Phylogenetic Analysis of *Dioon* Cycads Using Genes in Chloroplast is Inconclusive

By Ellie Coppock

# Abstract

Cycads are an understudied vascular plant group that includes the genus *Dioon.* The *Dioon* genus evolved during prehistoric aridification, making it an interesting case study in the evolution of traits related to water, like epidermal, vein, and photosynthetic traits. While other studies have looked into epidermal traits, not many have looked at the evolutionary history of photosynthetic genes. In this study, genes from the chloroplast are compared using phylogenetic inference tools to study the evolution of the genus. If the current understanding of their history is true, the clades in the phylogenetic analysis should mirror their current distribution. This knowledge can be used in current-day studies of conservation. The genus *Dioon* and other genera of cycads are noted as endangered or critically endangered. This study could give insight into the genus’ fate as the climate of their habitat continues to change and become warmer and more arid. The study of the evolution of *Dioon* in conjunction with its current distribution could help determine the best route for conservation for the genus. However, with inconclusive results, more work is necessary in order to make sense of the evolutionary history and current conservation of the genus.

# Introduction

Cycads are a type of ancient gymnosperm plant that includes the genus *Dioon.* The genus has a current distribution that spans Mexico and parts of Central America, specifically Honduras. *Dioon* includes 17 species.

*Dioon* is an ancient genus whose diversity peaked during the Jurassic period. The Jurassic period, which took place between approximately 200 and 145 million years ago, was characterized by a climate that was considerably warmer and more humid than the climate today (Nagalingum et al 2011). Currently, about 99.9% of *Dioon* of species that have existed are extinct (Condamine 2015). While species diversity reached its peak during the Jurassic period, most of the current species alive today evolved during the Neogene period. The Neogene, which took place between approximately 25 to 2.5 million years ago, is known to be far drier and cooler than the Jurassic period. This means that *Dioon* evolved largely during a period of global aridification. This means that the study of genes related to water usage like epidermal traits and traits related to photosynthesis can be helpful in making inferences about the evolution of the genus. Some studies show phylogenetic divergences related to epidermal traits (González-Astorga et al, 2008). This study can be used as a base for the phylogenetic analysis done in this paper.

Currently, *Dioon* exists on an aridity gradient in Mexico and parts of Central America. Mexico is known to have an aridity gradient where Northern regions are characteristically more arid and southern regions are more humid. In turn, different species within *Dioon* populate these varying regions. The genus spans from Sonora to Chiapas in Mexico with a species found in Honduras (*D. mejiae*). In Mexico, Oaxaca is known to be a region that is species-rich and holds 4 major complexes. Complex 1 is found in the Rio Santo Domingo Drainage area and includes 3 species: *D. purpusii, D. califanoi,* and *D. argenteum.* Complex 2 is found in the Pacific Coast and Upper Rio Balsas Drainage region and includes 2 species: *D. holmgrenii* and *D. caputoi.* Complex 3 is found in the Rio Tehuantepec Valley and Western Chiapas and includes 1 species: *D. merolae.* Lastly, Complex 4 is found in the Tropical Forests in Northern Oaxaca and includes 2 species: *D. spinulosum* and *D. rzedowskii.* Most of the distribution of species in Oaxaca has to do with the many river drainage areas found throughout the region. These habitats are always moving due to recent climate changes. Experts theorize that *Dioon* species migrate to river drainage areas and higher elevations during climate changes that result in warmer temperatures and more humid conditions, like the current climate change trends. This migration can lead to isolation and thus speciation within the *Dioon* genus.

The ancient genus of *Dioon* has proven to be an interesting case study in diversification during aridification events. This distribution of these plants can be analyzed in conjunction with aridification events and might prove to have implications for the current climate changes. Using phylogenetic inference tools on genes related to photosynthesis, specifically a gene located in the chloroplast, could prove helpful in completing the evolutionary puzzle of *Dioon.* Phylogenetic analysis can also highlight if species found within the current-day complexes are the most evolutionary similar or if they have evolved independently to occupy the same habitat. It can also give insight into conservation efforts by looking at evolutionary events that mirror the current climate crisis.

# Methods

The data used is an unpublished collection of cycad sequences from various species and genes. Sequences related to *Dioon* species were separated by hand and the sequences for chloroplast intergenic space were selected through a Blast search. The data set consists of 50 sequences.

For the alignment, two methods were used in order to ensure the most accurate alignment by comparing the results of multiple methods. The first alignment tool used was MAFFT (Katoh, 2002). MAFFT is a multi-sequence alignment tool that uses fast Fourier Transformation (FFT) to quickly detect homolog sequences within multiple sequences. MAFFT was chosen because of the ease of download and speed with the same accuracy as other notable methods like T-Coffee and ClustalW. The major limitation of MAFFT is, while the CPU time is reduced when compared to ClustalW and T-Coffee, the accuracy is similar, meaning that MAFFT might only be useful for trees with more taxa. This method assumes that the sequences are from the same gene sequence since the method searches for homologous sequences. The sequences used in this data set are from the same gene from organisms within the same genus.

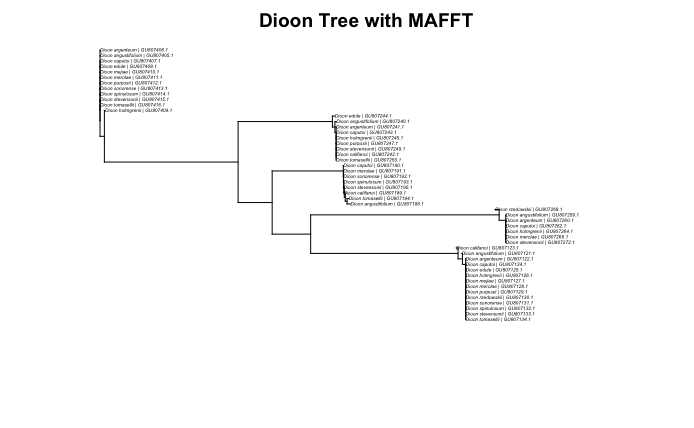
The second alignment method is MUSCLE (Edgar, 2004). MUSCLE is a multi-sequence alignment tool that uses k-mers to estimate evolutionary distances. MUSCLE was chosen because of the comparable accuracy to ClustalW and T-Coffee, both of which are less accessible and more difficult to download and use. The major limitations of MUSCLE are that test results indicating accuracy did not indicate significance between MUSCLE and other methods, despite claims otherwise. Additionally, in stage 1 of the alignment, speed is emphasized rather than accuracy. This method assumes a constant evolution rate (UPGMA), which can lead to higher accuracy alignments in some cases.

Using both alignments, 2 tree inference tools were used. The first one is Maximum Likelihood using IQ-Tree (Minh, 2020). IQ-Tree is a hill-climbing NNI inference tool. IQ-tree was chosen due to the accuracy and scalability of large data sets. Additionally, IQ-Tree 2 offers multiple evolutionary models and selects the best fit if an evolutionary model is not chosen for the data set. Some major limitations include that fast NNI can sometimes be less accurate than traditional NNI and that IQ-Tree, in some cases, can have a longer CPU time than other methods like RAxML. This method has few assumptions since IQ-Tree 2 was used instead of IQ-Tree 1, which assumes that there is one model of evolution. IQ-Tree 2 only assumes that the local optima identified in the search is the global optima of trees.

The second tree inference tool used is a distance method through R Studio using Ape. Distance methods are non-character-based methods using a specific model of evolution. The Distance method was chosen due to Distance methods being more accurate than other methods like parsimony and general familiarity with R Studio and R. The major limitation of this method is that the algorithm is largely based on estimations, though this does not necessarily mean it is completely inaccurate, it does indicate that it may have less solid calculations when compared to other methods. This method assumes a specific model of evolution that is set by the user. K80 (Kimera, 1980) was used in this analysis due to the rates of transitions and transversions being distinct and based on the structure of the nucleotide.

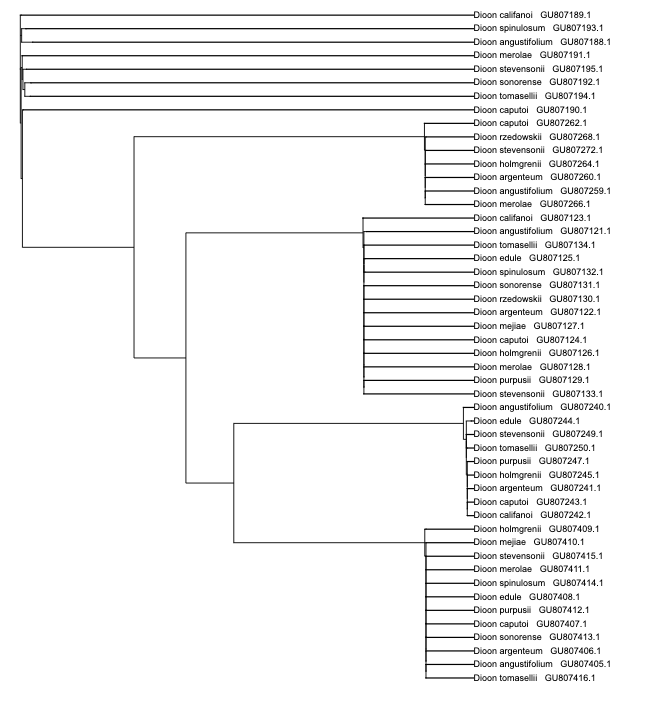
# Results

MAFFT Trees

MAFFT Distance Tree

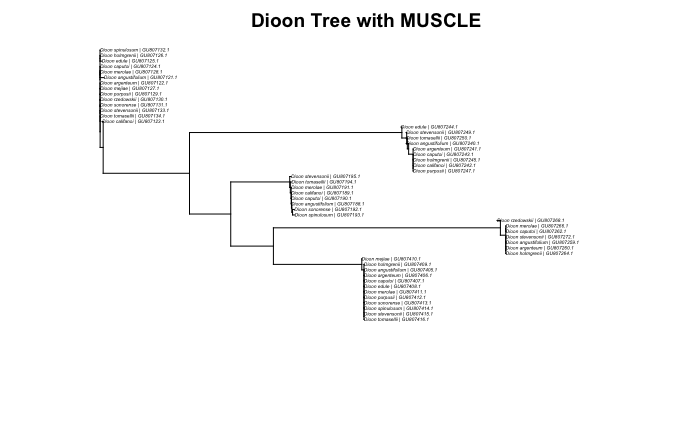
The MAFFT distance tree shows 5 clades. These clades are consistent with the MUSCLE distance tree but not the Maximum Likelihood trees.

MAFFT Maximum Likelihood Tree



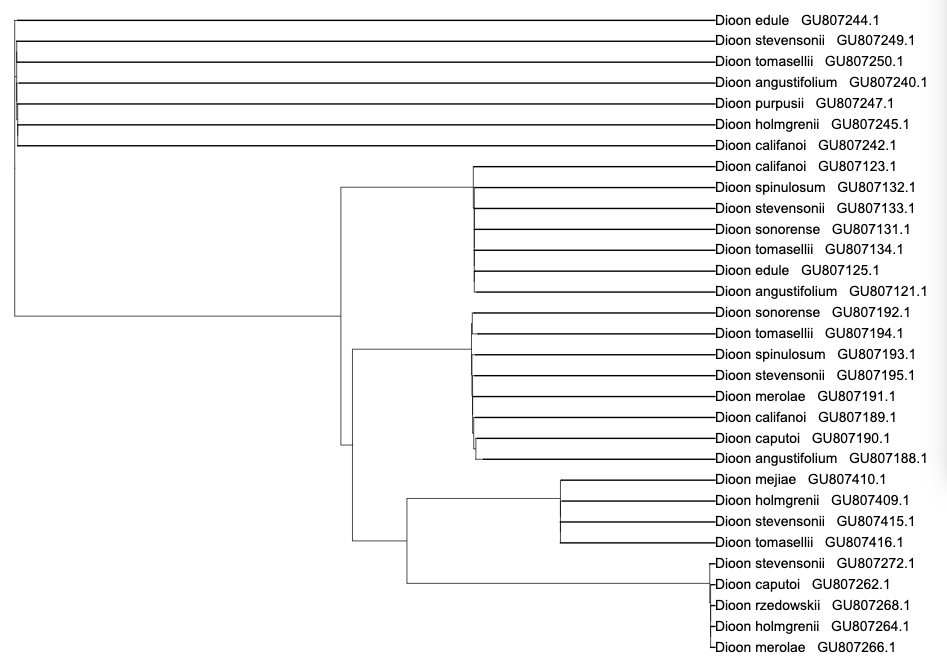
The MAFFT Maximum Likelihood tree also shows 5 clades. There is not much overlap between the previous MAFFT tree and this one.

MUSCLE Trees

MUSCLE Distance Tree

The MUSCLE Distance tree also has 5 clades and is consistent with the previous Distance tree using the MAFFT alignment.

MUSCLE Maximum Likelihood Tree



The Maximum Likelihood tree with the MUSCLE alignment has 5 clades like the other trees, but does not share much overlap with the other trees with either alignment.

# Discussion

Comparing the trees above, the only trees with consistency are the ones using the Distance method with the K80 evolutionary model. This might indicate that the Distance method is the most accurate when compared to the other methods. While the Distance method is clearly the method with the most consistency, the alignment methods are harder to differentiate. MAFFT appears to be better suited for nucleotide sequences than MUSCLE, which is better at protein sequences (nucleotides were used in this study). With Maximum Likelihood being inconclusive, it is difficult to determine the more accurate alignment method.

Compared to previous studies and the known current distributions in Mexico, the results of the Distance methods are not as similar as expected. This might be due to the lack of robust data used in similar studies, as the González-Astorga study only used 11 sequences and many of them were the same species from different sites. This also might be that the suspected trend of species in the same complexes evolving together is inaccurate. The most likely answer is that this phylogenetic analysis is not the best representation of the genus, whether that be from errors in method selection or errors in data processing.

Overall, while the results of this phylogenetic analysis using tree inference tools were inconclusive, tree inference tools are an incredibly useful tool in putting together genetic and evolutionary data and understanding the history of many organisms. In the future, more thorough analyses might be able to make more concrete conclusions about the evolutionary history of the genus. These future analyses will be incredibly useful in the conservation process of many species in the *Dioon* genus and of cycads as a plant group.

# 

# Work Cited

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, Robert Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era, *Molecular Biology and Evolution*, Volume 37, Issue 5, May 2020, Pages 1530–1534, <https://doi.org/10.1093/molbev/msaa015>

Condamine, F.L., Nagalingum, N.S., Marshall, C.R. *et al.* Origin and diversification of living cycads: a cautionary tale on the impact of the branching process prior in Bayesian molecular dating. *BMC Evol Biol* 15, 65 (2015). <https://doi.org/10.1186/s12862-015-0347-8>

JORGE GONZÁLEZ-ASTORGA, FRANCISCO VERGARA-SILVA, ANDREW P. VOVIDES, FERNANDO NICOLALDE-MOREJÓN, DÁNAE CABRERA-TOLEDO, MIGUEL ANGEL PÉREZ-FARRERA, Diversity and genetic structure of three species of *Dioon* Lindl. (Zamiaceae, Cycadales) from the Pacific seaboard of Mexico, *Biological Journal of the Linnean Society*, Volume 94, Issue 4, August 2008, Pages 765–776, <https://doi.org/10.1111/j.1095-8312.2008.01034.x>

Kazutaka Katoh, Kazuharu Misawa, Kei‐ichi Kuma, Takashi Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, Volume 30, Issue 14, 15 July 2002, Pages 3059–3066, <https://doi.org/10.1093/nar/gkf436>

Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci U S A. 1981 Jan;78(1):454-8. doi: 10.1073/pnas.78.1.454. PMID: 6165991; PMCID: PMC319072.

N. S. Nagalingum et al. ,Recent Synchronous Radiation of a Living Fossil. *Science* **334**, 796-799 (2011). DOI:[10.1126/science.1209926](https://doi.org/10.1126/science.1209926)

Robert C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, Volume 32, Issue 5, 1 March 2004, Pages 1792–1797, <https://doi.org/10.1093/nar/gkh340>