Multilocus Phylogenetic Tree Estimation Using Topic Modeling

Marzieh Khodaei[1], **Scott V. Edwards**[2], **and Peter Beerli**[1]

[1] *Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA*
[2] *Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, Cambridge MA 02138, USA*

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 aligned/unaligned DNA sequences. The method is implemented in the Python package TopicContml (`https://github.com/TaraKhodaei/TopicContml`). It is based on a two-phase pipeline: (1) it learns a probabilistic topic model from a multi-locus or genome-wide data set of DNA sequences and extracts the topic frequencies of sequences using Latent Dirichlet Allocation (LDA) model [2] for each locus (Fig. 1 left). (2) It uses these topic frequencies from multiple loci to estimate a species tree with Contml in the PHYLIP package [3] (Fig. 1 right).

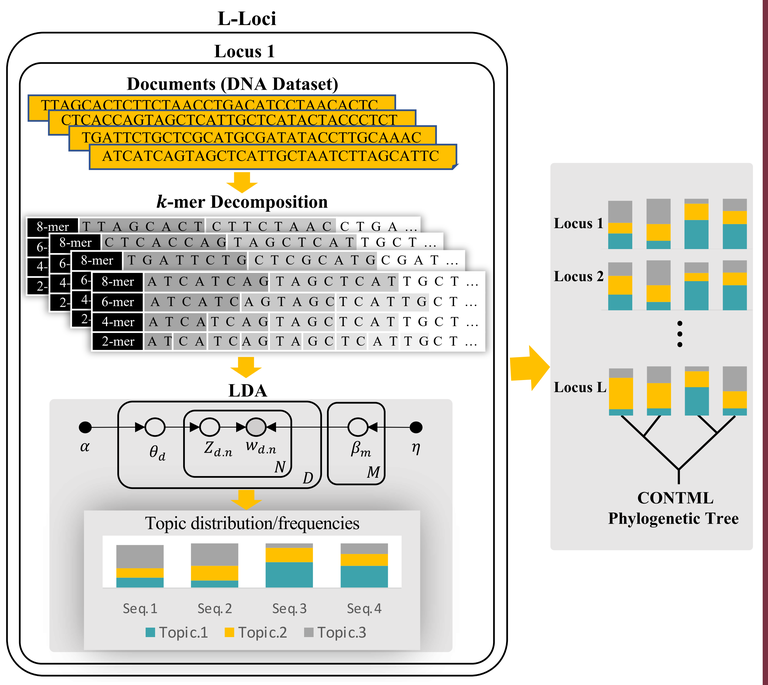
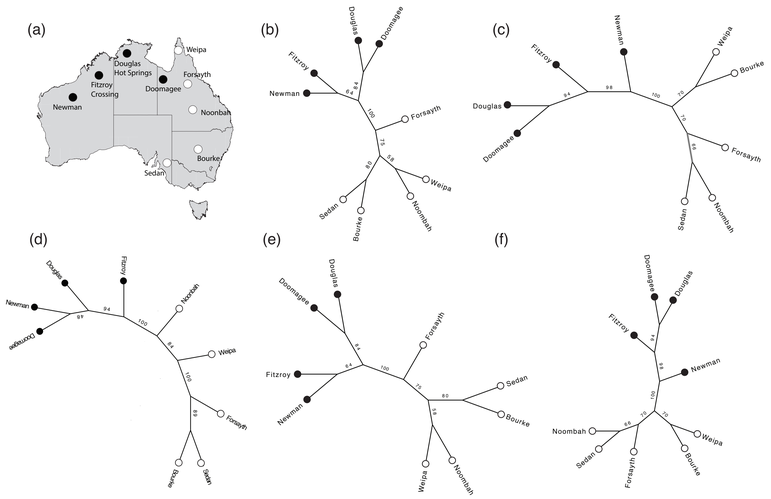


Figure 1. TopicContml workflow.

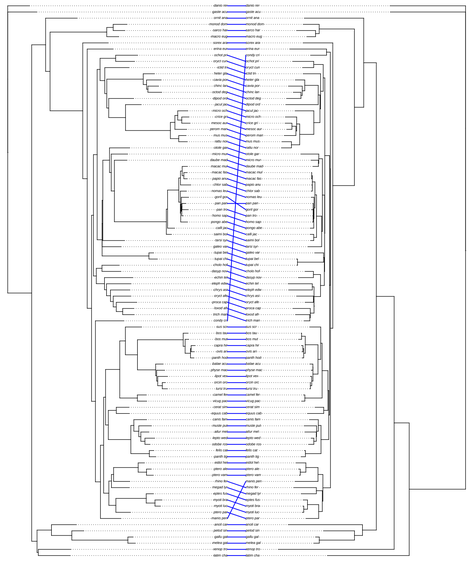
Application to Real Data Sets

- **Bird dataset –** The sequences are collected from 14 loci of 9 populations of 2 bird species, Australian brown treecreepers (white disks) and black-tailed treecreepers (black disks) [4]. For each locus, sequence length varied from 288 to 418 base pairs, and the number of aligned sequences varied from 78 to 92 haplotypes. To obtain unaligned sequences, we removed all gaps in each sequence.
  (a) The map of Australia showing 9 locations. (b) The best phylogeny constructed by TopicContml from unaligned sequences with bootstrap support. (c) The best phylogeny constructed by TopicContml from aligned sequences. Bootstrap support comparison with 1000 replicates: (d) Majority-consensus tree of aligned data analyzed by SVDquartets [5]. (e) Majority-rule consensus tree, using SumTrees in DendroPy [6], of unaligned data by TopicContml. (f) Majority-rule consensus tree of aligned data by TopicContml.
- **Mammal dataset –** The data set contains 90 vertebrate species focusing on mammals with 5162 loci. Figure shows the tanglegram of the TopicContml tree (left side) compared to the Maximum likelihood tree of [7] (right side). The left tree used TopicContml after all site columns containing gaps or 'N' were removed. Our tree and the maximum likelihood tree deliver similar answers, with a distance of 56. The runtime for 90 species and 5162 loci is substantial.



Bird dataset

Mammal dataset



References

[1] Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci,101(suppl_1):5228-35,2004.
[2] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res, 3:993–1022, 2003.
[3] Felsenstein J. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution, pages 1229–1242, 1981.
[4] Edwards SV, Tonini JF, Mcinerney N, Welch C, Beerli P. Multilocus phylogeography, population genetics and niche evolution of Australian brown and black-tailed treecreepers (Aves: Climacteris). Biol J Linn Soc, 138(3):249–273, 2023.
[5] Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics, 30(23):3317–3324, 2014.
[6] Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics, 26(12):1569–1571, 2010.
[7] Liu et al. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc Natl Acad Sci, 114(35):E7282–E7290, 2017.

Inferring the evolutionary history of species or populations with genome-wide data is gaining ground, but computational constraints still limit our abilities in this area. We developed an alignment-free method to infer the genome-wide species tree and implemented it in the Python package TopicContml. The method uses probabilistic topic modeling (specifically, Latent Dirichlet Allocation or LDA) to extract 'topic' frequencies from k-mers, which are derived from multilocus DNA sequences. These extracted frequencies then serve as an input for the program Contml in the PHYLIP package, which is used to generate a species tree.

Scan the QR code and download a copy of the bioRxiv paper.

This project was partly funded by NSF grant DBI 2019989 to Peter Beerli, and by NSF grant number MCB-2344806 to Marzieh Khodaei.