**A preliminary inquiry into the phylogenetica relationships of the Xylaria genus.  
Roberto Regalado**

**Abstract**

The Xylaria genus encompasses a diverse group of sordariomycete fungi with a broad ecological distribution, from tropical rainforests to temperate woodlands[10]. Known for their production of unique secondary metabolites, Xylaria species have garnered attention for their potential biomedical applications, having had several hundreds of novel bioactive compounds isolated from Xylaria sp.[9]. Despite their ecological significance and pharmacological potential, the evolutionary trajectory and phylogenetic relationships within the genus remain relatively unexplored. This study aims to address this gap by characterizing the phylogenetics of Xylaria species using the highly variable Internal Transcribed Spacer (ITS) region of the ribosomal RNA gene cluster. Leveraging modern phylogenetic analysis techniques, including Bayesian Markov Chain Monte Carlo (MCMC) inference implemented in MrBayes and maximum likelihood estimation in IQ-TREE2, we conduct a preliminary investigation into the evolutionary dynamics of selected Xylaria species. Additionally, we utilize Tracer software to assess the convergence and mixing properties of the MCMC analyses. Our findings shed light on the evolutionary history and ecological niches of Xylaria fungi while providing a framework for future genus-wide phylogenetic studies.  **Introduction**  
The Xylaria genus comprises a diverse array of sordariomycete fungi within the *Xylariaceae* family characterized by the production of perithecial ascocarps. Spanning tropical rainforests to temperate woodlands, the genus exhibits a broad ecological distribution. Notably, Xylaria species are recognized for their synthesis of numerous uncommon secondary metabolites, garnering significant attention for their prospective biomedical applications. However, despite the genus' ubiquitous ecological presence and its pharmacological potential, its evolutionary trajectory and phylogenetic relationships have remained relatively uncharacterized.  
  
The Internal Transcribed Spacer (ITS) region, a non-coding region of DNA nestled within the ribosomal RNA gene cluster, is highly conserved in fungi and yet, it displays high variability between fungal species. For this reason, it has emerged as a cornerstone in fungal phylogenetic studies.Top of Form

Bottom of Form

Characterizing the phylogenetics of the Xylaria genus using the ITS region of individual species in the genus can help to shed insight on its evolutionary history while also shedding light on the ecological niches it both has and potentially could carve out. Leveraging modern phylogenetic analysis techniques can yield valuable insights into the evolutionary dynamics of this understudied fungal taxa., In this paper, a preliminary inquiry into characterizing the phylogenetics of a small number of *Xylaria sp*. and strains using the ITS region is performed in hopes of providing a template for a larger genus-wide effort at a future date.

**Materials and Methods**

**Acquiring datasets**

To run phylogenetic analysis on *Xylaria sp*., a data set is required. To date, the largest repository of genetic Information on *Xylaria sp*. resides on the National Center for Biotechnology Information internet data base (NCBI) [[1](https://www.ncbi.nlm.nih.gov/)]. Searching the “all databases” searchbar for the term “Xylaria ITS” on NCBI returned the accession number AF163036.1. This Accession number is correlated to the internal transcribed spacer region of sample of *Xylaria hypoxylon* strain CBS 590.72. Using the blastn function of the NCBI website, a nucleotide blast was performed on accession number AF163036.1. This resulted in an output of dozens of the ITS sequences of *Xylaria sp*. (Table1). Previous phylogenetic studies of the Xylaria genus identify the Neurospora genus as a closely clustering independently of Xylaria species while still having deep evolutionary relationships with them[[2](https://scholarsbank.uoregon.edu/xmlui/handle/1794/23459)]. It is for this reason that the Neurospora genus was selected as an outgroup for phylogenetic analysis. The fasta files for all of these taxa were downloaded from NCBI.

|  |  |
| --- | --- |
| Accession number | Organism |
| MN219591.1 | *Xylaria corniformis* |
| MG098261.1 | *Xylaria longipes* strain NW-FVA2228 |
| KP133498.1 | *Xylaria scruposa* isolate 866 |
| AF163036.1 | *Xylaria hypoxylon* CBS 590.72 |
| JX256828.1 | *Xylaria schweinitzii* voucher HMJAU 22751 |
| MT661488.1 | *Xylaria polymorpha* strain DSM 105756 |
| MG098262.1 | *Xylaria polymorpha* strain NW-FVA2229 |
| MW447065.1 | *Xylaria polymorpha* strain FeC105 |
| MN846336.1 | *Xylaria polymorpha* isolate TW07032019\_02 |
| MK595684.1 | *Xylaria sp.* isolate MBD\_3013 |
| NR\_159859.1\* outgroup | *Neurospora cratophora* CBS 558.94 |

**Table 1** Dataset for phylogenetic analysis

**Alignment of dataset**

The sequences downloaded from NCBI were unaligned fasta files. In order to perform accurate phylogenetic analysis, a computational technique used to align three or more biological sequences called Multiple sequence alignment (MSA) is performed in order to identify similarities, differences, and conserved regions among them. For this study, two Multiple sequence alignment software’s were used for this process, the first of which are MAFFT (Multiple Alignment Using Fast Fourier Transform)[[3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603318/)], which utilizes progressive alignment in addition to iterative refinement protocols

MAFFT employs state-of-the-art algorithms employing Fast Fourier transforms to produce accurate alignments, even for large datasets and highly divergent sequences. It utilizes progressive alignment strategies combined with iterative refinement methods to improve alignment accuracy. This software is designed for user friendliness and has excellent scalability. For this reason it was selected as one of the softwares used for this study

**Installing MAFFT**

A thorough guide to installing MAFFT can be found in the Supplementary materials [S1 section 1] containing reproducible script. Since MAFFT relies more heavily on pairwise iterate alignment a scoring matrix was not used opting instead for the default algorithm states.

The second alignment software used was MUSCLE (Multiple Sequence Comparison by Log-Expectation)[[4](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-113)]. This alignment software Initiates its multiple sequence alignment by making a seed alignment using either Kmer or pairwise techniques which it progressively re-aligns. The software employs guide trees to facilitate the progressive alignment steps. In contrast to MAFFT, MUSCLE produces alignments much more rapidly and is more sensitive to local sequence similarities. In comparison to MAFFT, muscle is also more ideal for smaller datasets. It is for this reason that MUSCLE was selected for the relatively small dataset involved.

**Installing MUSCLE**

A thorough guide to installing MUSCLE can be found in the Supplementary materials [S1 section 2] containing reproducible script. For the purposes of simplicity the default scoring matrix was used when performing the MUSCLE alignment

**Maximum likelihood analysis Software**

To infer a more probable evolutionary tree from a set of different sequences, Maximum likelihood phylogenetic analysis should be performed. Maximum likelihood analysis assesses the tree topology and branch lengths that provide the maximum probability of sequences using a preselected evolutionary model. For the purposes of this study, IQtree 2[5] was the maximum likelihood software used. IQ-TREE 2 incorporates advanced model selection techniques to identify the best-fit substitution models for sequence evolution while also using bootstrap capabilities to help validate trees  
  
For both MAFFT and MUSCLE alignments, In each maximum likelihood analysis, the evolutionary model of TNe+G4 was used.  
  
For both MAFFT and MUSCLE alignments, In each maximum likelihood analysis, the number of bootstrap iterations performed were 1000

**IQtree 2 installation**  
A thorough guide to installing IQtree 2 can be found in the Supplementary materials [S1 section 3] containing reproducible script. For the purposes of simplicity the default scoring matrix was used when performing the MUSCLE alignment

**RStudio Installation for visualizing trees**Rstudio[6] is a coding studio for the programming language known as R that provides a user-friendly accessible interface for editing R-code. A wide variety of output files generated by phylogenetic applications in the terminal can be visualized in Rstudio including Fasta files (.fasta, .fa, .fna) Newick Tree Files (.tree, .nwk, .treefile) Nexus Tree Files (.nex) and many others. In order to visualize the trees from the phylogenetic analysis performed, Rstudio was used.

**Installing Rstudio**  
A thorough guide to installing MUSCLE can be found in the Supplementary materials [S1 section 4] containing reproducible script.

**Bayesian analysis software**

Bayesian Markov Chain Monte Carlo (MCMC) analysis is often used in phylogenetic analysis because it uses iterative algorithms to generate parameter values that will eventually arrive at the posterior distribution, it also incorporates prior information. This allows it to accurately handle very large datasets. It is for these and other reasons that MCMC is superior to maximum likelihood and other phylogenetic analysis methods. For the purposes of this study The Bayesian inference software MrBayes[7] was used to complete the analysis. For both MAFFT and MUSCLE alignments, In each MCMC analysis, the evolutionary model of TNe+G4 was used.  
  
For both MAFFT and MUSCLE alignments, In each MCMC analysis 1,000,000 generations were run.  
  
**Installing MrBayes**

A thorough guide to installing MUSCLE can be found in the Supplementary materials [S1 section 5]

**Using tracer to analyze MCMC**  
Tracer[8] is a software tool commonly used in molecular phylogenetics to assess the convergence and mixing properties of Markov Chain Monte Carlo (MCMC) analyses  
**Installing Tracer**

A thorough guide to installing Tracer can be found in the Supplementary materials [S1 section 6] containing reproducible script

**Results**

**DISTANCE AND PARSIMONY TREES**After aligning the data sets with both MUSCLE and MAFFT softwares, RStudio was used to plot both distance and parsimony trees from the aligned files. Unrooted MAFFT distance and parsimony trees showed significant differences in structure, node groups and node lengths[Fig 1]. . While unrooted MUSCLE aligned trees were identical [Fig 2]

A screenshot of a computer screen

Description automatically generated

**Fig 1** unrooted distance and parsimony trees created in RStudio derived from MAFFT multiple sequence alignment.

Unrooted Muscle distance and parsimony trees showed significant differences in structure, node groups and node lengths.

A screenshot of a computer screen

Description automatically generated

**Fig 2** unrooted distance and parsimony trees created in rstudio derived from MUSCLE multiple sequence alignment.

**MAXIMUM LIKELYHOOD TREES**

Maximum likelihood analysis was performed on both MAFFT and MUSCLE alignments[Fig 3]. From these alignments, recommended evolutionary models and bootstrap values were acquired. .contree output files were loaded into RStudio to visualize the trees with their respective bootstrap values. The average of all bootstrap values of the MAFFT alignment was 70.2, while the average of all bootstrap values for the muscle alignment was 68.2

A screenshot of a computer

Description automatically generated

**Fig 3** rooted Iqtree2 trees created in RStudio derived from MAFFT and MUSCLE multiple sequence alignment.

**Bayesian analysis**

Using MrBayes, Bayesian analysis was conducted on both MUSCLE and MAFFT alignments. Bayesian analysis yielded identical trees for both alignment softwares[Fig 4,5]. All posterior probability values were 100%

A computer screen shot of a tree

Description automatically generated

**Fig 4** rooted tree conducted on MAFFT alignment

MCMC posterior probabilities for MAFFT aligned files

|  |  |  |
| --- | --- | --- |
| Node | Name | Posterior Probability % |
| 1 | *NR\_159859.1\_Neurospora\_cratophora\_CBS\_558.94* | 100 |
| 2 | *MN219591.1\_Xylaria\_corniformis\_small* | 100 |
| 3 | *MG098261.1\_Xylaria\_longipes\_strain\_NW-FVA2228* | 100 |
| 4 | *KP133498.1\_Xylaria\_scruposa\_isolate\_866* | 100 |
| 5 | *AF163036.1\_Xylaria\_hypoxylon\_CBS\_590.72* | 100 |
| 6 | *JX256828.1\_Xylaria\_schweinitzii\_voucher\_HMJAU\_22751* | 100 |
| 7 | *MT661488.1\_Xylaria\_polymorpha\_strain\_DSM\_105756* | 100 |
| 8 | *MG098262.1\_Xylaria\_polymorpha\_strain\_NW-FVA2229* | 100 |
| 9 | *MW447065.1\_Xylaria\_polymorpha\_strain\_FeC105* | 100 |
| 10 | *MN846336.1\_Xylaria\_polymorpha\_isolate\_TW07032019\_02* | 100 |
| 11 | *MK595684.1\_Xylaria\_sp\_\_isolate\_MBD\_3013* | 100 |

**Table 2** Node locations and Posterior probabilities of MAFFT aligned files

A screenshot of a computer

Description automatically generated

**Fig 5** rooted tree conducted on MUSCLE alignment.

MCMC posterior probabilities for muscle aligned files

|  |  |  |
| --- | --- | --- |
| Node | Name | Posterior Probability % |
| 1 | *NR\_159859.1\_Neurospora\_cratophora\_CBS\_558.94* | 100 |
| 2 | *MN219591.1\_Xylaria\_corniformis\_small* | 100 |
| 3 | *MG098261.1\_Xylaria\_longipes\_strain\_NW-FVA2228* | 100 |
| 4 | *KP133498.1\_Xylaria\_scruposa\_isolate\_866* | 100 |
| 5 | *AF163036.1\_Xylaria\_hypoxylon\_CBS\_590.72* | 100 |
| 6 | *JX256828.1\_Xylaria\_schweinitzii\_voucher\_HMJAU\_22751* | 100 |
| 7 | *MT661488.1\_Xylaria\_polymorpha\_strain\_DSM\_105756* | 100 |
| 8 | *MG098262.1\_Xylaria\_polymorpha\_strain\_NW-FVA2229* | 100 |
| 9 | *MW447065.1\_Xylaria\_polymorpha\_strain\_FeC105* | 100 |
| 10 | *MN846336.1\_Xylaria\_polymorpha\_isolate\_TW07032019\_02* | 100 |
| 11 | *MK595684.1\_Xylaria\_sp\_\_isolate\_MBD\_3013* | 100 |

**Table 3** Node locations and Posterior probabilities of Muscle aligned files

After trees were constructed using Bayesian analysis, tracer plots were created for both the MUSCLE AND MAFFT aligned files to test for convergence and mixing. Tracer plots for both alignments showed that convergence and proper mixing was achieved [Fig 6,7]

A screenshot of a computer

Description automatically generated

**Fig 6** A tracer plot of the muscle aligned file.

**A screenshot of a computer

Description automatically generated**

**Fig 7** A tracer plot of the muscle aligned file.

**Discussion**

**Distance and parsimony trees**

In this study, we investigated the phylogenetic relationships within several species of the genus Xylaria using Maximum likelihood analysis, Distance and phylogenetic trees, and Bayesian MCMC analysis based on molecular data. Our results function as a preliminary exploration into this genus in hopes of a true genus wide analysis at a future date. Distance trees of MAFFT aligned files showed significant difference from their parsimony tree counterparts, specifically with respect to which node included the Neurospora outgroup. This result may be attributable to the simple difference in how distance and parsimony trees are constructed, however this does not account for the fact that the distance and parsimony trees from the MUSCLE aligned files were identical. This perhaps can be explained by the fact that MUSCLE is generally regarded as more ideal for smaller datasets and could possibly have produced more accurate trees than MAFFT in this instance.

**Maximum likelihood trees**

Maximum likelihood analysis produced trees that were identical, with the exception that, when taking all of the bootstrap values and dividing by ten (the number of organisms sans the outgroup) MAFFT aligned files had an average bootstrap value 2 points higher than MUSCLE aligned files. Both trees correctly plotted the Neurospora outgroup as such, indicating that both alignment softwares were generally in agreement about the data set, indicating support for the trees being accurate. It should be noted that both MUSCLE and MAFFT maximum likelihood trees differed significantly from the distance and parsimony trees. This is likely due to the fact that maximum likelihood trees tend to be inherently more accurate. Another observation worth pointing out is that the bootstrap values were consistently high (100%) for *X. longipes* and *X. corniformis*, but started to hover around 50% for all of the X. polymorpha strains, indicating that the confidence for these nodes being correct is not particularly high.

**Bayesian analysis**

The Bayesian MCMC analysis allowed us to infer a well-supported phylogenetic tree that resolved the relationships among the sampled Xylaria species with high posterior probabilities. The use of Bayesian inference provided several advantages over traditional methods such as maximum likelihood and distance-based approaches. Furthermore, the posterior probabilities obtained from the Bayesian MCMC analysis provided valuable information about the confidence levels of inferred phylogenetic relationships. High posterior probabilities (>0.95) supported most of the major clades in our tree, indicating support for these evolutionary relationships.

**Supplemental Materials**

Complete Reproducible script as well as an index for installation of all software [S1] can be found at the following path

/Users/robertoregalado/Desktop/final-project-repository/Scripts/Finalized\_Reproducible\_Script.md

**References**

[1] *National Center for Biotechnology Information*. (n.d.). https://www.ncbi.nlm.nih.gov/

[2] Ly, A. (2015). *Xylaria primers for phylogenetic reconstruction*. https://scholarsbank.uoregon.edu/xmlui/handle/1794/23459

[3] Katoh, K., & Standley, D. M. (n.d.). *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability*. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/mst010

[4] Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1). https://doi.org/10.1186/1471-2105-5-113

[5] Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and Efficient Methods for phylogenetic inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

[6] *R: The R Project for Statistical Computing*. (n.d.). https://www.r-project.org/

[7] Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D., Darling, A. E., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

[8] *Tracer | BEAST Documentation*. (n.d.). https://beast.community/tracer

[9] Chen, W., Yu, M., Chen, S., Gong, T., Xie, L., Liu, J., Bian, C., Huang, G., & Zheng, C. (2024). Structures and Biological Activities of Secondary Metabolites from Xylaria spp. *Journal of Fungi*, *10*(3), 190. https://doi.org/10.3390/jof10030190

[10] Suwannasai, N., Sangvichien, E., Phosri, C., McCloskey, S., Wangsawat, N., Thamvithayakorn, P., Ruchikachorn, N., Thienhirun, S., Mekkamol, S., Sihanonth, P., Whalley, M. A., & Whalley, A. (2023). Exploring the Xylariaceae and its relatives. *Botanical Studies*, *64*(1). https://doi.org/10.1186/s40529-023-00389-6