Molecular Studies in Selected Oak Species: What Have We Learned About Oak Evolution?

Lucia Vazquez
Biology Department,
University of Illinois at Springfield, Springfield, IL 62703,
phone 217-206-7337, fax 217-206-6162
e-mail vazquez.lucia@uis.edu

INTRODUCTION

The genus *Quercus* is the largest in the family Fagaceae and consists of approximately 500 species of trees and shrubs. Two subgenera, *Cyclobalanus* and *Quercus,* have been recognized based on morphological characters, with *Cyclobalanus* native to Asia and *Quercus* distributed in tropical and subtropical regions of the Northern Hemisphere. Subgenus *Quercus* has been further subdivided into sections *Quercus* (white oaks), *Lobatae* (red and black oaks), and *Protobalanus* (intermediate oaks) with the two latter sections endemic to the Americas, and section *Quercus* found in both the Old- and the New World (Nixon, 1993).

Taxonomic problems in oaks still persist due to their extensive foliar polymorphism, convergence, and occurrence of hybridization and introgression. Despite these challenges, good progress has been made in the morphological characterization of oak species through the publication of local and regional floras (e.g. Flora of California, Flora of North America, and Flora of China). Morphological and molecular studies of representative species in the sections of subgenus *Quercus* have also been conducted and have provided insight into the phylogenetic relationships at the intersectional level (Nixon, 1985, 1993; Manos et al., 1999); however, these two character sources (i.e. morphological and molecular) have not yielded enough information to understand the phylogenetic relationships of oaks at the species level. Detailed phylogenetic analyses, at the intraspecific level and based on morphological characters, have not been conducted due to the limited number of vegetative and reproductive characters available (Manos, 1999; Vazquez 2001). Consequently, researchers of oak systematics have turned to DNA sequences as possible sources of additional characters; nonetheless, searches in genbank as well as in the literature show that only a limited number of genes in oaks have been sequenced.

The intent of this paper is to present the advances in the research of molecular variation in oaks, as well as to assess the amount of nucleotide variation in nuclear and organelle genes from oak species.

METHODS

Two general types of methods were used: 1) analysis of oak sequences available in genbank (a public database: http://www.ncbi.nlm.nih.gov/, and 2) analysis of data obtained in the laboratory.

Table 1. Variation of chloroplast genes in oak species

Gene name	Region analyzed	Nucleotide Sequence divergence
rbcL	Coding region	0 – 0.0122
matK	Coding region	0.0007-0.0135
rps16	intron	0 to 0.0273
trnL	intron	0 to 0.0279
trnL-trnF	intergenic spacer	0 to 0.0586

Table 2. Variation of nuclear genes/regions in oak species

Gene name	Region analyzed	Nucleotide Sequence divergence
Cinnamyl alcohol dehydrogenase.	Coding region	0.0328
Legumin	Intron	0 to 0.0540
Legumin	Exon	0 to 0.0455
Glyceraldehyde 3-phosphate dehydrogenase (GA3-P).	Intron and exon	0 to 0.0183
Internal transcribed spacers (ITS) of the nuclear ribosomal DNA.	Internal transcribed spacers and 5.8S	Large paralogue- 0.0530 Medium paralogue-0.0230 Small paralogue-0.0130

For the first method, representative oak sequences available in genbank were downloaded and aligned using the default parameters in Clustal W (Thompson et al. 1994). When necessary, sequence alignment editing was carried out with the program Seaview (Galtier et al., 1996). Estimation of distances was performed with the program Phylip (Felsenstein, 1989) using the Kimura 2-parameter for DNA sequences.

The second method for the analysis of nuclear genes began with the collection of plant material in several localities throughout Mexico. Consequent lab work included isolation of total genomic DNA, design of primers for the amplification of nuclear genes, isolation of genes using polymerase chain reaction (PCR), cloning of PCR products, DNA sequencing, DNA editing, and DNA analysis. The

genes/regions studied in the laboratory include the internal transcribed spacer of the nuclear ribosomal genes (ITS), introns of the nuclear genes *leafy*, *agamous*, *pistillata*, *slg*, and the intron of the chloroplast gene *rps16*. Phylogenetic analyses were conducted with Winclada 2.0 (Nixon, 1999); only informative characters where analyzed and gaps were coded following the simple indel coding method of Simmons and Ochoterena (2000).

RESULTS

Mitochondrial genes. Only four oak mitochondrial genes were found in genbank: orfx, ribosomal protein small 3 (rps3), maturase (matR), and NADH dehydrogenase subunits 4 & 7, and their introns 3 & 2, respectively. Only the last two genes had at least two sequences to allow some comparisons. DNA sequence alignment of the mitochondrial gene matR in *Quercus engleriana* and *Q. multinervis* showed that the coding sequence of this gene is 1743 base pairs (bp) long and shows no nucleotide differences in these species. The third intron of the mitochondrial gene NADH dehydrogenase subunit 4 in *Q. glauca* is 541 bp long and revealed only 2 nucleotide substitutions; one of the four sequences analyzed showed an 8-bp gap. The second intron of the mitochondrial gene NADH dehydrogenase subunit 7 in *Q. glauca* is 702 bp long and displayed only two nucleotide substitutions.

Chloroplast genes. Sequences of the following oak chloroplast genes were found in genbank: ndhF intergenic region, ycf6-psbM intergenic spacer, NADH dehydrogenase subunit F, ATP synthase beta subunit, atpB-*rbcL* spacer region, trnD-trnT intergenic spacer, *matK*, *rbcL*, tRNA-Leu (trnL)-tRNA-Phe (trnF), and rps16. Of these genes, only the genes *rbcL*, *matK*, the intron and intergenic spacer of trnL-trnF, and rps16 intron, had more than two sequences to allow any comparisons and generalizations.

rbcL. Analysis of 32 *rbcL* protein sequences corresponding to 27 oak species available in genbank indicate that this protein is 482 amino acids long, and it showed differences at only 3 amino acid sites. The sequence analysis of 31 *rbcL* nucleotide sequences corresponding to 30 species, revealed sequence divergences ranging from zero to 0.0122. No inversions were detected in the sequences analyzed, and 2 indels ranging in size from 1 to 3 bp were detected in a sequence from *Quercus virginiana*.

matK. Analysis of 59 chloroplast maturase (*matK*) protein sequences from 29 oak species showed that this protein is 431 amino acids long and 1293 nucleotides long. At the protein level, these sequences show substitutions at 11 amino acid sites, one insertion that is 6 amino acids long, and genetic distances ranging from zero to 0.026. A comparison of *matK* nucleotide sequences from 12 oak species showed nucleotide substitution at 30 sites and distances ranging from 0.007 to 0.0135. This apparently higher level of protein sequence divergence compared to nucleotide divergence may be due to the larger number of protein sequences analyzed. Nevertheless, the levels of nucleotide sequence divergence are still low compared to nuclear genes.

trnQ-trnS, rps16 & trnL-trnF. A comparison of eight aligned *trnQ-trnS* oak sequences available in genbank also revealed low amounts of sequence variation with only one nucleotide substitution and a few deletions. Vazquez (2001) sequenced the *rps16* intron, and the intron and intergenic spacer of the *trnL-trnF* gene from several Mexican red oak species. Analysis of the sequences generated

by Vazquez (2001) and those available in genbank revealed that the sequence divergence of the rps16 intron ranges from zero to 0.0270. Similarly, analysis of the *trnL-F* sequences showed that the nucleotide sequence divergence for the intron ranges from zero to 0.0279, while the sequence divergence of the intergenic spacer varies from zero to 0.0586.

Oak nuclear genes and regions

As in the case of chloroplast genes, only a limited number of nuclear genes have been studied in oaks. The sequences of nuclear genes available in genbank included in this study are *cinnamyl alcohol dehydrogenase*, *legumin*, *glyceraldehyde 3-phosphate dehydrogenase* (GA3-P), and the internal transcribed spacers (ITS) of the ribosomal genes. Ninety-five additional sequences of ITS generated by the author of this paper were added to the analysis, as well as several sequences from the transposon En/Spm. *Cinnamyl alcohol dehydrogenase*. The gene cinnamyl alcohol dehydrogenase in *Quercus* is 326 amino acids long and their corresponding DNA sequences are 978 nucleotides long. Alignment of two protein sequences from *Q. ilex* and *Q. suber* showed amino acid differences at 12 sites with most of these changes being conserved and semiconserved. Analysis of the corresponding nucleotide sequences shows differences at 38 nucleotide sites including two 3-bp insertions, and a nucleotide sequence divergence of 0.0328.

Legumin. Examination of 13 *legumin* oak sequences from genbank indicates that this nuclear gene consist of an intron 105 bp long and an exon 366 bp long. Aligned sequences showed 30 nucleotide substitutions with 9 of them located in the intron and the remaining 21 in the coding region of the exon. Two of the examined sequences differ from the rest by the presence of a 5-bp insertion. The nucleotide divergences in the *legumin* intron range from zero to 0.0540, and in the *legumin* exon they vary from zero to 0.0455.

Glyceraldehyde 3-phosphate dehydrogenase (GA3-P). The oak sequences available in genbank consist of 3 noncoding regions and two exons. Comparisons with rosid GA3-P sequences suggest that these oak coding sequences correspond to exons 8 and 9. The beginning of the oak GA3-P sequences available in genbank consists of a noncoding region that is 80 bp long. A noncoding region 408 bp long is located between exons 8 and 9; downstream of exon 9 there is a third noncoding region 163 bp long. Alignment of oak GA3-P sequences revealed that most of the 45-nucleotide substitutions observed occur in the longest noncoding region located between exons 8 and 9. Furthermore, some of these sequences have a 3-bp insertion and a 11-bp insertion. The average nucleotide divergence of this region ranges from zero to 0.0183.

Internal transcribed spacers (ITS) of the nuclear ribosomal DNA. Fifty-three ITS sequences were downloaded from genbank, and 50 additional sequences were generated by the author. It is worth mentioning that although over 100 oak ITS sequences are available in genbank, only complete sequences of different species were selected for this analysis. In contrast to the genes discussed above, the oak ITS region has been previously studied by several authors. A phylogenetic study of the ITS region by Manos et al. (1999) supported the previously recognized monophyletic sections *Quercus* (white oaks *sensu stricto*) and *Lobatae* (red oaks). Manos et al. (1999) also showed that the Eurasian white oaks form a monophyletic group different from *Quercus sensu stricto*.

In a subsequent study, Vazquez (2001) focused on the red oaks and sampled 30 species mainly native to Mexico. In contrast to Manos et al. (1999), the results of Vazquez (2001) showed that there are at least 3 different ITS paralogues of sizes 613 bp (large paralogue), 600 bp (medium paralogue), and 497 bp (small paralogue). Furthermore, sequence divergence of the ITS region ranged from 0.013 in the small paralogue to 0.053 in the large paralogue; the average divergence of the medium-size paralogue was 0.023. Interestingly, intraspecific sequence divergences ranged from zero to 0.29. These results suggest that the ITS region in the Mexican red oak taxa examined has not undergone full concerted evolution. The use of the ITS region, therefore, is unlikely to yield information about the phylogenetic relationships of Mexican red oak taxa.

Other nuclear regions under study. In the search for genomic regions that could provide phylogenetic information for reconstructing oak phylogenies, Vazquez (unpublished) has focused on the study of the introns from the nuclear genes pistillata, cauliflower, and leafy, as well as the DNA sequence variation of the En-Spm transposons. Vazquez's preliminary results (unpublished) show that there are two to four copies of the *pistillata* intron, depending on the species, with sizes of 800-900 bp long. The second intron of *leafy* seems to consist of at least two copies of sizes 700 and 800 bp. The intron of *agamous* is about 2000 bp long and of all the nuclear regions examined, appears to be the only one that is a single copy gene.

Transposons, also called jumping genes, are very important in genome structure and function (Bennetzen, 2000). Given that transposons have never been studied in oaks, Vazquez (unpublished) has isolated and sequenced En/Spm transposons from selected *Quercus* (oak) species using a PCR-based approach. The results show that the oak En/Spm transposons sequences show conserved regions, as well as highly variable regions. A preliminary phylogenetic analysis of these oak En/Spm sequences suggests that more than one copy of this transposon may be present in the oak genome.

DISCUSSION AND CONCLUSIONS

Analysis of the oak mitochondria genes available in genbank showed low nucleotide variation rates. This result is in agreement with other angiosperms studies, which have concluded that the variation rates of mitochondrial sequences are "roughly 4 times slower than in land plant chloroplast DNA (cpDNA)" (Palmer & Herbon, 1988).

Analysis of the oak chloroplast genes in this study indicates that the trnL-F intergenic spacer has a higher divergence level than the chloroplast introns examined. This result agrees with studies in other angiosperm taxa, which have shown that chloroplast intergenic spacers are more variable than chloroplast introns (Shaw et al. 2007).

Regarding the amount of nucleotide divergence in *matK*, previous studies have shown that this gene is more variable than *rbcL* (Hilu & Liang 1997); however, the amount of sequence divergence of these genes in oaks may suggest that is similar. Nevertheless, one has to keep in mind that the number of sequences analyzed for *matK* was smaller than those of *rbcL*, which may skew the results obtained in this analysis.

In general, chloroplast coding and noncoding sequences have been useful in the reconstruction of phylogenies at the genus and family levels in a variety of plant taxa (see Mummenhoff et al., 2001; Olmstead & Palmer, 1992; Hall et al., 2002); however, the use of chloroplast genes in oaks is unlikely to provide valuable information for phylogeny reconstruction at the interspecific level. The documented high levels of gene flow and reticulate histories in oak species make chloroplast data inappropriate for the reconstruction of phylogenetic relationships. Chloroplast genes are inherited maternally and if used in the phylogeny reconstruction of oaks, they would only provide partial information on the evolutionary history of the taxa sampled (Whittemore and Schaal, 1991). Despite the limitations for phylogenetic studies, chloroplast genes are suitable for the estimation of genetic diversity, as well as for phylogeographic studies, as has been shown in numerous studies (Ferris et al., 1993; Lumaret et al., 2002; Petit et al., 1993, 2002; Romero-Severson et al., 2003; Magni et al., 2005, among others).

The nuclear genes/regions examined showed sequence divergences ranging from 0.0183 in the *glyceraldehyde 3 phosphate dehydrogenase* gene, to 0.0540 in the *legumin* intron. Of the sequences compared, the coding region of the *Cinnamyl alcohol dehydrogenase* gene showed moderate sequence divergence with a value of 0.0328. Although the large copy of the ITS region also showed a relatively high level of nucleotide divergence with a value of 0.0530, the use of this molecule for phylogeny reconstruction of Mexican red oak taxa is not recommended due to the presence of paralogues.

Of the nuclear gene sequences examined, the most promising for providing phylogenetic information are the exon and intron of the legumin gene. However, before these genes are used, it must be verified that they are present as single copy genes in the oak genome. If present in low-copy numbers, then the orthologous sequences must be identified before carrying out any type of analysis.

In summary, mitochondrial genes are not suitable for reconstructing the phylogeny of oak species due to their low rate of evolution and the frequency of structural arrangements, that have been documented in other angiosperms (Wolfe et al., 1987). Although chloroplast genes evolve at a faster rate than mitochondrial genes, their use in phylogeny reconstruction is not advisable as gene flow is very common in oak species, and it would only reveal the maternal parentage in oaks.

Regarding nuclear genes, there is a limited number of genes available with potential to recover the evolutionary history of oaks, such as the *legumin* and the *cinnamyl alcohol dehydrogenase* genes. Even though the ITS region has been used in the analysis of North American, European, and Asian oak species, its use should be avoided in Mexican taxa due to the presence of paralogues. Despite the large number of gene entries in the public gene databases, the number of genes useful for phylogeny reconstruction in oaks remains very small. Uncovering the evolutionary history of oaks at the interspecific level will require the active collaboration of oak researchers to find additional and multiple single copy or low-copy genes with enough information to reconstruct the phylogenetic relationships of this challenging but yet interesting group of plants.

LITERATURE CITED

Bennetzen, J. L. 2000. Transposable element contributions to plant gene and genome evolution. Plant Molecular Biology 42 (1): 251-269.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

Ferris C, R. P. Oliver, A. J. Davy, G. M. Hewitt. 1993. Native oak chloroplasts reveal an ancient divide across Europe. Molecular Ecology 2:337-344.

Galtier, N., Gouy, M. & Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci., 12:543-548.

Hall, J. C., K. J. Sytsma and H. H. Iltis. 2002. Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. American Journal of Botany, 89:1826-1842.

Hilu, K. W. and H. Liang. The *matK* gene: sequence variation and application in plant systematics. Am. J. Bot. 84(6):830-839.

Mummenhoff K., H. Brüggemann. and J. L. Bowman. 2001. Chloroplast DNA phylogeny and biogeography of Lepidium (Brassicaceae). American Journal of Botany 88:2051-2063.

Lumaret et al. 2002. Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* L.). Molecular Ecology 11: 2327-2336.

Magni et al. 2005. Chloroplast DNA variation of *Quercus rubra* L. in North America and comparison with other Fagaceae. Molecular Ecology 14: 513-524.

Manos, P., J. J. Doyle, K.C. Nixon. 1999. Phylogeny, biogeography, and processes of molecular differentiation in *Quercus* subgenus *Quercus* (Fagaceae). Mol. Phyl. Evol. 12: 333-349.

Nixon, K. C. 1985. A biosystematic study of *Quercus* series *Virentes* (the live oaks) with phylogenetic analysis of Fagales, Fagaceae and *Quercus*. Dissertation thesis. University of Texas. Austin, Texas. 392 p.

Nixon, K.C. 1993. Infrageneric classification of *Quercus* (Fagaceae) and typification of sectional names. Annals Sci. For. 50 Suppl 1: 25s-34s

----------. 1999. Winclada (beta) ver. Published by the author. Ithaca, NY.

Olmstead, R.G., and J. D. Palmer. 1992. A Chloroplast DNA Phylogeny of the Solanaceae: Subfamilial Relationships and Character Evolution. Ann. Missouri Bot. Gard. 79: 346-360.

Palmer, J. D. & L. A. Herbon. 1988. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J. Mol. Evol. 28: 87-97.

Petit R. J., A. Kremer and D. B. Wagner. 1993. Geographic structure of chloroplast DNA polymorphisms in European oaks. Theoretical and Applied Genetics. 87:122-128.

Romero-Severson J, P. Aldrich, Y. Feng, W. Sun, and C. Michler. 2003. Chloroplast DNA variation of northern red oak *(Quercus rubra L.)* in Indiana. New Forests 26: 43-19.

Shaw, J., E. B. Lickey, E. E. Schilling, and R. L. Small. 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot. 94(3):275-288.

Simmons, M. P. and H. Ochoterena. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol 49:369-381.

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

Vázquez, M.L. 2001 Molecular and morphological studies on Mexican red oaks (*Quercus* sect. *Lobatae*). Ph D. Dissertation. Cornell University, New York, 290 pp.

Whittemore, A. T. and B. A. Schaal. 1991. Interspecific gene flow in sympatric oaks. Proc. Natl. Acad. Sci. 88: 2540-2544.

Wolfe, K.H., W.H. Li, and P.M. Sharp. 1987. Rates of nucleotide substitution vary greatly among mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. 84: 9054-9058.