving relationships at higher taxonomic levels, the *GA20ox1* can be considered a useful nuclear coding region able to resolve phylogenetic relationships at lower taxonomic levels, the level that most systematists work (Small *et al.*, 2004). The phylogenetic analysis of this gene may also highlight the different evolutionary pathways followed by the land plant groups.

ACKNOWLEDGEMENTS

This work was partially supported by financial support of FAPERGS (Process 1013351) and CNPq (Process 471812/2011-0). The authors thank CNPq and CAPES for grant and scholarship. Thanks are due to Dr. J. R. I. Ribeiro and Dr. J. Boldo for the insightful discussions.

REFERENCES

APG III (2009) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APGIII. Bot J Linn Soc 161:105–121

Avise JC (1989) Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution, 43:1192-1208

Belinky F, Szitenberg A, Goldfarb I, Feldstein T, Wörheide G, Ilan M, Huchon D (2012) ALG11 – A new variable DNA marker for sponge phylogeny: Comparison of phylogenetic performances with the 18S rDNA and the COI gene. Mol Phyl Evol 63:702–713

Chaw SM, Chang CC, Chen HL, Li WH (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58:424-441

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32**:**1792-1797

Eriksson ME, Israelsson M, Olsson O, Moritz T (2000) Increased gibberellins biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. Nature 18:784-788

Hedden P, Phillips AL (2000) Gibberellin metabolism: new insights revealed by the genes. Trends Plant Sci 5:523-530

Hilu K W, Black C, Diouf D, Burleigh J G (2008) Phylogenetic signal in *matK* vs. *trnK*: A case study in early diverging eudicots (angiosperms). Mol Phylogenet Evol 48:1120-1130

Huerta L, Garcia-Lor A, Garcia-Martinez JL (2009) Characterization of gibberellins 20- oxidases in the citrus hybrid Carrizo citrange. Tree Physiol 29:569-577

Posada D (2008) jModelTest: Phylogenetic Model Averaging. Mol Biol Evol 27:1253-1256

Prychitko TM, Moore WS (1997) The utility of DNA sequences of an intron from the b-fibrinogen gene in phylogenetic analysis of woodpeckers (Aves: Picidae). Mol Phylogenet Evol 8:193-204

Qiu YL, Lee J, Quadroni FB, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainenk V, Chasek MW (1999) The earliest angiosperms: evidence frommitochondrial, plastid and nuclear genomes. Nature 402:404-407

Quicke DLJ (1993) Principle and techniques of contemporary taxonomy. Chapman & Hall, Glasgow, London UK pp 311

Small RL, Cronn RC, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. Aust Syst Bot 17:145-170

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-404.

Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoop SB, Fay MF, Axtell M, Swensen SM, Prince LM, Kress JW**,** Nixon KC, Farrise JS (2000) Angiosperm phylogeny inferred from 18s rDNA, *rbcL*, and *atpB* sequences. Bot J Linn Soc 133:381-461

Swofford DL (1998): PAUP*: Phylogenetic analysis using parsimony and other methods. Illinois Natural History Survey, Champaing, IL.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

4 CONCLUSÃO

Avaliando os resultados obtidos sobre as relações filogenéticas entre uma amostra de espécies de eudicotiledôneas, através das sequências do gene nuclear *GA20ox1*, pode-se considerar este gene uma região nuclear útil para este tipo de estudo, sendo capaz de resolver as relações filogenéticas em níveis taxonômicos inferiores. O grupo das eudicotiledôneas é composto por cerca de 300 famílias e apresenta aproximadamente 1650 espécies, neste estudo foram utilizadas as sequências de 43 espécies disponíveis no *GenBank*, correspondendo a 18 famílias. Apesar da baixa amostragem, a análise possibilitou avaliar a capacidade informativa deste gene. Além disso, a análise desse gene sugere ser possível avaliar os diferentes caminhos evolutivos seguidos pelos grupos de plantas terrestres. Essas informações servirão de subsídio para o desenvolvimento de futuras pesquisas visando avaliar as relações filogenéticas em diferentes grupos de plantas, com base neste gene nuclear.

5 REFERÊNCIAS BIBLIOGRÁFICAS

APG. An ordinal classification for the families of flowering plants. **Annals of the Missouri Botanical Garden**. v. 85, p. 531-553, 1998.

APG II. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APGII. **Bot J Linn Soc**.v.141, p. 399–436, 2003.

APG III. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APGIII. **Bot J Linn Soc**.v.161, p. 105–121, 2009.

BAXEVANIS, A. D., OUELLETTE, B. F. F. **Bioinformatics**: A practical guide to the analysis of genes and proteins. Ed. Wiley-interscience. 2ed. 2001.

COLES, J. P. et al. Modification of gibberellin production and plant development in Arabidopsis by sense and antisense expression of gibberellin 20-oxidase genes. **The Plant Journal**. p. 547-556, jan. 1999.

DAYHOFF, M.O. et al. **Atlas of Protein Sequence and Structure**. Vol. 1. Maryland, EUA: National Biomedical Research Foundation/Silver Spring, 1965.

DDBJ. Disponível em: <http://www.ddbj.nig.ac.jp/> Acesso em: 21 out. 2012.

DEVEREUX, J., HAEBERLI, P., SMITHIES,O. A comprehensive set of sequence analysis programs for the VAX. **Nucleics Acid Res**. v.12, p. 387-395, 1984.

DUVALL, M. R. et al. Phylogenetic analysis of rbcL sequences identifies *Acorus calamus* as the primal extant monocotyledon. **Proc. Natl. Acad. Sci**. v. 90, p. 4641-4644, 1993.

EMBL. Disponível em: <http://www.ebi.ac.uk/embl/> Acesso em: 21 out. 2012.

ERIKSSON, M.E. et al. Increased gibberellin biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. **Nature Biotechnology**. v. 18. p. 784-788, jul. 2000.

GENBANK. Disponível em: <http://www.ncbi.nlm.nih. gov/Genbank/> Acesso em: 20 out. 2012.

HASEBE, M. et al. *rbcL* gene sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. **Proc. Nail. Acad. Sci**. v. 91, p. 5730-5734, 1994.

HEDDEN, P.; KAMIYA, Y. Gibberellin Biosynthesis: Enzymes, Genes and Their Regulation. **Annual Review Planta Physiology Molecular Biology**, n. 48, p. 431-460, 1997.

HEDDEN, P., PHILLIPS, A.L. Gibberellin metabolism: new insights revealed by the genes. **Trends Plant Sci** v. 5, p. 523-530, 2000.

HEENAN, P. et al. Phylogenetic analyses of ITS and *rbcL* DNA sequences for sixteen genera of Australian and New Zealand *Brassicaceae* result in the expansion of the tribe *Microlepidieae*. **Taxon**. v. 61, p. 970-979, 2012.

HENNIG, W. Grundzüge einer Theorie der phylogenetischen Systematik. Berlin: Deutscher Zentralverlag, 1950. **Phylogenetic systematics**. Urbana: University of Illinois Press, 1966.

HUERTA, L. et al. Characterization of gibberellin 20-oxidases in the citrus hybrid *Carrizo citrange*. **The Physiology**, n. 29, p. 569-577, jan. 2009.
KAMIYA, Y.; GARCÍA-MARTÍNEZ, J. L. Regulation of gibberellin biosynthesis by light. **Current Opinion in Plant Biology**, n. 2, p. 398-403, 1999.

LEE, D.J.; ZEEVAART, J.A.D. Regulation of gibberellin 20-oxidasel expression in spinach by photoperiod. **Planta**. 236, p. 35-44, 2007.

MADDISON, W. P. et al. Outgroup Analysis and Parsimony. **Oxford Journals**. v. 33, p. 83-103, 1983.

PUSTELL, J., KAFATOS, F. C., A high speed, high capacity homology matrix: zooming through SV40 and polyoma. **Nucleic Acids Res.** v.10, p.4765-4782, 1982.

RIEU I et al. The gibberellin biosynthetic genes AtGA20ox1 and AtGA20ox2 act, partially redundantly, to promote growth and development throughout the Arabidopsis life cycle. **The Plant Journal**. v. 53, 488–504, 2008.

SANTOS, C. M. D., CALOR, A. R. Using the logical basis of phylogenetics as the framework for teaching biology. **Papéis Avulsos de Zoologia**. v. 48, p. 199-211, 2008.

SAVOLAINEN, V. et al. Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid *atpB* and *rbcL* Gene Sequences. **Sysl. Biot**. v. 49, p.306-362, 2000.

SCHUH, R. T., BROWER, A. V. Z. **Biological systematics: principles and applications**. New York: Cornel University Press, 2009.

STADEN, R. Sequence data handling by computer. **Nucleic Acids Res**. v. 4, p. 4037-4051, 1977.

STOESSER, G. et al. The EMBL Nucleotide Sequence Database. **Nucleic Acids Res.**, v.30, p. 21–26, 2002.

SWISS-PROT. Disponível em: <http://www.expasy.ch/sprot/> Acesso em: 21 out. 2012.

WHEELER, D. L. et al. Database resources of the national center for biotechnology information. **Nucleic Acid Res**. v.35, D5–D12. 2007.

WOLFE, K. H., LI, W.H. SHARP, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. **Proc. Natl. Acad. Sci.** v. 84, p. 9054-9058, 1987.

UNIPROT. Disponível em: <http://www.expasy.uniprot.org/> Acesso em: 21 out. 2012.