Multilocus Evolutionary Analysis Using Probabilistic Topic Modeling

Marzieh (Tara) Khodaei and Peter Beerli

Department of Scientific Computing, Florida State University, Tallahassee FL
EXPO 2023

Methodology

We present a new computational approach using k-mers and probabilistic topic modeling [1], an unsupervised machine learning approach based on natural language processing, to construct evolutionary relationships among species from unaligned DNA sequences. We base our development on a software module of the PHYLIP computer package by Felsenstein, CONTML (**Cont**inuous Characters **M**aximum **L**ikelihood method) [3], which estimates phylogenies from frequency data.
Figure 1 illustrates the key steps of our method: First, it learns a probabilistic topic model from a dataset of gene sequences and extract the topic frequencies of sequences using Latent Dirichlet Allocation (LDA) technique [2]. Second, it estimates the phylogeny using the trained topic frequencies and CONTML.

Application to Real Data

We evaluate our approach using a dataset that was previously published [4]: The sequences are collected from 14 loci and 9 different locations. For each locus, the length of each sequence varies from 288 to 418 base pairs, and the number of sequences varies from 78 to 92 individuals. For each locus, we applied LDA on a continent-wide (Australia) scale across populations and extract the topics for 9 locations. Then, we applied these topic frequencies of 14 loci in CONTML to generate the tree.

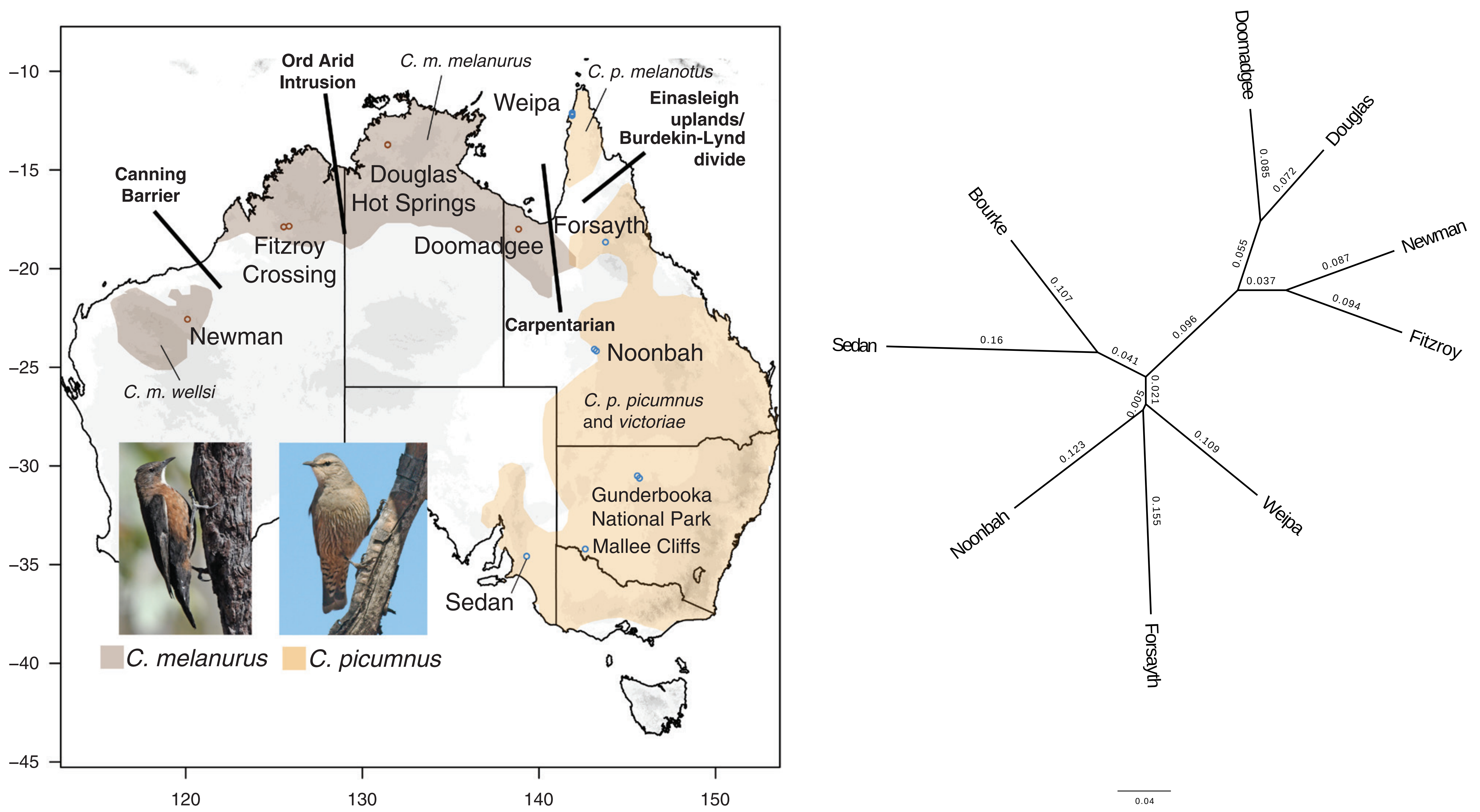

Figure 2. Left: 9 locations at which the birds were collected [4]; Right: phylogeny constructed using our method

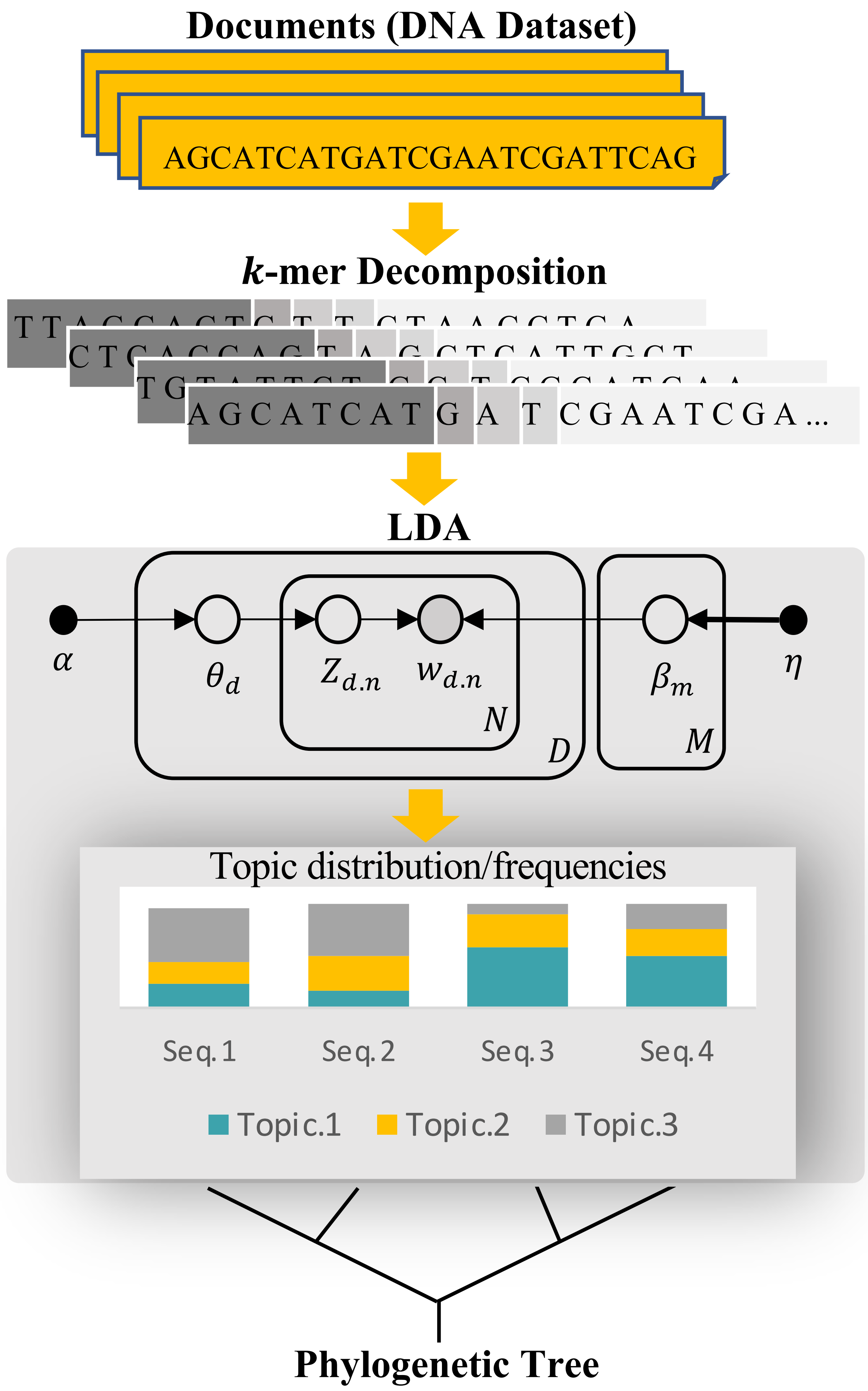

Figure 1. Workflow of topic modelling to generate topic frequencies and the corresponding phylogeny.

Biological data classification and taxonomic identification have a significant role in evolutionary biology and bioinformatics. Most current approaches use a two-step procedure to classify biological data: (1) alignment of the biological sequences, (2) analysis of this alignment. The alignment-free approaches are gaining more attention from the scientific community because of their ability to overcome the drawbacks of sequence alignment techniques. We present an alignment-free technique based on probabilistic topic modeling to extract topic frequencies for the dataset, and then we apply topic frequencies as an input to CONTML to generate the phylogeny.

References

[1] Griffiths, T.L. and Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl_1), pp.5228-5235
[2] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3(Jan), pp.993-1022.
[3] Felsenstein, J., 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution, pp.1229-1242.
[4] Edwards, S.V., Tonini, J.F., Mcinerney, N., Welch, C. and Beerli, P., 2023. Multilocus phylogeography, population genetics and niche evolution of Australian brown and black-tailed treecreepers (Aves: Climacteris). Biological Journal of the Linnean Society, 138(3), pp.249-273